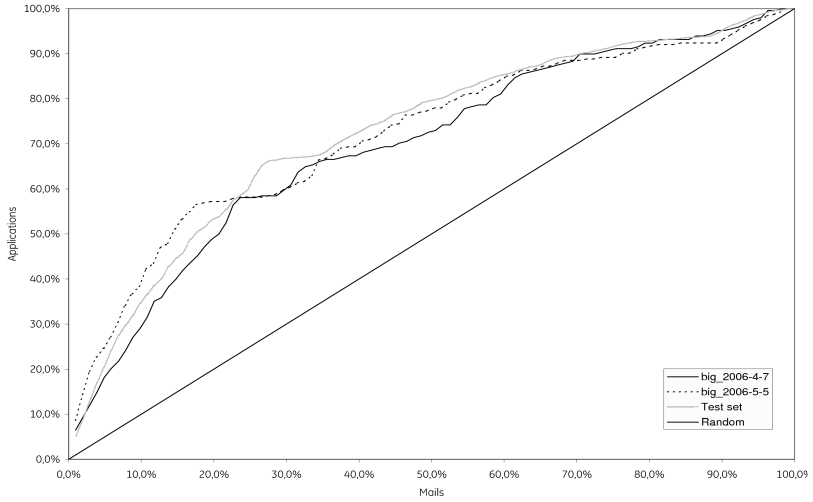
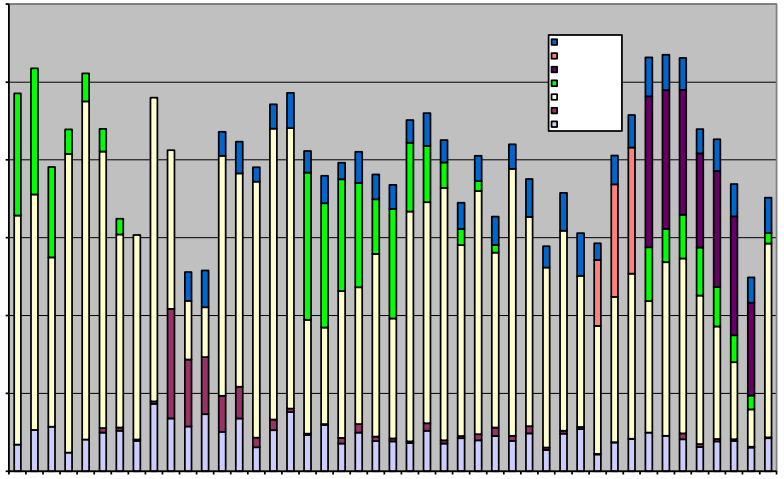
Open Source Data Mining mit WEKA





Dr. Alexander K. Seewald

SEEWALD SOLUTIONS

Was ist Data Mining?

DATA MINING

"Data Mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."

(Fayyad, Piatetsky-Shapiro & Smyth, 1996)

MACHINE LEARNING

"The field of machine learning is concerned with the questions of how to construct computer programs that automatically improve with experience."

(Tom M. Mitchell, 1997)

SEEWALD SOLUTIONS
alex@seewald.at
http://www.seewald.at

Was ist WEKA? (1)

Waikato **E**nvironment for **K**nowledge **A**nalysis

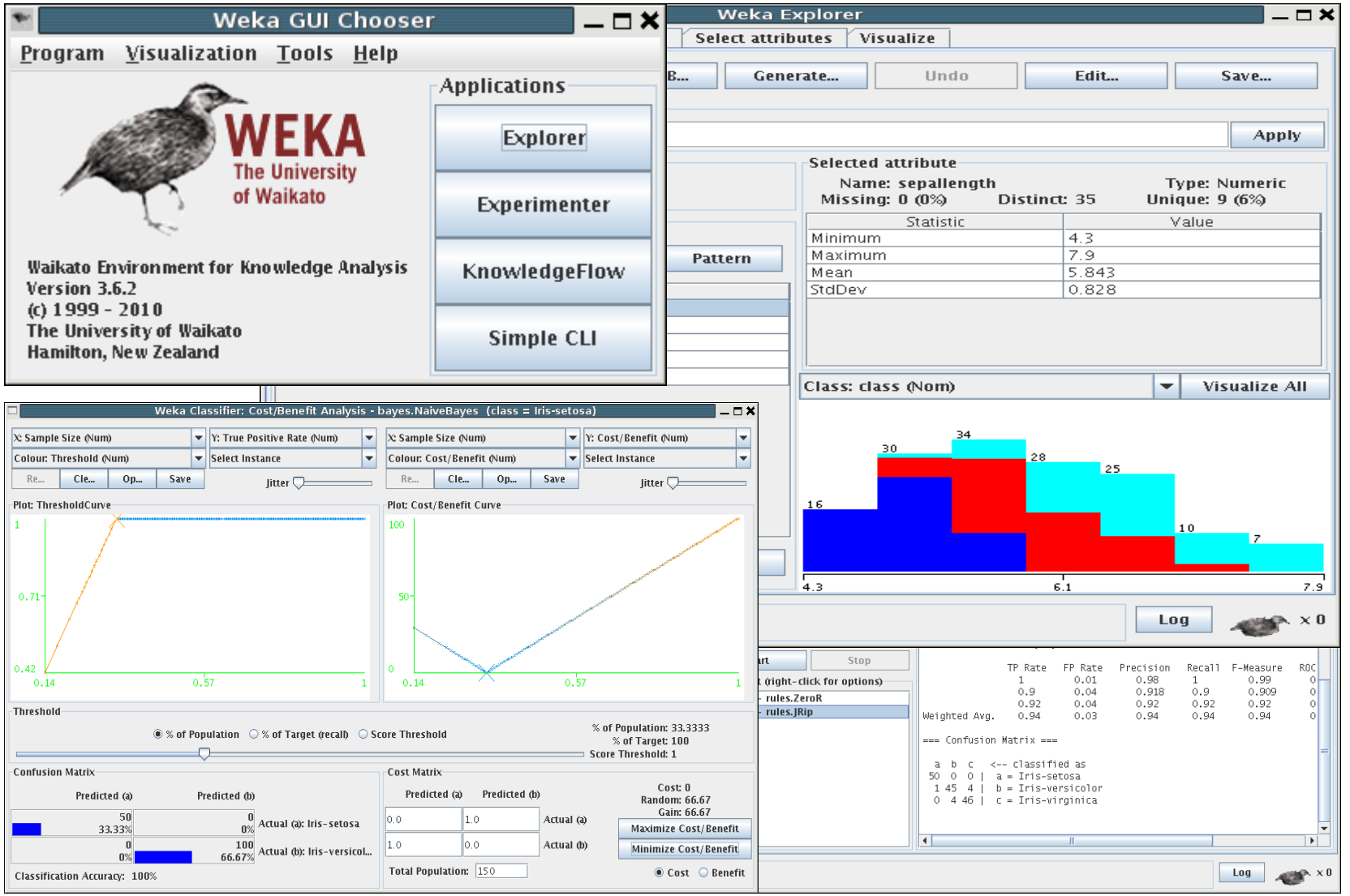
Die weitverbreiteste Data Mining Suite, für Anwendung, Lehre und Forschung

