ALGODAN Algorithmic Data Analysis

Data Mining and Computational Creativity

Prof. Hannu Toivonen
Discovery Group
University of Helsinki and HIIT

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Discovery Group

1 ■ Prof. Hannu Toivonen

+1 ■ Alessandro Valitutti, Postdoc (until Dec 2013)

- Ping Xiao from 2014

+3

■ Laura Langohr, PhD student

■ Oskar Gross, PhD student

■ Jukka Toivanen, PhD student

■ Fang Zhou, PhD 2012

■ Mika Timonen, PhD 2013 (VTT, Techn. Res. Centre of Finland)

■ Esther Galbrun, PhD 2014 (co-supervised with Mikko Koivisto)

■ Joonas Paalasmaa, PhD 2014 (Beddit.com Ltd.)

Affiliate members

■ Assoc. Prof. Antoine Doucet (U. Caen, France)

■ Dr. Tommi Opas (entrepeneour)

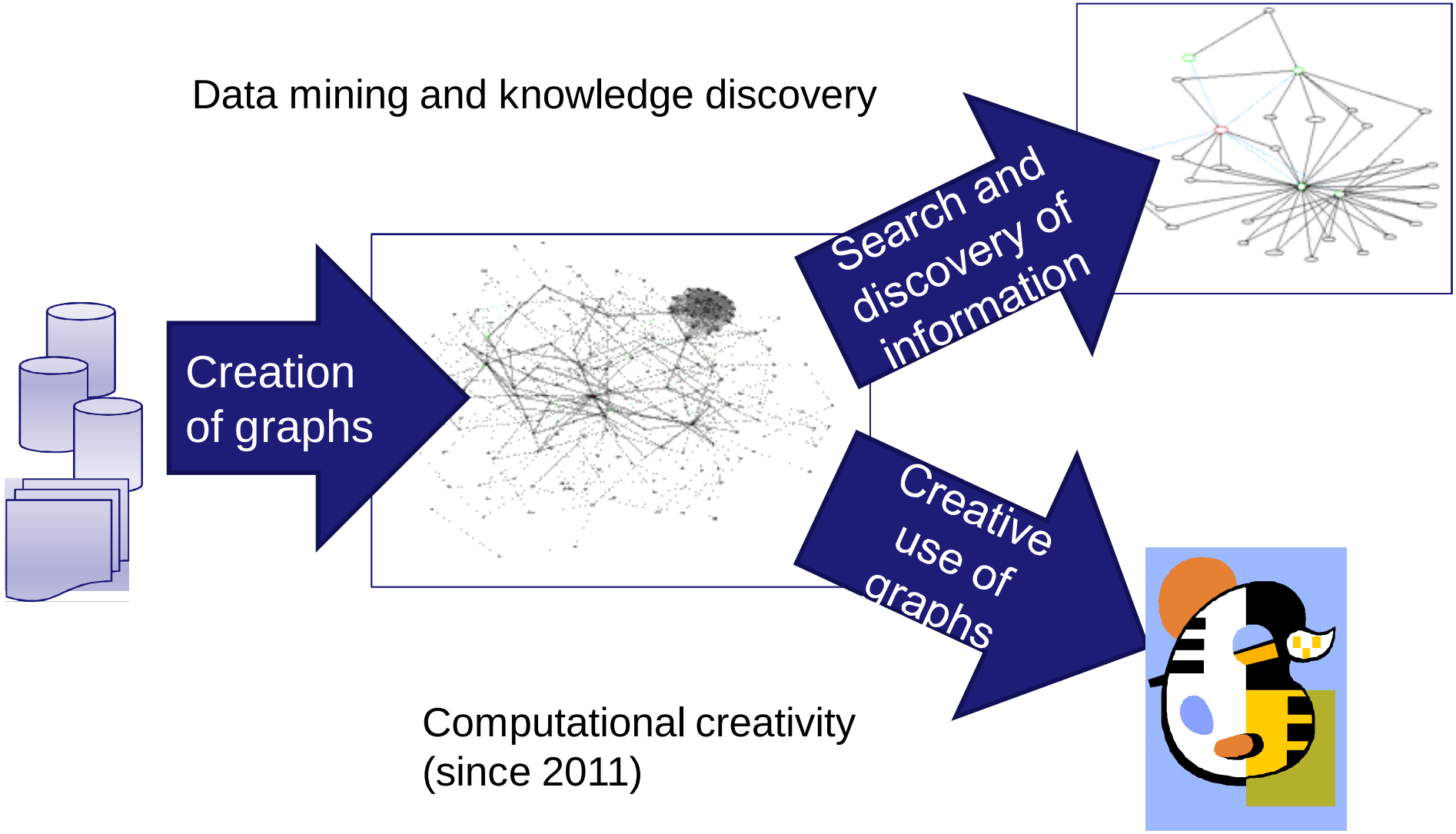
Mission

We develop novel methods and tools for data mining and computational creativity.

- Algorithms for discovering links and patterns in data
- Their use in creative systems
  - E.g. poetry generation

Evolution of research topics

Data mining and knowledge discovery



Computational creativity
(since 2011)

Computational creativity

1. Creative computers
2. Computers supporting human creativity
3. Studies of creative computational processes

Why study computational creativity?

- An ultimate AI challenge
- A test bed for AI/ML/DM methods
- Applications: games, user interfaces, creativity support
- An intellectual challenge

International conference ICCC since 2010

Computational poetry

Music swells, accent practises, theatre hears!
Her delighted epiphanies bent in her universe:
– And then, singing directly a universe she disappears!
An anthem in the judgements after verse!

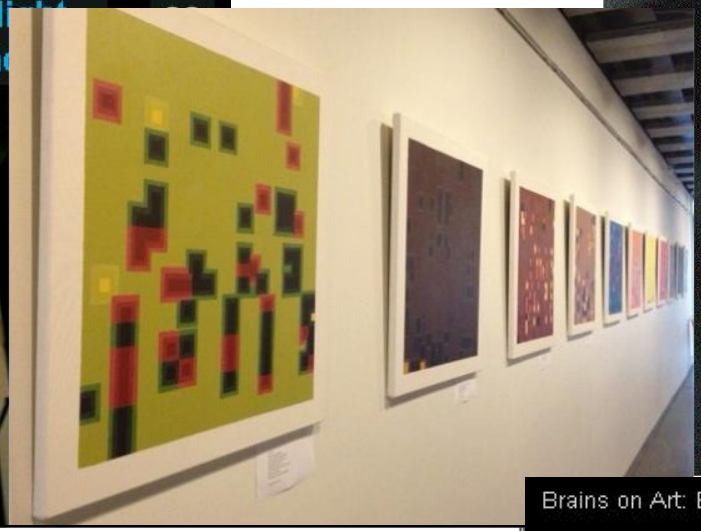
Computational poetry (ICCC 2012, 2013)

- No explicit grammars, rules or semantics given
  - Except for rhymes in this example
- Instead, utilization of existing texts both for the form (syntax) and the content (lexical selection)
- Endless data mining opportunities
  - Analysis of language use beyond simple co-occurrence
  - Analysis of example poetry, learning styles etc.
- Applications e.g. in supporting creativity
  - An interactive tool for playful practice of writing in grammar schools

Future work

- Focus on computational creativity
  - Using data mining and graph mining
  - Learning and adaptive creative systems
  - “Concept creation technology”, EU FP7, 2013-16
- Establish further contacts and collaboration with scientists in applied fields (e.g. literature, cinema)
  - “Promoting scientific exploration of computational creativity”, EU FP7 CSA, 2013-16

Societal impact: Contributions to arts



Societal impact: Press

The Times, 22 Jan 2014

The Mail (Online), 22 Jan 2014

Yle TV2, 14 Sep 2012

Helsingin Sanomat, 8 Sep 2012

New Scientist, 13 Oct 2012

CNET (Online), 12 Sep 2012

CBC Radio, 11 Nov 2012

Technical and economical impact

Impact comes from data mining research with companies:

- Collaboration with Finnish media companies on news analysis
- Sleep analysis research commercialized by Beddit Ltd, see the following talk by Joonas Paalasmaa

Publications 2011-2013

Computational creativity

- ICCC (Computational Creativity) x 3
- KICSS (Creativity Support)
- IDA, Frontier award winner
- ACL
- CICLing

Data mining in bioinformatics

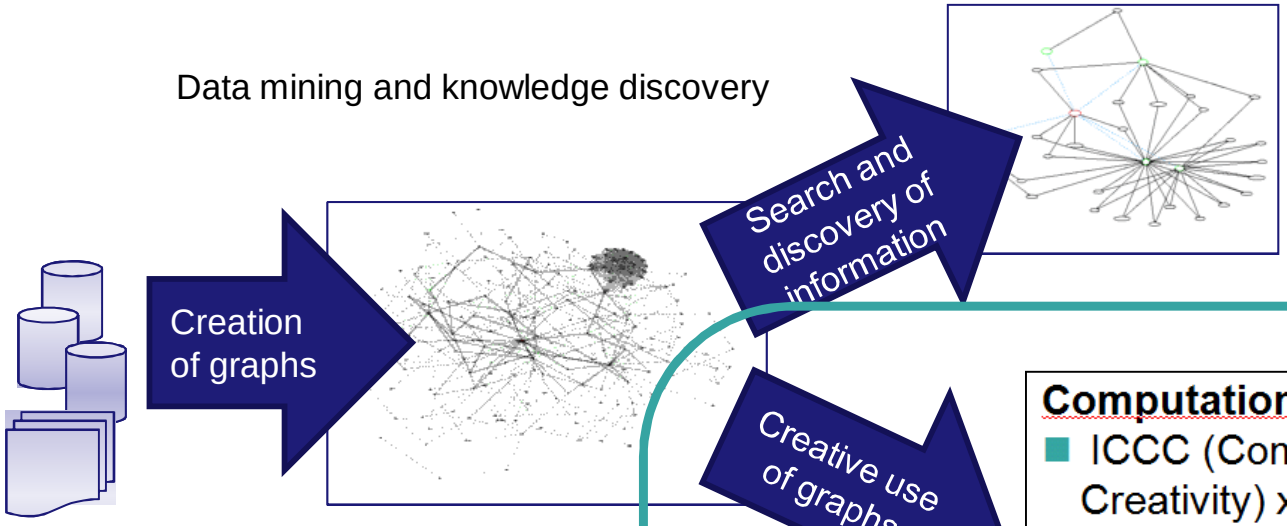
- BMC Bioinformatics x 2
- PloS One
- IEEE EMBS

Data mining

- Machine Learning
- Statistical Analysis and Data Mining
- Computer x 2
- Print Media Technology Research
- SIGKDD x 2
- SDM
- DS
- ASONAM
- Bisociative Knowledge Discovery (book) x 5

Impact of CoE funding

Data mining and knowledge discovery



Computational creativity
(since 2011)

Computational creativity

- ICCC (Computational Creativity) x 3
- KICSS (Creativity Support)
- IDA, Frontier award winner
- ACL
- CICLing