http://www.cs.waikato.ac.nz/~ml/weka

- Stabilität, Verfügbarkeit und Qualität der Lernalgorithmen noch immer weit jenseits von kommerziell verfügbaren Tools.
- 1000+ Contributers seit 1999, GPL, vielfach ausgezeichnet (InfoWorld 2007)
- Benannt nach einem neugierigen flügellosen Vogel, der in Neuseeland heimisch ist und unter Naturschutz steht

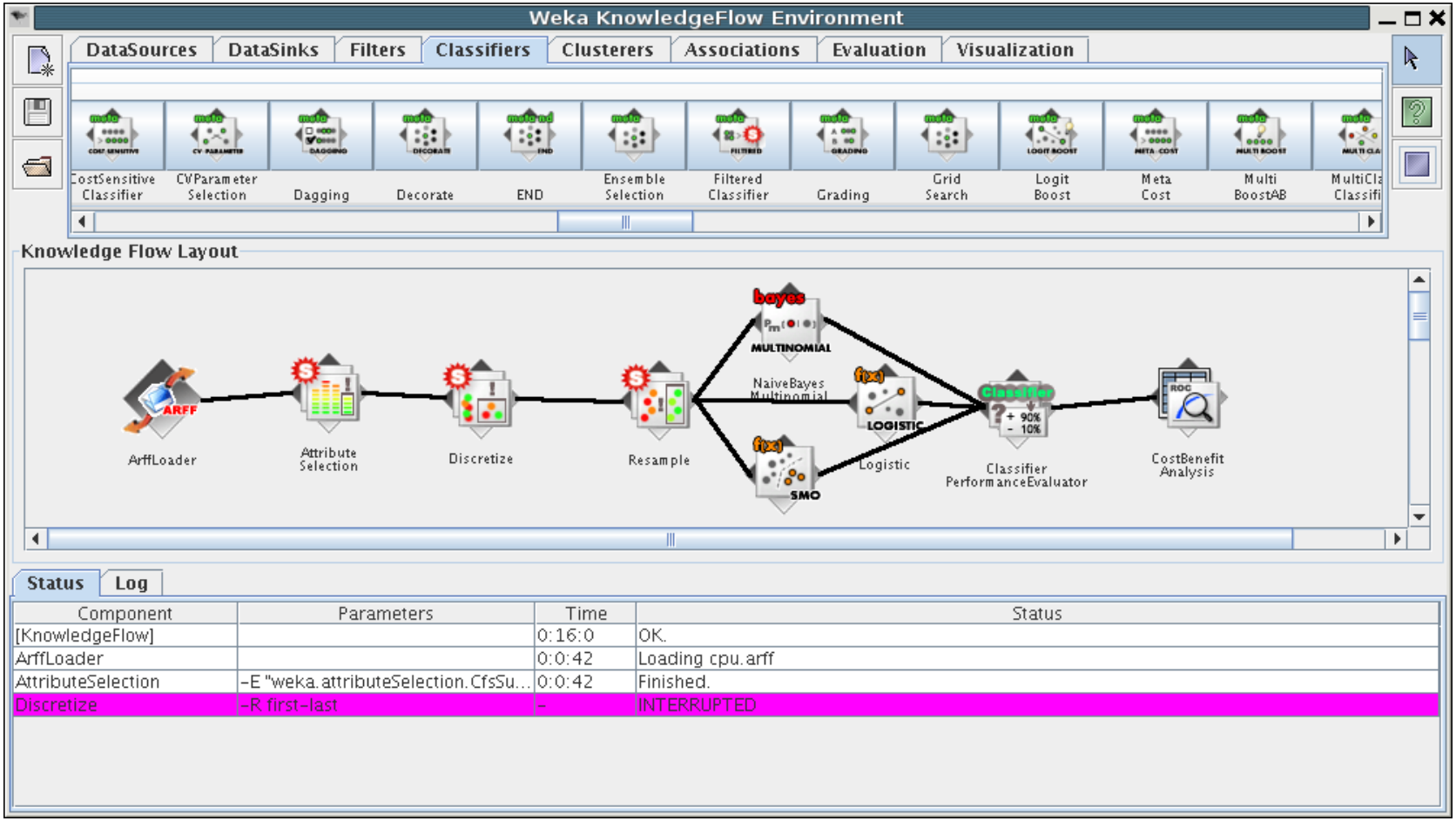
alex@seewald.at
http://www.seewald.at

Was ist WEKA? (2)



alex@seewald.at
http://www.seewald.at

Was ist WEKA? (3)



alex@seewald.at
http://www.seewald.at

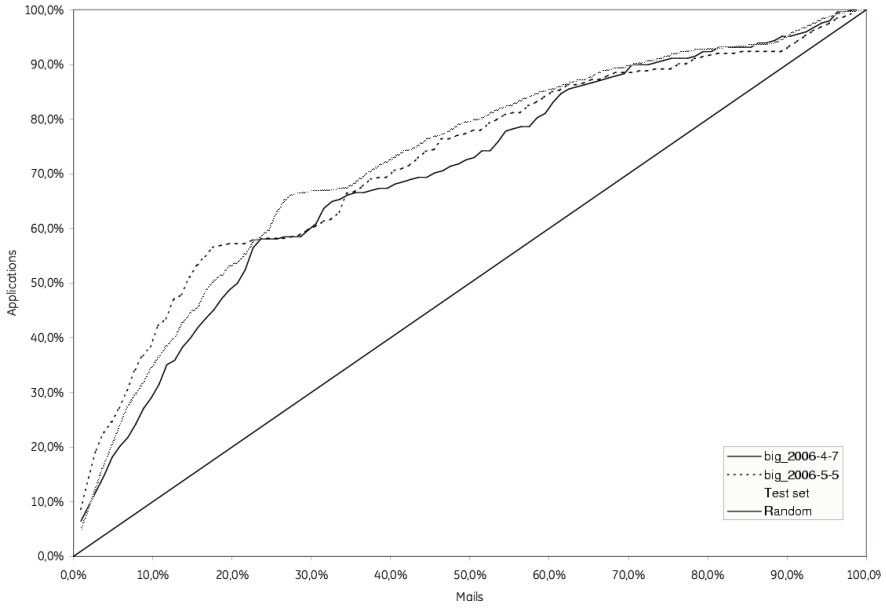
Übersicht

- **Marketing-Effizienz erhöhen, Banken**
- **Validierung Marketingmaßnahmen, Banken**
- **BioMinT – Biologisches Text-Mining**
- **Ein Frühwarnsystem für Bot-Netze**
- **IGO 2 – Image-Mining mit WEKA**
- **Watching C. elegans Think**

SEEWALD SOLUTIONS
alex@seewald.at
http://www.seewald.at

Marketing-Effizienz erhöhen (1)

- **Problem:** Nicht genug Kapazität, um alle Kunden prer Post Info-Mail anzuschreiben
- **Lösung:** Erhöhung der Effizienz mittels eines gelernten Rücklauf-Modells (Response Model)



White Paper

Seewald A.K.: Improving the Effectiveness of Mailings by Building a Response Model for Inactive Customers. Technical Report, Seewald Solutions, Wien, 2007.

publications.seewald.at

SEEWALD SOLUTIONS

alex@seewald.at
http://www.seewald.at

Marketing-Effizienz erhöhen (2)

Trained a response model for inactive customers, based on historical data (07/2005 – 03/2006). Trained to determine customers who apply for a loan.

Data: About 300,000 past responses — about 1% are positive, 99% negative.

Training = Downsampling to 1:1 class distribution (50% of positive, 0.5% of negative responses)

Testing = Rest of the data (50% of positive, 99.5% of negative responses)

Additionally, tested on two recent inactive customer mailings in April and May 2006.

Using NaiveBayes-derived classifier HNB on a subset of 74 partner, contracts and mailing-based features. Feature subset was chosen by extensive feature subset selection using this classifier.

HNB estimates the propabilities of attribute values, given the class and a weighted sum of dependent attribute probabilities, from training data.

$$P(f|TD) = \frac{P(TD|f)P(f)}{P(TD)}$$

alex@seewald.at
http://www.seewald.at

Marketing-Effizienz erhöhen (3)