Societal impact: Contributions to arts

THE TIMES
Mail Online
yle
HS
NewScientist
c|net
CBCradio

A risky move to renew science, made feasible by CoE funding

ALGODAN Algorithmic Data Analysis

Monitoring Sleep with Force Sensor Measurement

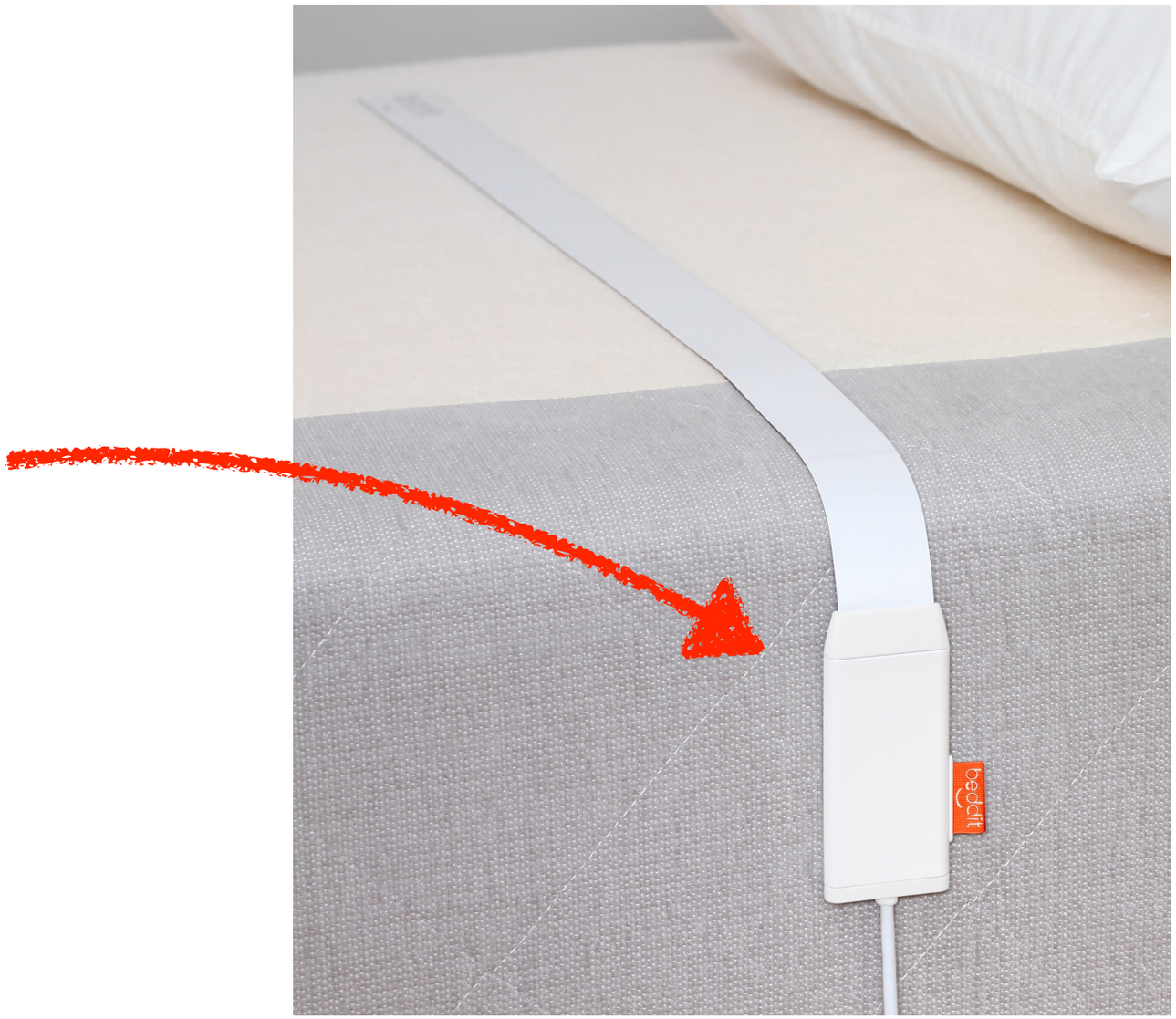
Joonas Paalasmaa

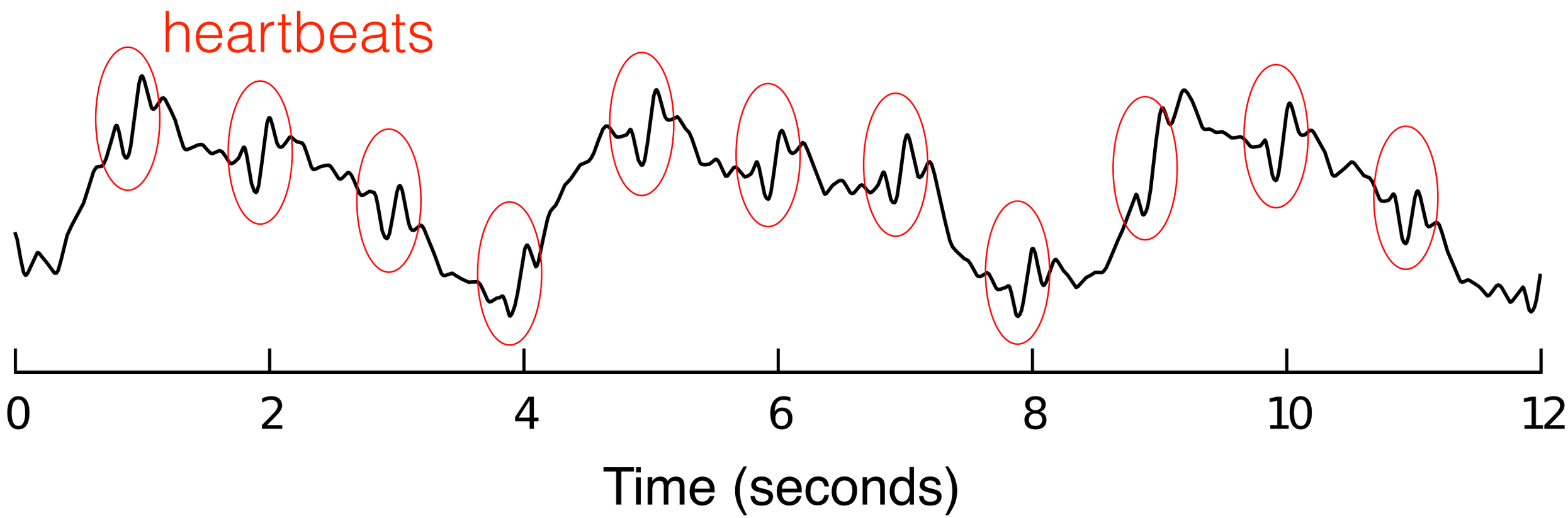
Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

My PhD thesis from last month: *Monitoring Sleep with Force Sensor Measurement*

Developing sleep measurement signal processing methods for this kind of sleep sensors




heartbeats
0
2
4
6
8
10
12
Time (seconds)

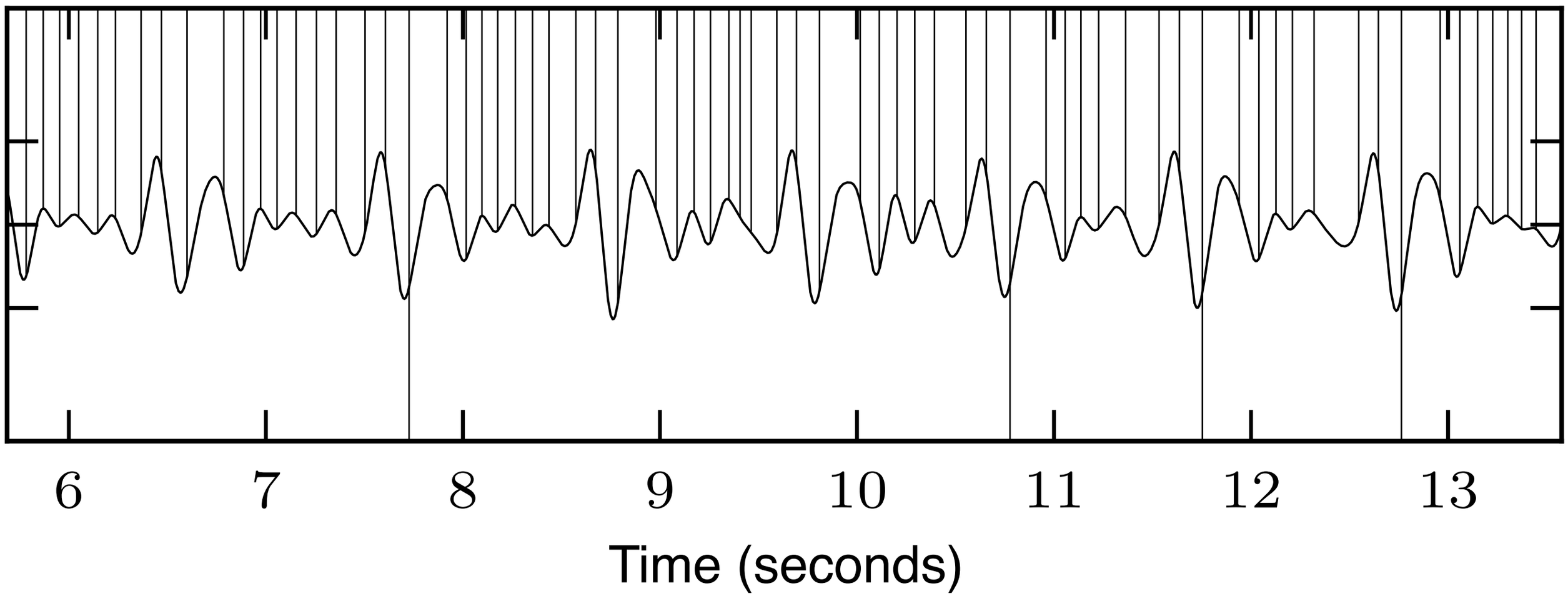
Signal processing problems to solve

- Detect from the signal: **heart rate**, **respiration**, **movements**

- It is very difficult, because signal properties vary by person, bed, etc. **Unsupervised** methods are needed.

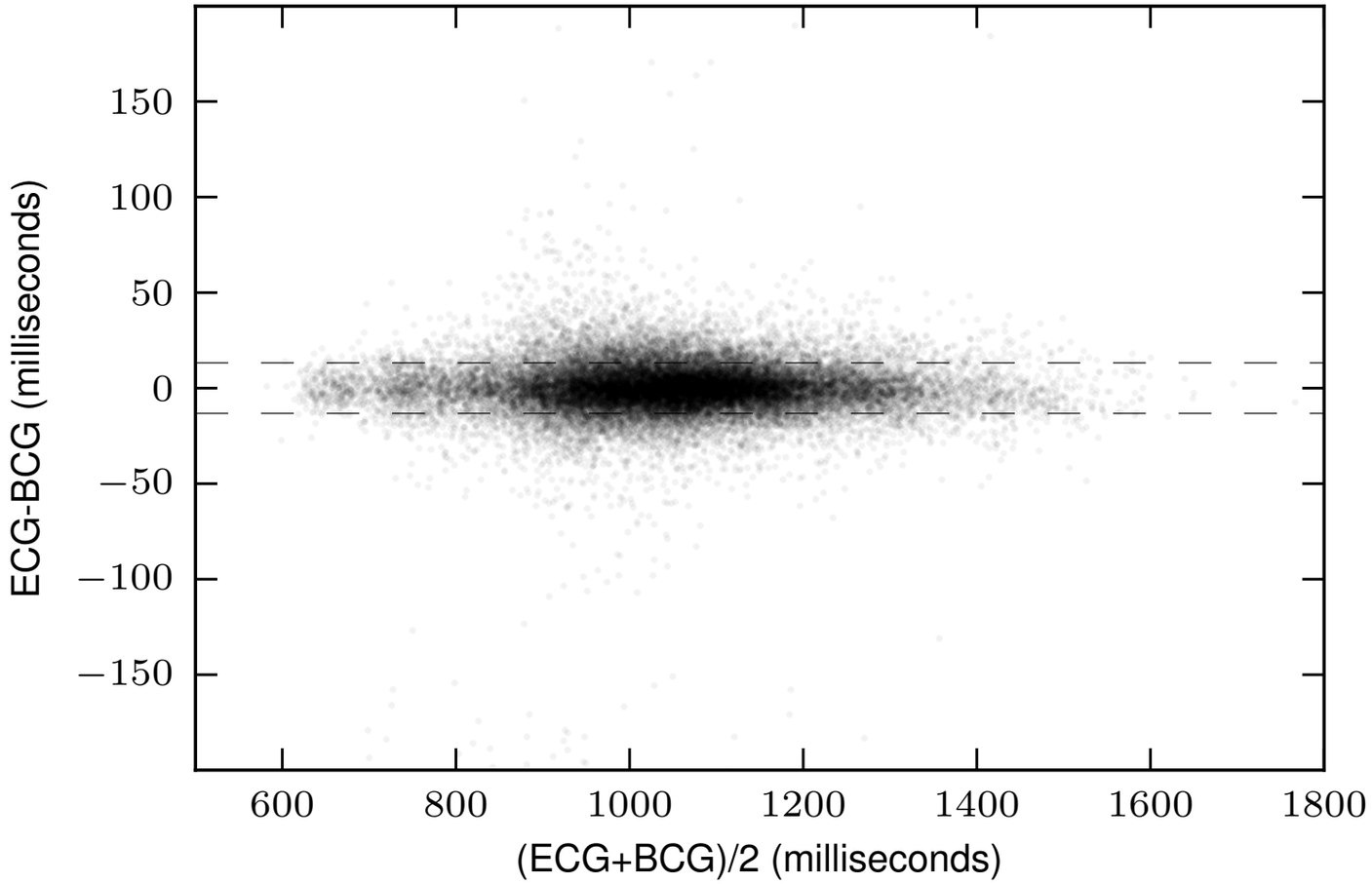
Other machine learning problems to solve

- Based on heart rate, respiration and movements, classify sleep into: **sleep stages**, **sleep cycles**, **snoring**, …

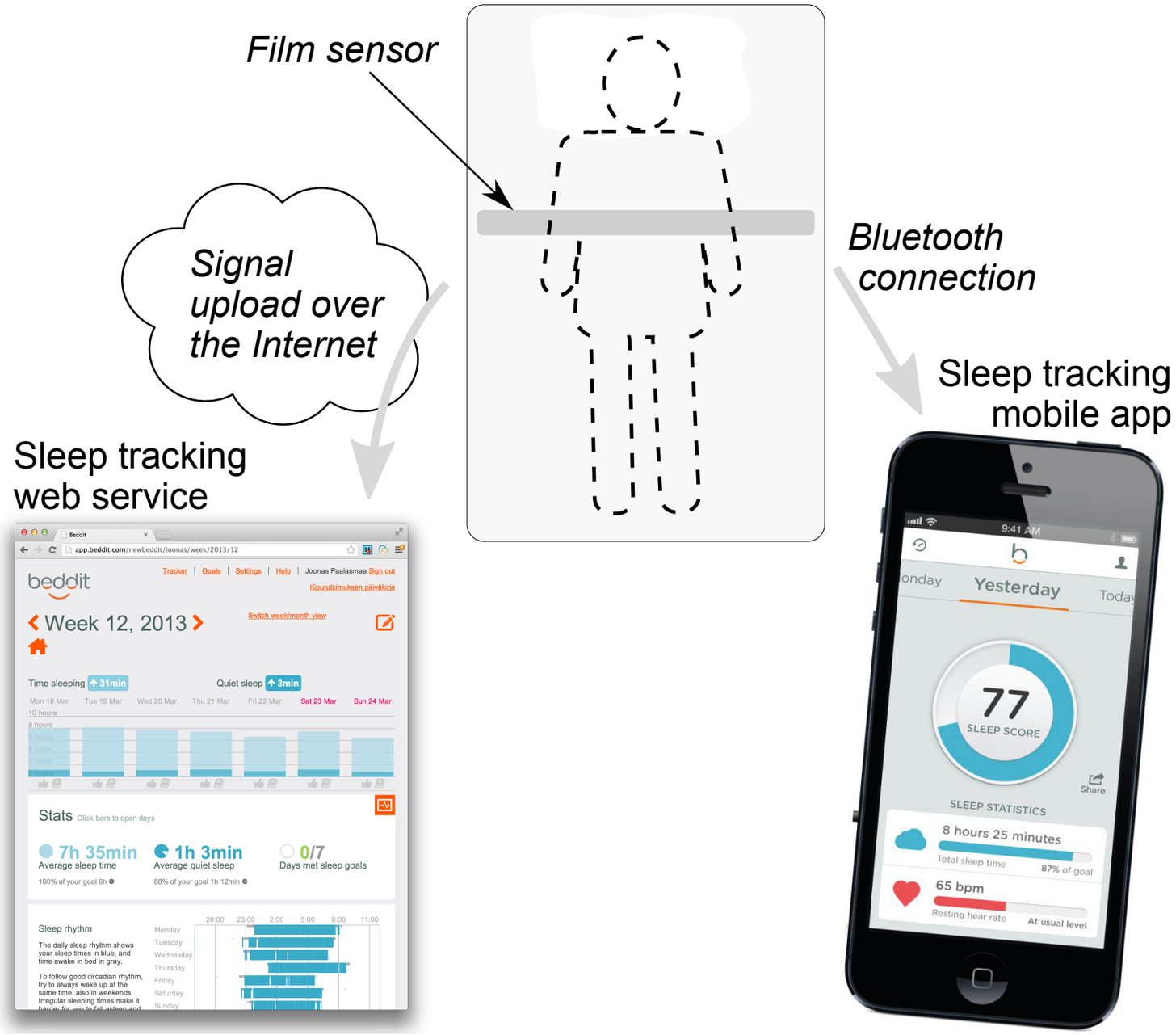

6
7
8
9
10
11
12
13
Time (seconds)

Clinical validation

- 40-person clinical study in 2010 (patients at the lab)
- 20-person clinical study in 2012 (volunteers at their homes)



Turning research into business — Beddit Ltd



Beddit

https://app.beddit.com/newbeddit/joonas/week/2014/6

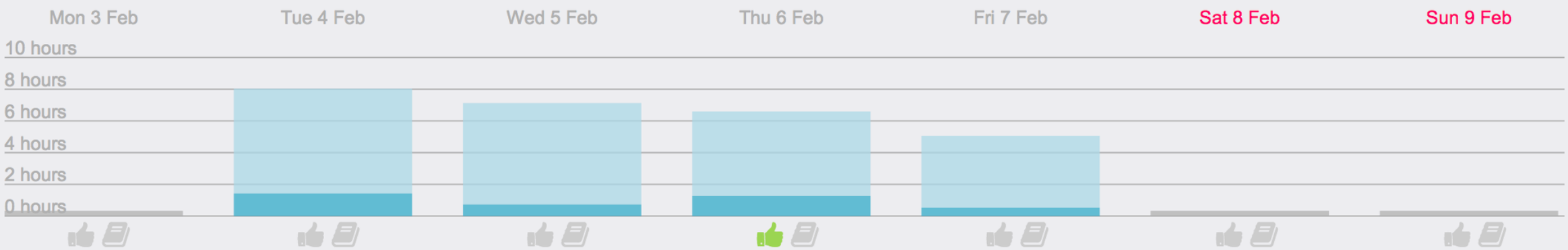
beddit

Tracker | Goals | Settings | Help | Joonas Paalasmaa Sign out

Week 6, 2014

Switch week/month view

Time sleeping 6h 42min



Stats

Click bars to open days

6h 36min
Average sleep time
100% of your goal 6h

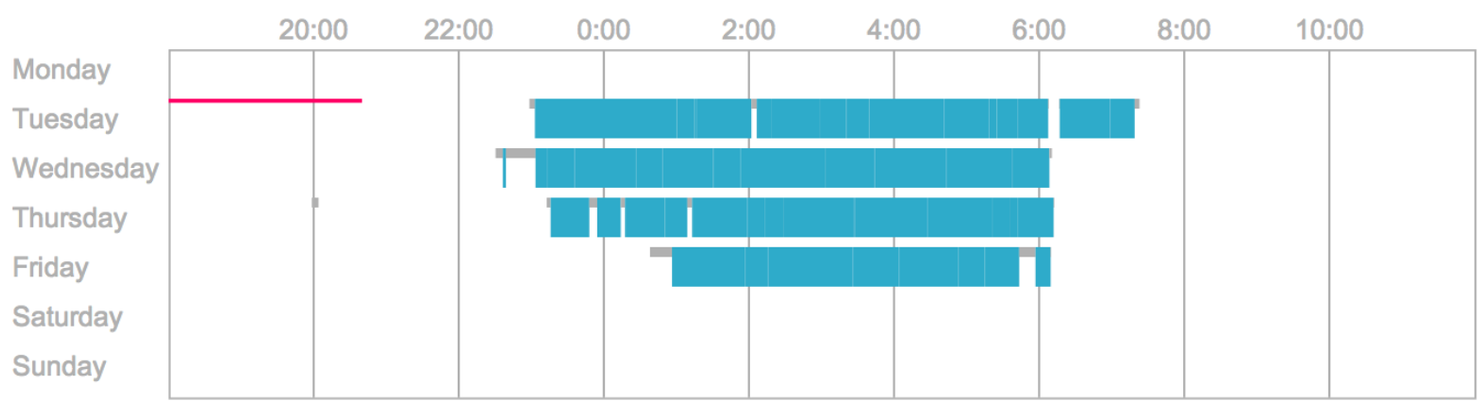
59min
Average quiet sleep
82% of your goal 1h 12min

2/7
Days met sleep goals

Sleep rhythm

The daily sleep rhythm shows your sleep times in blue, and time awake in bed in gray.