Final Input Features

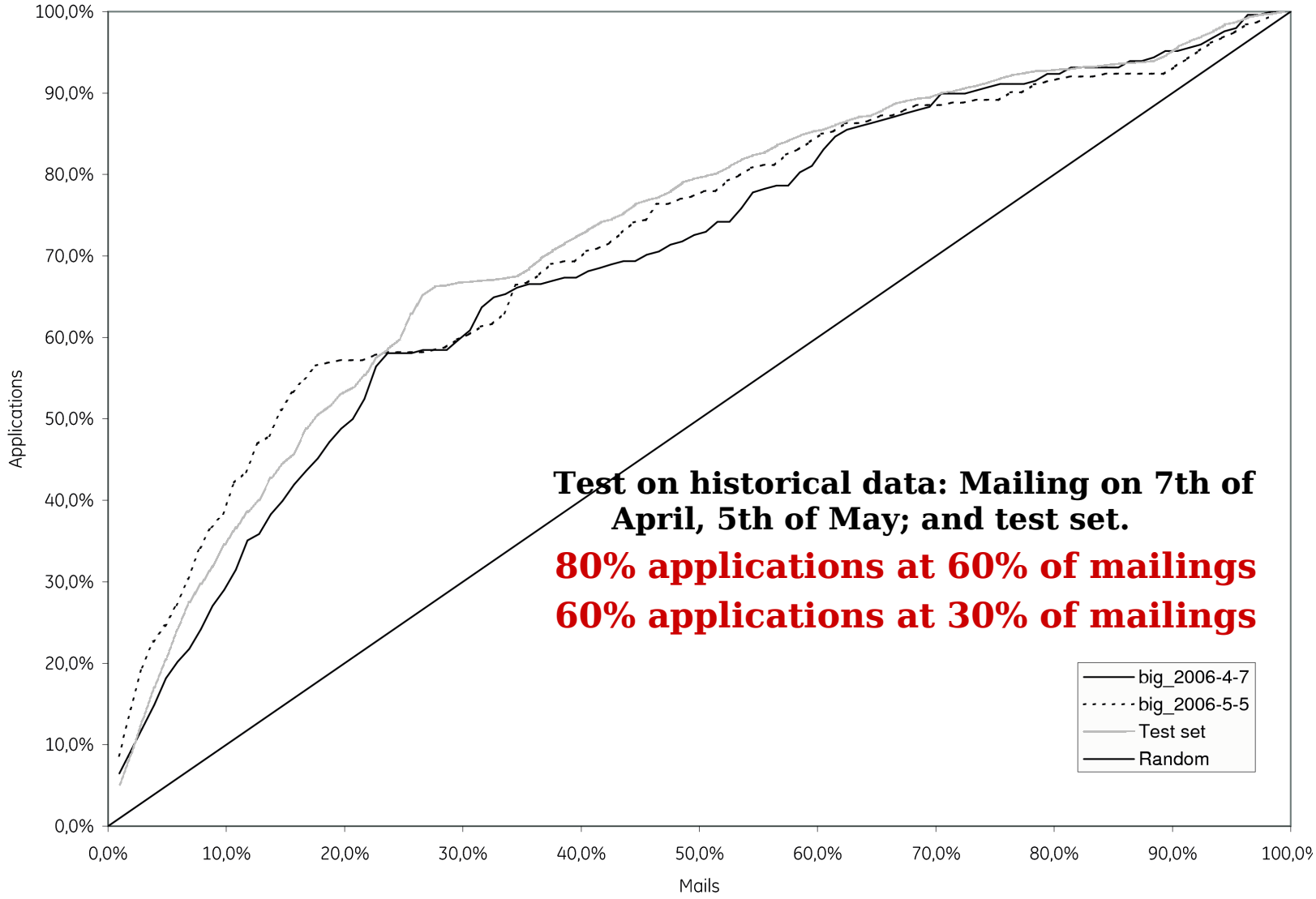
• totalPastAppl	No. of applications for CL F2F last 180 days
• Dependant_partner	Spouse with / without income
• No_accounts	Total number of past contracts
• Worst_payment	Worst paying_score on all contracts
• No_deferrals_not_liquidated	number of deferrals on all active contracts
• Industry	"Hauptbranche"
• Net_Income	Latest net income of customer
• Written_Prove_Salary_Available	Net income is proven by written receipt
• Tel_Type	Type of telephone (fixed-line, mobile phone)
• Reminder_Status	"Mahnstatus"
• MOB	months on book, from most current contract
• Loantermcov	MOB/contract_term
• MeanOverpayment13	mean overpayment of last three months

Target Variable

- At least one application within 60 days of mailing send-out date (similar to Marketing report)

SEEWALD SOLUTIONS
alex@seewald.at
http://www.seewald.at

Marketing-Effizienz erhöhen (4)

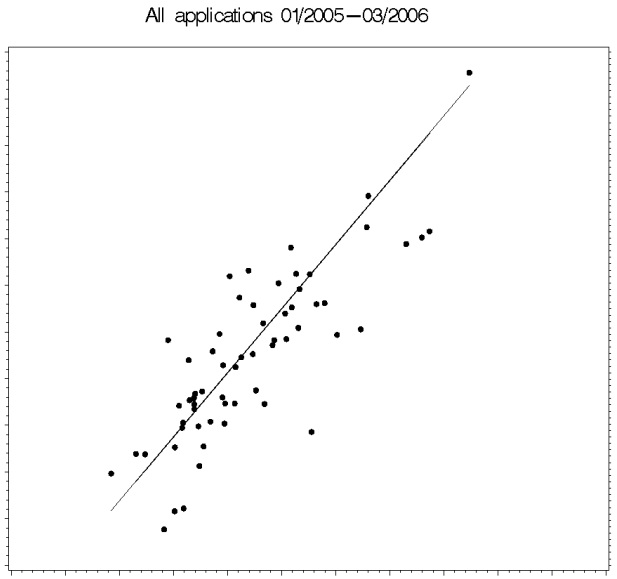


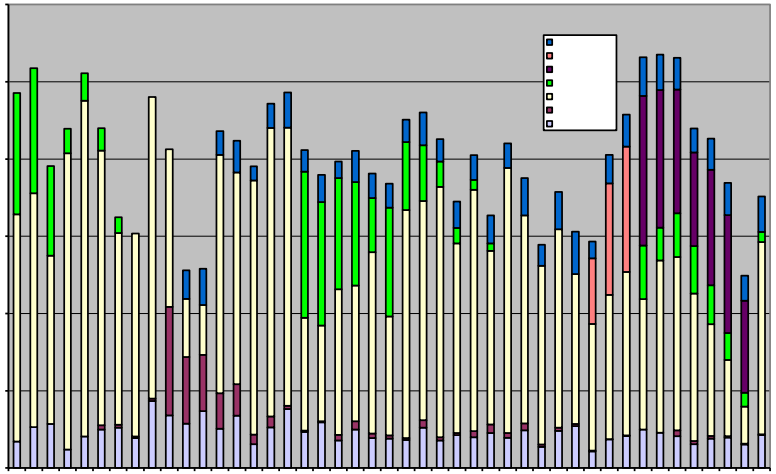
SEEWALD SOLUTIONS

alex@seewald.at
http://www.seewald.at

Validierung Marketing-Maßnahmen (1)