To follow good circadian rhythm, try to always wake up at the same time, also in weekends. Irregular sleeping times make it harder for you to fall asleep and to wake up.



Beddit – Automatic sleep a

http://www.beddit.com/static/img/sensor-phone-screenshot-tip.jpg

indiegogo browse | learn | create Sign Up | Log In search by title

We're celebrating International Women's Day all week long! Check out these campaigns empowering women & girls! BROWSE #WOMENSDAY CAMPAIGNS

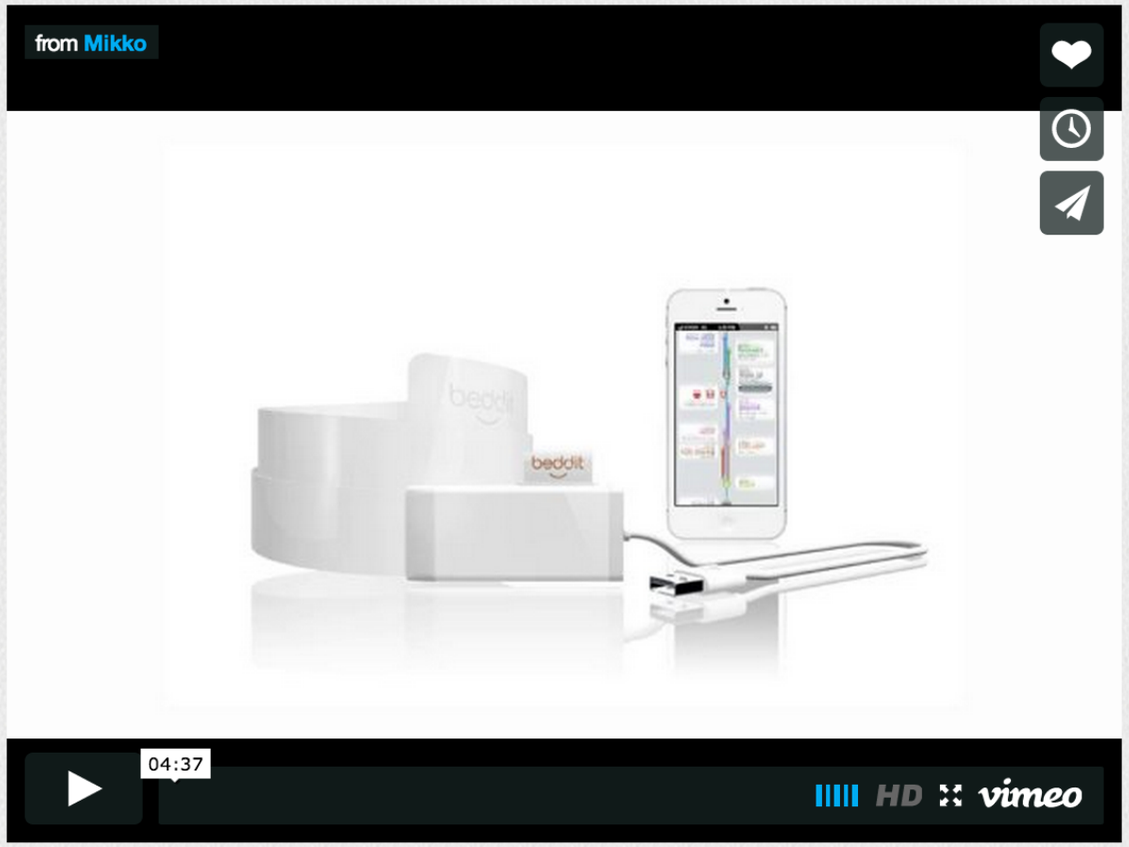
Beddit - Automatic sleep and wellness tracker. Turn your bed into a smart bed.

The Beddit sensor tracks your sleep quality, heart rate, and breathing under the sheet while you sleep. The app coaches you to improve sleep and performance.

Technology – Saratoga, California, United States

Campaign Home | Updates / 20 | Comments / 803 | Funders / 3981

from Mikko



Share This Campaign: http://igg.me/at/beddit-sleep-tracker Follow

$503,571 USD

Raised of $80,000 Goal

0 time left

Flexible Funding

This campaign has ended and will receive all funds raised. Funding duration: August 02, 2013 - October 15, 2013 (11:59pm PT).

Select a Perk *for your contribution*

$99 USD Featured

Beddit

You will receive the Beddit ultra thin film sensor that automatically tracks your sleeping patterns, heartbeats, and breathing under

Summary

- **Scientific contributions**: signal analysis methods (*Physiological Measurement*, 2010. *IEEE EMBS*, 2011, 2012. *IEEE J-BHI*, under review.)

- **Medical contributions**: clinical validation of methods

- **Commercial activity**: raised over $500 000 in a crowdfunding campaign in 10 weeks. 10000 units produced so far.

Exact algorithms

Petteri Kaski
Department of Information and Computer Science
Aalto University, Helsinki

ALGODAN Scientific Advisory Board
18 March 2014

Aalto University
School of Science

ALGODAN
Algorithmic Data Analysis

Combinatorics, Algebra, and Computing (CO-ALCO)

Members

- Mikko Koivisto, Academy Research Fellow (8/2008-10/2013), Professor (1/2013-), Co-leader
- Petteri Kaski, Academy Research Fellow (9/2011-), 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

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.

ALGODAN themes

F = Foundations of algorithmic data analysis

D = Discovery of hidden structure in data

CO-ALCO Highlights, 2011 & beyond

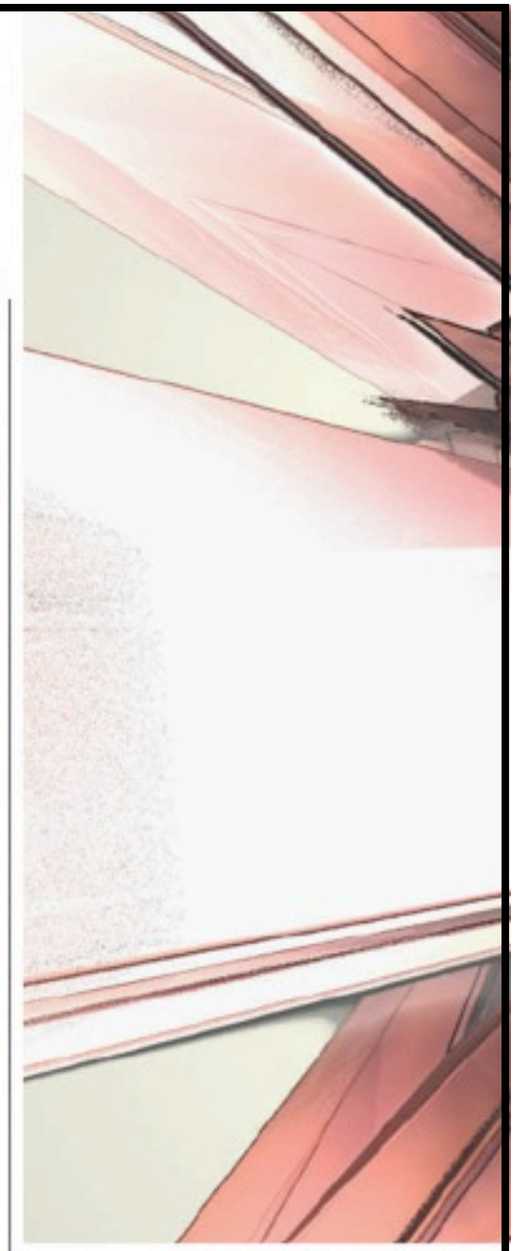
- Graphical models from data — Theme D
  (exact algorithms for polytrees & for Bayesian networks parameterized by treewidth; AAAI'12, AISTATS'13, JMLR 2013)
- Discovering connected motifs in graphs — Theme D
  (linear-time in the size of the host graph; STACS'13)
- Fast "Fourier analysis" on partially ordered sets — Theme F
  (fast Möbius inversion on lattices, counting thin subgraphs; SODA'12 & SODA'14)

DOI:10.1145/2428556.2428575

Discovering surprises in the face of intractability.