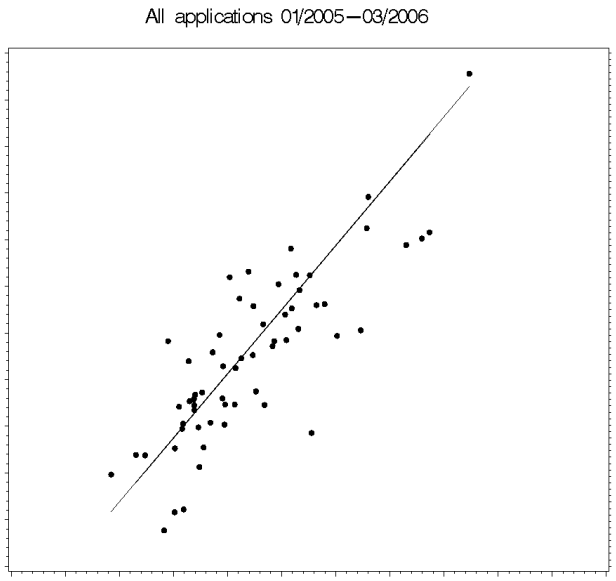
- **Problem:** Uneinheitliches Reporting – keine definitive Effektgröße pro Marketing-Maßnahme
- **Lösung:** Input/Output-Analyse aller Marketing-Maßnahmen über ein Jahr





alex@seewald.at
http://www.seewald.at

Validierung Marketing-Maßnahmen (2)



model=0.0028*Big
+0.0177*Pre
-0.0339*Liq
+0.0299*Lza
+578.6685

To cross-check mailing performance, we determined a model of applications vs. sent mailings. This was based on the following assumptions:

1. Mailings have the largest effect in the week after they are sent out. This effect decreases geometrically by a factor of 1.5 per week for 6 weeks, after which it can be neglected.

2. Each mailing type has an initial effect which is linearly proportional to the number of mails sent out.

SEEWALD SOLUTIONS

alex@seewald.at
http://www.seewald.at

Validierung Marketing-Maßnahmen (3)

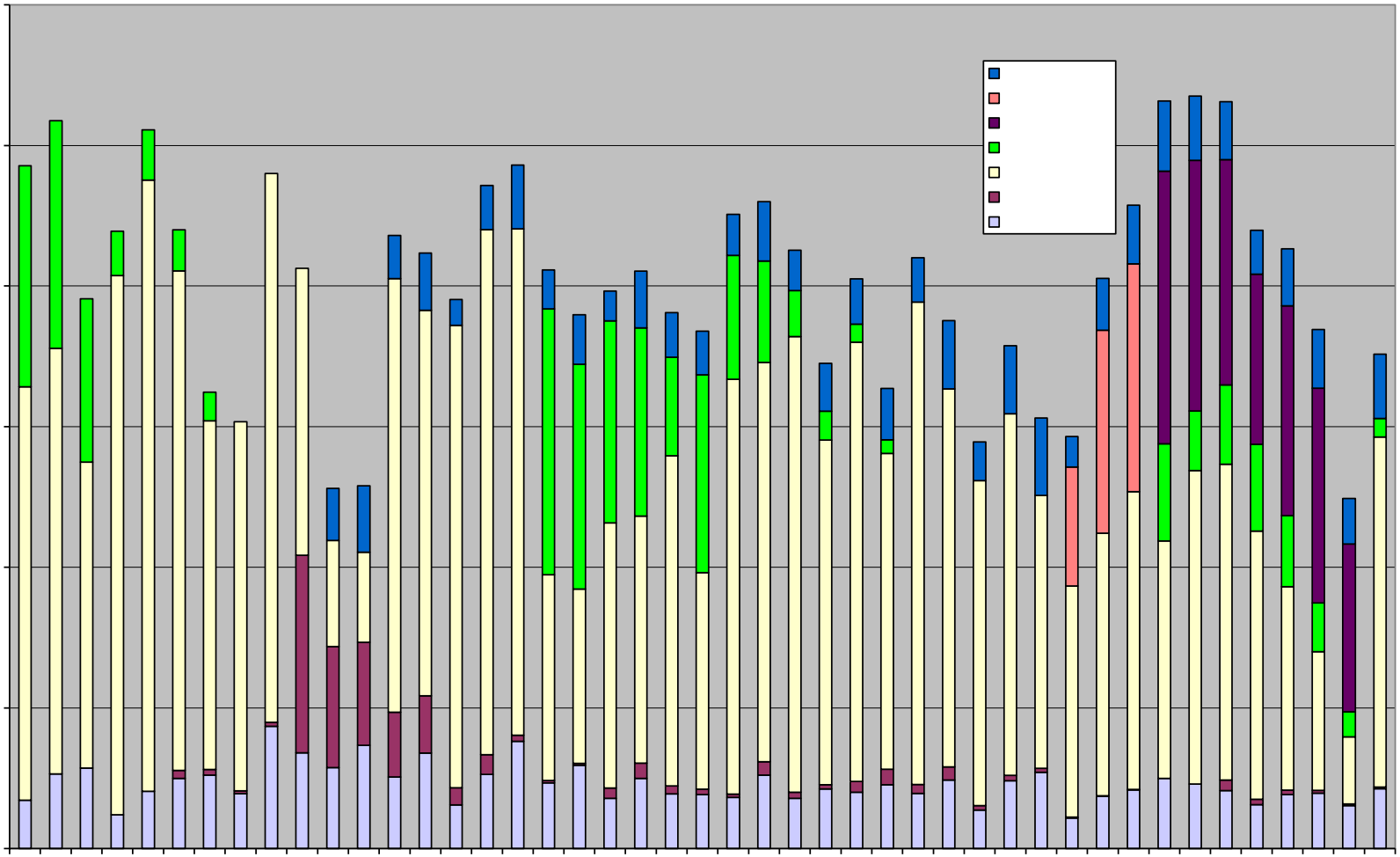
Extending the Mailing Response Model

- Integrates all Marketing activities (billboards, radio, print, branch changes, „postwurf“, mailings)
- Models distribution of activities (inputs) and applications (outputs) spatially by postal district and temporally by week – a spatiotemporal model.
- Assumes mostly linear effects depending only on marketing activity type (except mailing response)

Determine effectiveness of marketing activities in a straight-forward, quantitative manner.

alex@seewald.at
http://www.seewald.at

Validierung Marketing-Maßnahmen (4)



SEEWALD SOLUTIONS

alex@seewald.at
http://www.seewald.at

BioMinT: Biological Text Mining (1)

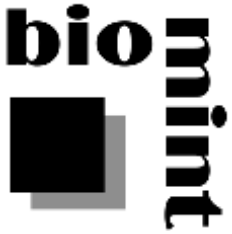
- **Problem:** Forscher verwenden extrem viel Zeit auf das Updaten von Online-Protein-Datenbanken.
- **Lösung:** Verringerung des Aufwands durch Erstellung eines webbasierten Systems, das alle wichtigen Schritte laufend unterstützt.

<u>Research project funded by the EU (2003 – 2005)</u>

- Generic text mining tool for content-based and knowledge-intensive information retrieval and extraction
- Applied to the annotation of the Swiss-Prot and PRINTS proteomics databases with information mined from scientific papers; and to build human-readable reports

In-silico research/curator assistant

biomint.pharmadm.com



SEEWALD SOLUTIONS

alex@seewald.at
http://www.seewald.at

BioMinT: The BioMinT Tool (2)

General workflow

1. User enters protein / gene name
2. Name is looked up in comprehensive Gene and Protein Synonym Database (GPSDB). Selection criteria: species, taxonomic range, source database and source field.

   This expands Name with (almost) all known synonyms.
3. Generate & execute PubMed query with all synonyms.
4. **Retrieve references, filter and rank by relevance.**
5. **Extract information for annotation purposes (PRINTS,SP)**

SEEWALD SOLUTIONS
alex@seewald.at
http://www.seewald.at

BioMinT: Species from MEDLINE (3)

Predict the species of an organism from MEDLINE

- 19.0% Baseline (most common class *Human*)
- 26.5% Rule based on single word *Fungal* (WEKA)
- **75.5% Human domain expert's rules**
- 76.4% NaiveBayes (WEKA)
- 79.6% Mapping MeSH Terms to species
- 88.9% JRip Rule Learner, 172 rules (WEKA)
- **89.3% Support Vector Machine (SMO, Weka)**

Comparing JRip rules to domain expert rules

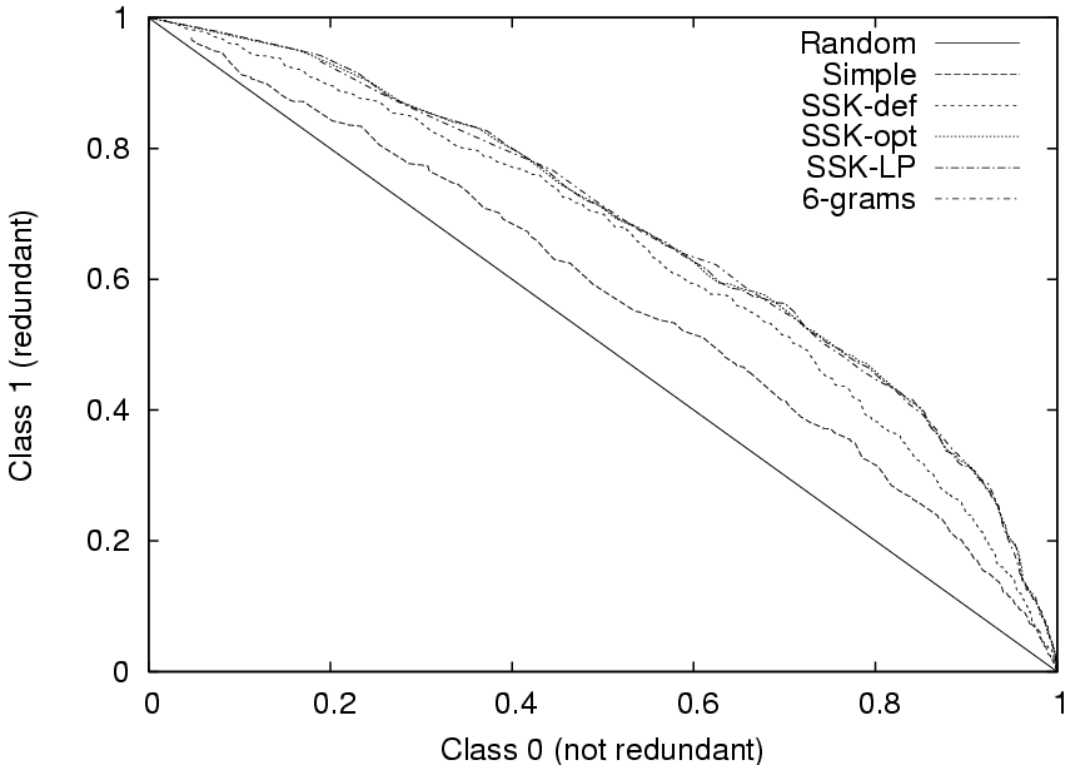
- Expert: + precision, – recall; – – F-Measure
- JRip: – precision, + recall; ++ F-measure