BY FEDOR V. FOMIN AND PETTERI KASKI

Exact Exponential Algorithms



MANY COMPUTATIONAL PROBLEMS have been shown to be intractable, either in the strong sense that no algorithm exists at all—the canonical example being the undecidability of the Halting Problem—or that no *efficient* algorithm exists. From a theoretical perspective perhaps the most intriguing case occurs with the family of *NP*-complete problems, for which *it is not known* whether the problems are intractable. That is, despite extensive research, neither is an

of non-parameterized instances of intractable problems? At first glance, the general case of an *NP*-complete problem is a formidable opponent: when faced with a problem whose instances

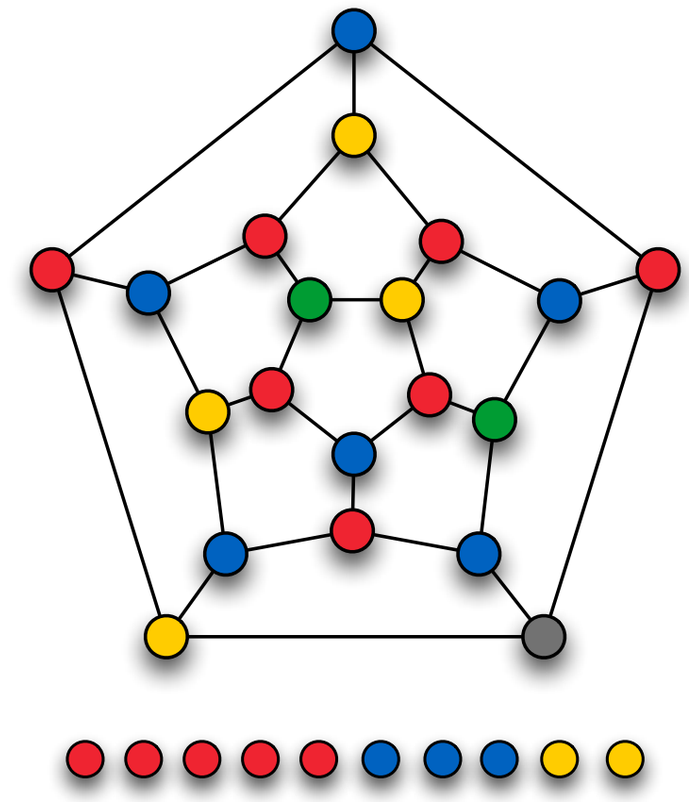
F. V. Fomin & P. Kaski
Communications of the ACM
March 2013, pp. 80–88.

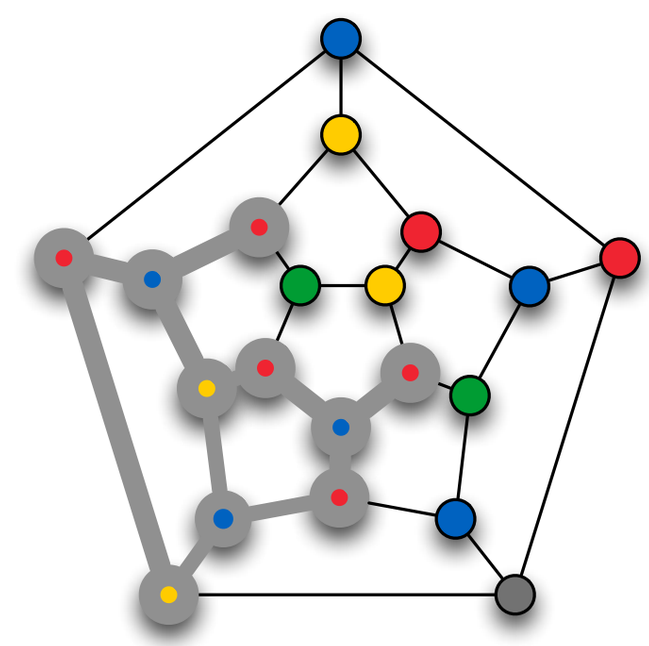
But what can we say about finding exact solutions

surprises have emerged recently.

Graph motifs

- **Input**
  - A vertex-colored host graph H
  - A multiset M of colors (**the motif**)
- **Question**
  Does H have a *connected* subgraph whose vertex colors agree with M?
- **Main result** (Björklund, K., & Kowalik, STACS'13)
  - Let H have n vertices and e edges, $e \geq n$
  - Let M have size k
  - There is a randomized algorithm that runs in time $O(2^k k^3 e)$ with
    (i) no false positives, and
    (ii) false negatives with probability ≤ 0.001





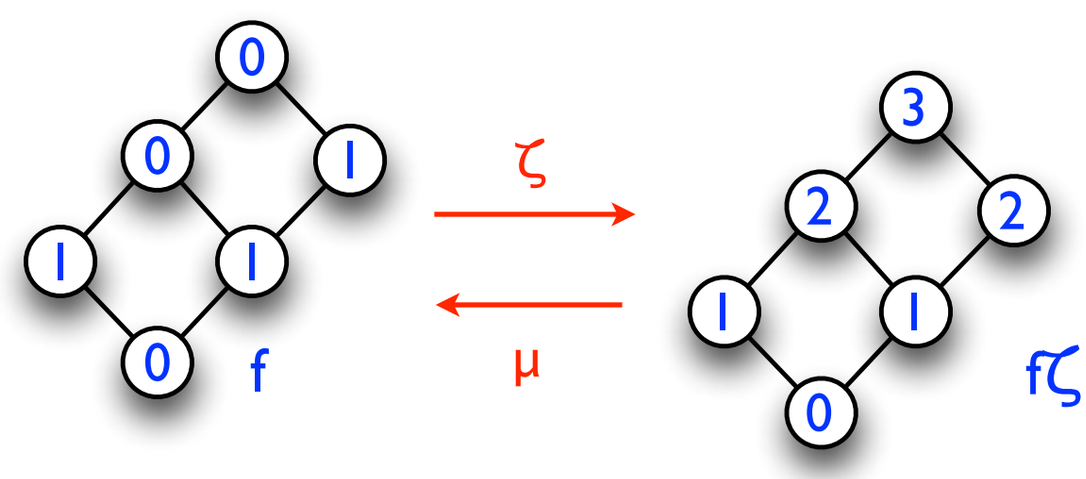
1) Linear time in the size e of the host graph
2) NP-hard problem, but exponentiality only in k, $k << n$
3) There is evidence that an $O(1.9999^k \text{ poly}(k,e))$ algorithm is unlikely

Fast Möbius inversion ("Fourier analysis") on partially ordered sets

(Björklund, Husfeldt, K., Koivisto, Nederlof, Parviainen; SODA'12 & ACM TALG, to app.)