alex@seewald.at
http://www.seewald.at

BioMinT: Redundancy Recognition (4)

- For purposes of automated Information Extraction, sentence classification models were created. To summarize the output, we investigated redundancy recognition via String Subsequence Kernels.

Kernels were contributed to WEKA, see [Seewald&Kleedorfer, 2007].



alex@seewald.at
http://www.seewald.at

Ein Frühwarnsystem für Bot-Netze (1)

- **Problem:** Spam wird von Bot-Netzen ausgesendet, deren Lebenszyklus noch kaum erforscht ist.
- **Lösung:** Rein passives Verfolgen von Bot-Netzen durch Darknets zur Identifizierung/Früherkennung

<u>Forschungsprojekt im Bereich IT Security (2008)</u>

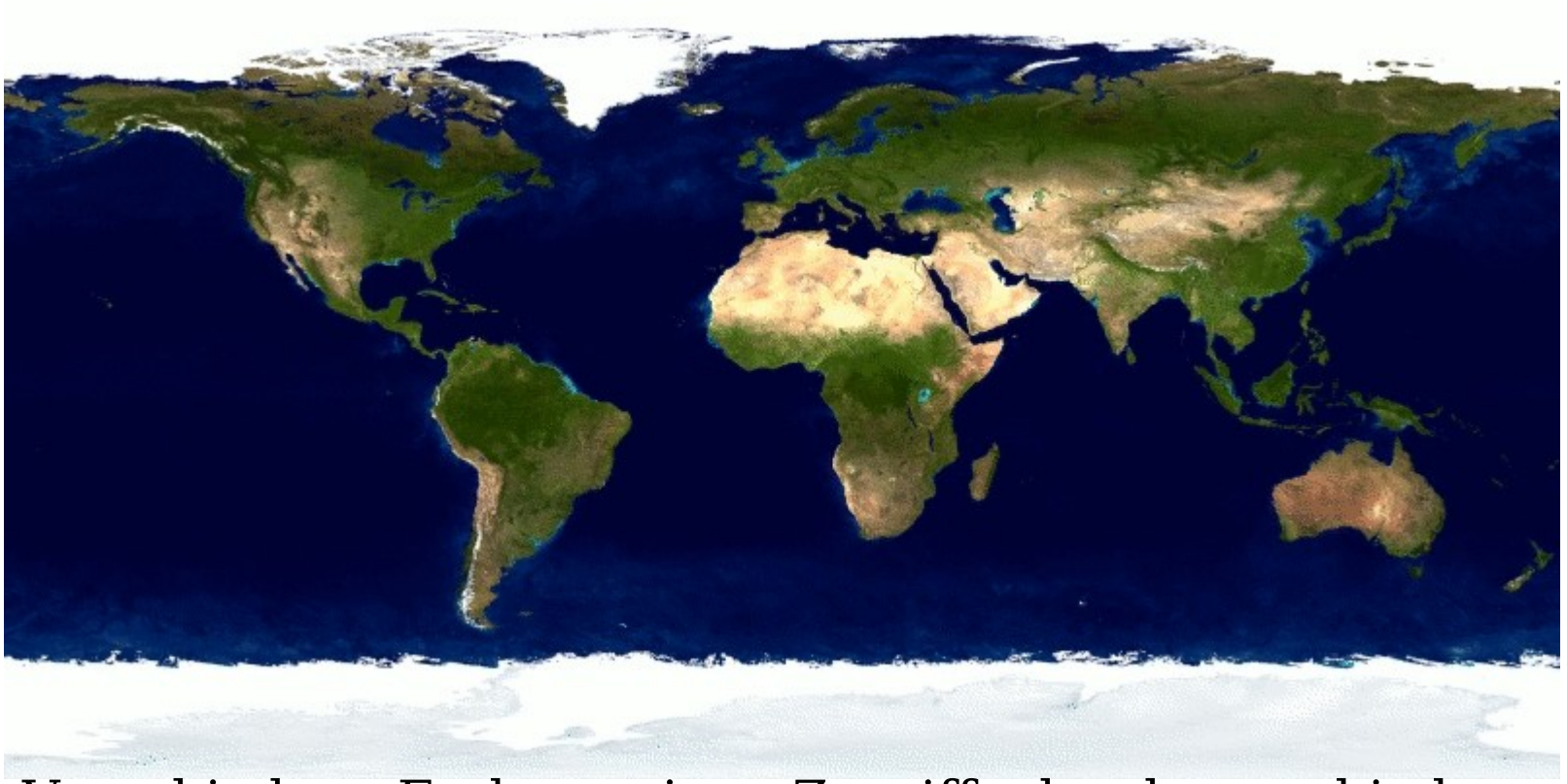
- Referenzdaten zu bekannten Bots- und Bot-Netzen
- Trainieren von Lernmodellen zur Erkennung des TCP/IP-Traffic eines bestimmten Bots
- Validierung und Test

Basiert vollständig auf Open-Source Software; alle Lernmodelle & Vorverarbeitung in WEKA.