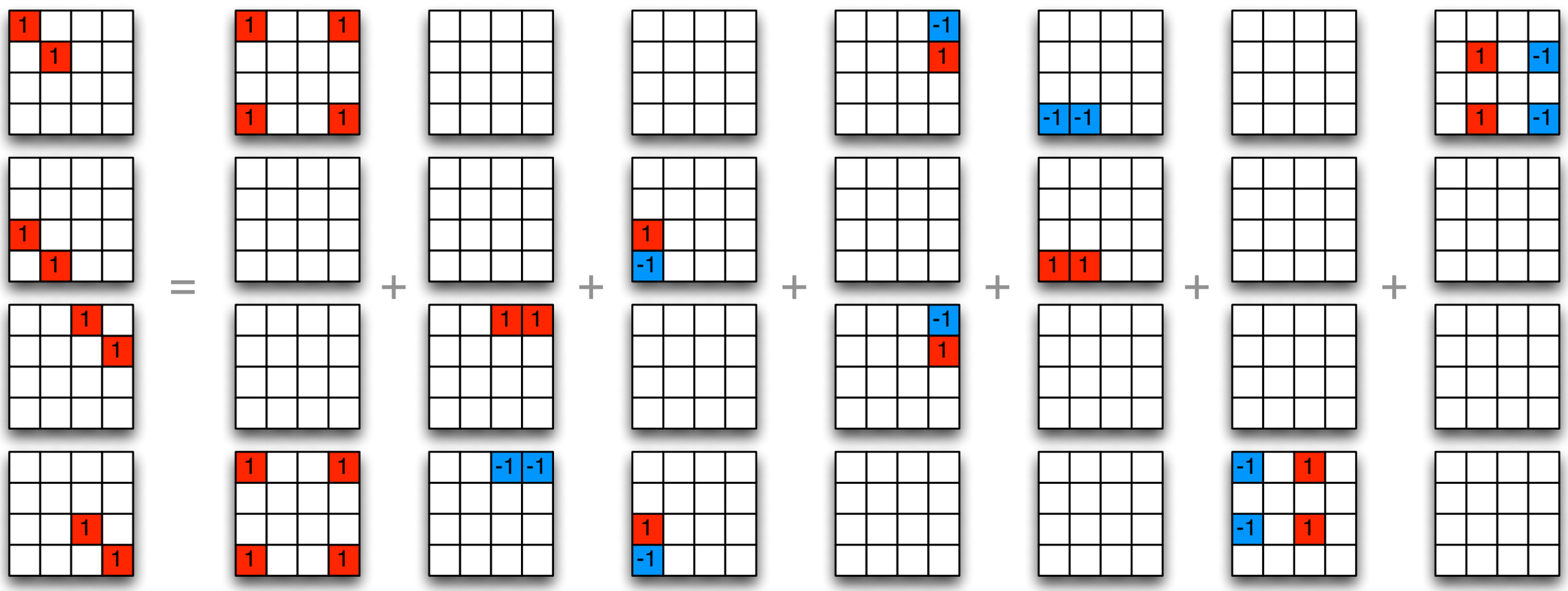
	analogy	
• Discrete Fourier analysis	⟷	• Möbius inversion
• (Fast) Fourier transform	⟷	• (Fast) zeta transform ζ
• (Fast) inverse Fourier transform	⟷	• (Fast) Möbius transform μ
• (Fast) convolution	⟷	• (Fast) join product
... on a finite group		... on finite lattices

(e.g. the cyclic group
= classical discrete
Fourier analysis)



Previous slide summarized: The structural tensor of the join product has “low (tensor) rank”

... but so has the structural tensor of the square matrix product ...



(Strassen’s 1969 low-rank tensor decomposition for 2x2 matrix product)

Counting (thin subgraphs) in three parts faster than "meet-in-the-middle" time

(Björklund, K., Kowalik SODA'14)

- Given as input $f, g, h : \binom{[n]}{k/3} \to \mathbb{Z}$ with values bounded in bit-length by a polynomial in n, we can in time $O(n^{(1/2-\tau)k+c})$ compute

$$\Delta(f,g,h) = \sum_{\substack{A,B,C \in \binom{[n]}{k/3} \\ A \cap B = \emptyset \\ A \cap C = \emptyset \\ B \cap C = \emptyset}} f(A)g(B)h(C)$$

where

$$\tau = \begin{cases} \frac{(3-\omega)(1-\alpha)}{36-6(1+\omega)(1+\alpha)} & \text{if } \alpha \leq 1/2 \\ \frac{1}{18} & \text{if } \alpha \geq 1/2 \end{cases}$$

Summary
— what next ?

- Graphical models from data — Theme D
  (exact algorithms for polytrees & for Bayesian networks parameterized by treewidth; AAAI'12, AISTATS'13, JMLR 2013)
- Discovering connected motifs in graphs — Theme D
  (linear-time in the size of the host graph; STACS'13)
- Fast "Fourier analysis" on partially ordered sets — Theme F
  (fast Möbius inversion on lattices, counting thin subgraphs; SODA'12 & SODA'14)

Theory and Practice of Advanced Search and Enumeration

(ERC StG 338077 “TAPEASE”)

Petteri Kaski
Department of Information and Computer Science
Aalto University, Helsinki

(1 Feb 2014 – 31 Jan 2019)

Aalto University
School of Science

ALGODAN Algorithmic Data Analysis

Machine Learning for Metabolite Identification

Juho Rousu
Department of Information and Computer Science
Aalto University

Aalto University

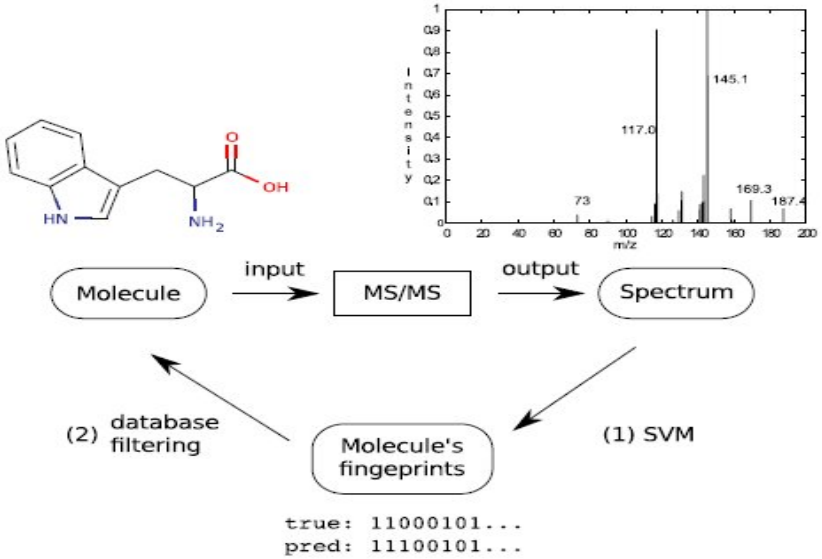
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Metabolite identification (MID)

- Given a biological sample, identification of which molecular species are present is a major bottleneck in metabolomics research
- Tandem Mass Spectrometry (MS/MS) is one of the key measurement techniques, but gives a convoluted signal that requires further processing
- The classical approach to MID is spectral matching:
  - look up a most similar spectrum to the query
  - (only) works well if there is a database spectrum of the same molecule, that was measured with same MS/MS type, similar conditions

Our Approach: Machine Learning

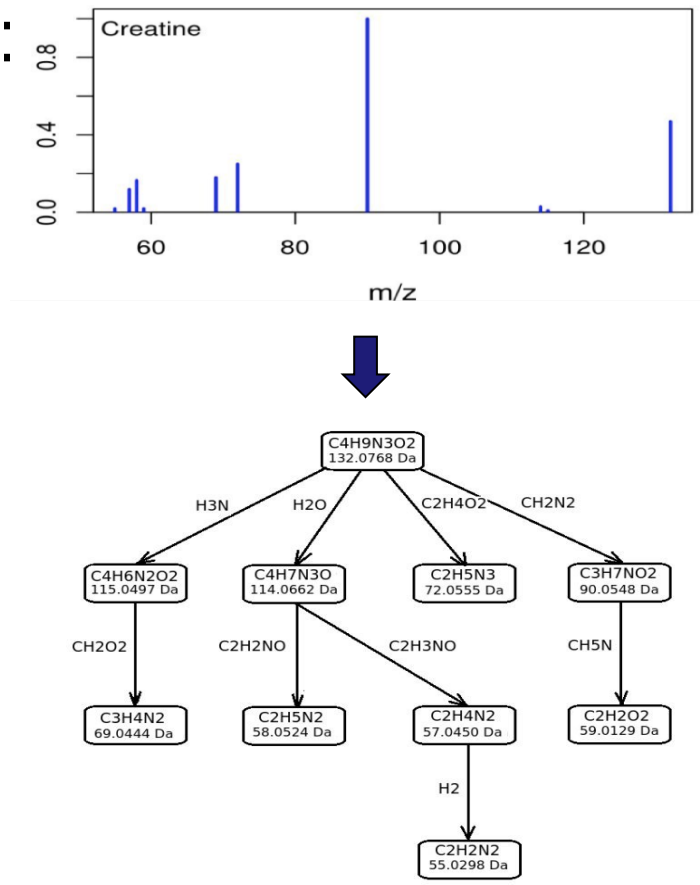
- Basic idea:
  - SVM learning of mappings between MS/MS spectra and molecular features.
  - Retrieve molecules with the predicted features from a large molecular database (PubChem)
- First MID approach using modern machine learning (Heinonen et al. 2012)



Metabolite identification and molecular fingerprint prediction through machine learning (2012). M Heinonen, H Shen, N Zamboni, J Rousu. Bioinformatics 28 (18), 2333-2341

Adding prior knowledge: Fragmentation Trees (FT)

- Fragmentation trees (FT):
  - model the process of breakage of the molecule inside MS/MS
  - Help molecular formulae identification, not full MID
- We use FTs to define **kernels** for MS/MS spectra
- Collaboration with prof. Sebastian Böcker (Jena)



Huibin Shen, Kai Duehrkop, Sebastian Boecker and Juho Rousu. Metabolite Identification through Multiple Kernel Learning on Fragmentation Trees. ISMB 2014, accepted

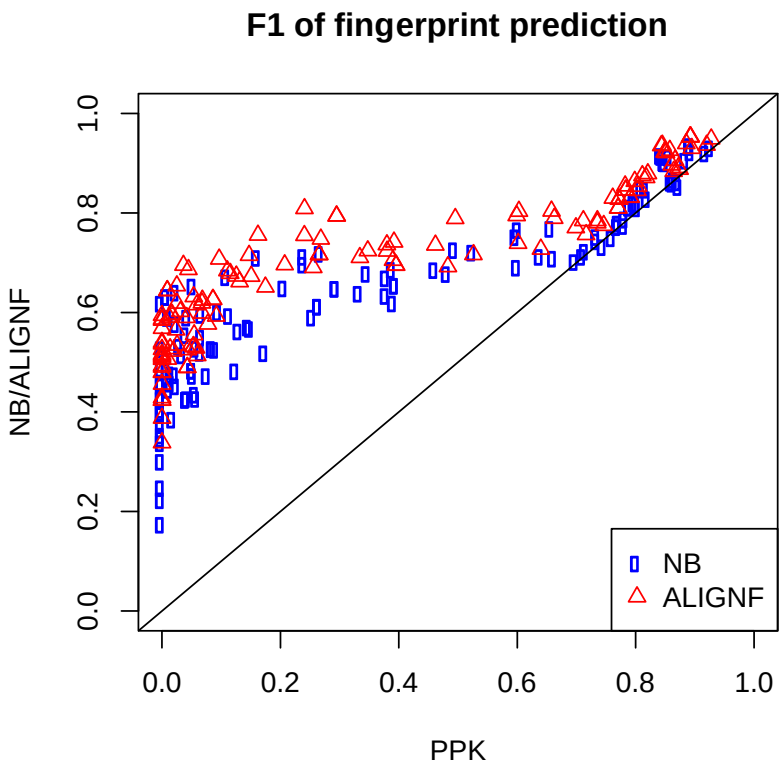
Multiple Kernel Learning

- Kernels between two trees: count co-occurring labeled nodes, edges, paths, subtrees
  - Edges, nodes: inner product of feature maps
  - Paths, subtrees: dynamic programming
- Probability Product Kernel for “raw” MS/MS spectra (Heinonen et al .2012)
- Multiple Kernel Learning: combine the base kernels optimally
  - Uniform combination
  - Lp block-norm approaches
  - **Kernel alignment approaches (Cortes et al. 2012)**
  - Non-linear combination

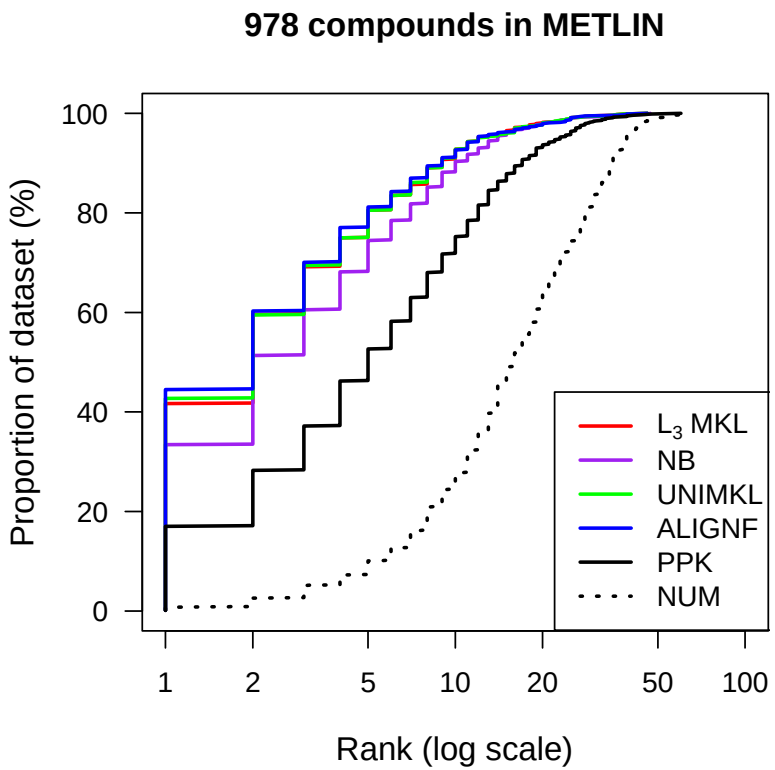
Huibin Shen, Kai Duehrkop, Sebastian Boecker and Juho Rousu. Metabolite Identification through Multiple Kernel Learning on Fragmentation Trees. ISMB 2014, accepted

Prediction performance

Molecular fingerprints



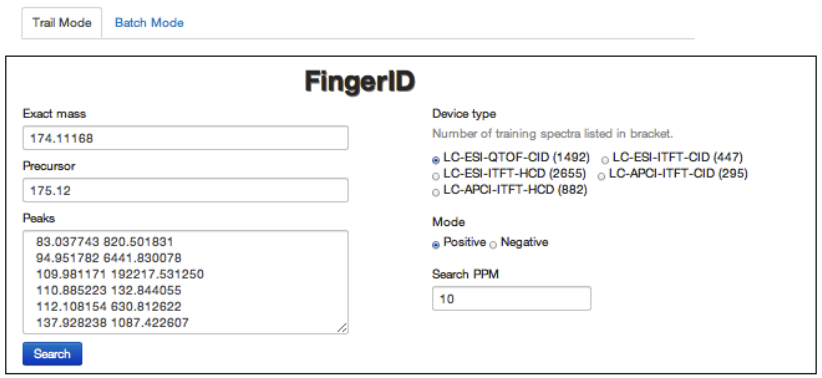
Metabolite Identification



Huibin Shen, Kai Duehrkop, Sebastian Boecker and Juho Rousu. Metabolite Identification through Multiple Kernel Learning on Fragmentation Trees. ISMB 2014, accepted

MID Software

- Our algorithms are available for research community
  - Software package (source forge)
- FingerID tool
  - Easy user interface for metabolomics researchers
  - http://research.ics.aalto.fi/kepaco/fingerid/



> > >

Score	Name	Formula	Exact Mass	Structure	Database ID
0.833	Amino acid(Arg-); Arginine; 2-Amino-5-guanidinovaleric acid	C6H14N4O2	174.1117	C02385	C02385
0.359	D-Arginine; D-2-Amino-5-guanidinovaleric acid	C6H14N4O2	174.1117	C00792	C00792
0.359	L-Arginine; (S)-2-Amino-5-guanidinovaleric acid; L-Arg	C6H14N4O2	174.1117		C00062

Perspectives on Metabolite Identification

- Current results are state-of-the-art of automatic MID
- Google search analogy: 80% of correct molecular structures within top 5 candidates – already useful in practise!
- Current and future themes:
  - Multilabel and structured output prediction methods for MID
  - Joint identification of metabolites from metabolomics samples
- Supported by Academy of Finland grant “Metabolite Identification through Algorithms and Statistical Learning (MIDAS)”, 2013-2017

Activities of KEPACO group: see posters

- More MID:
  - Huibin Shen: Metabolic Identification through Multiple Kernel Learning on Fragmentation Trees (ISMB 2014, to appear)
- Things I did not tell you about:
  - Hongyu Su: Random graph ensembles for multilabel learning (ACML 2013)
  - Jana Kludas: Protein-protein interactions in the protein transport pathways (BIOLEDGE, FP7 STREP)
  - Anna Cichonska: Predicting Drug-Target interactions through KroneckerRLS learning (Collaboration with Finnish Institute for Molecular Medicine, FIMM)