Top downloaded journal paper in Q4/2009

SEEWALD SOLUTIONS
alex@seewald.at
http://www.seewald.at

Ein Frühwarnsystem für Bot-Netze (2)



Verschiedene Farben zeigen Zugriffe durch verschiedene Spambots an. GPL code:
http://botnetz-tracker.seewald.at/

Hintergrund: Visible Earth (NASA), IP-Positionsbestimmung durch IP Address Location.
Spambot Trainingsdaten zur Verfügung gestellt von Marshal Trace.

alex@seewald.at
http://www.seewald.at

Image-Mining mit WEKA

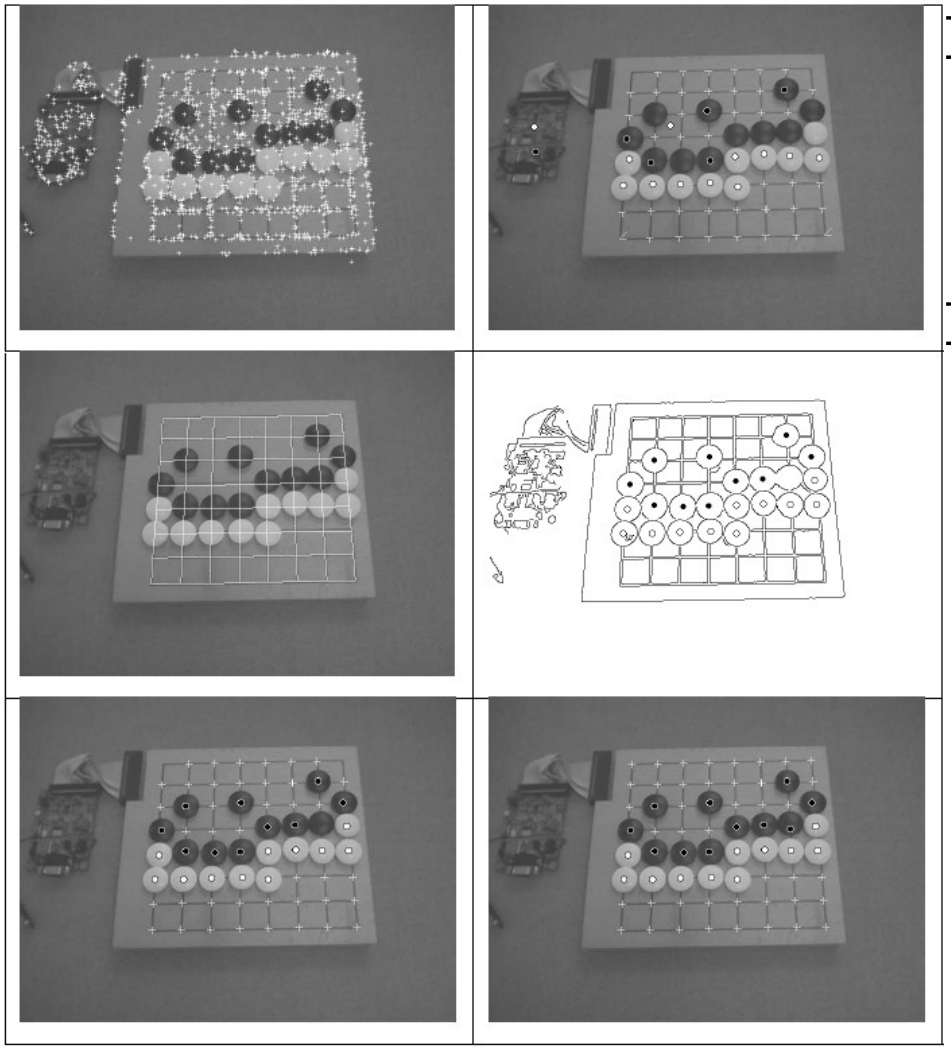


Figure 1: Steps 1-6 with sample images after each step, left-to-right, top-to-bottom.

Problem: Go-Spieler haben keine Zeit, die eigenen Spiele mitzuschreiben.

Lösung: Automatische laufende Erkennung der Spielposition über Handy-Bilder.

- Pro Einzelbild 98.4% genau, 4/6 Schritte verwenden WEKA [Seewald, 2010]

alex@seewald.at
http://www.seewald.at

Watching C. Elegans Think (1)

- **Problem:** Bestehende Lernalgorithmen sind gegenüber den Lernfähigkeiten von Tieren und Menschen meistens nicht konkurrenzfähig.
- **Lösung:** Wir nehmen uns ein Vorbild an der Natur.

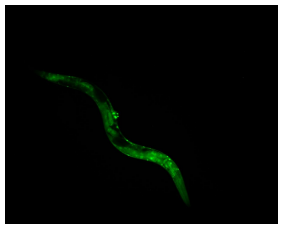
Four Objectives

• Engineering	*Real-time tracking nerve cells*
• Methodological	*Validate nervous cell models*
• Holistic	*Understand complete N.S.*
• Insight	*Better learning algorithms*

Model organism: C. elegans

~ 1000 cells, ~ 300 nerve cells

Might be feasible to simulate



alex@seewald.at
http://www.seewald.at

Watching C. Elegans Think (2)

Results of an automated analysis of C.Elegans images (data by Prof. T. Johnson's group)

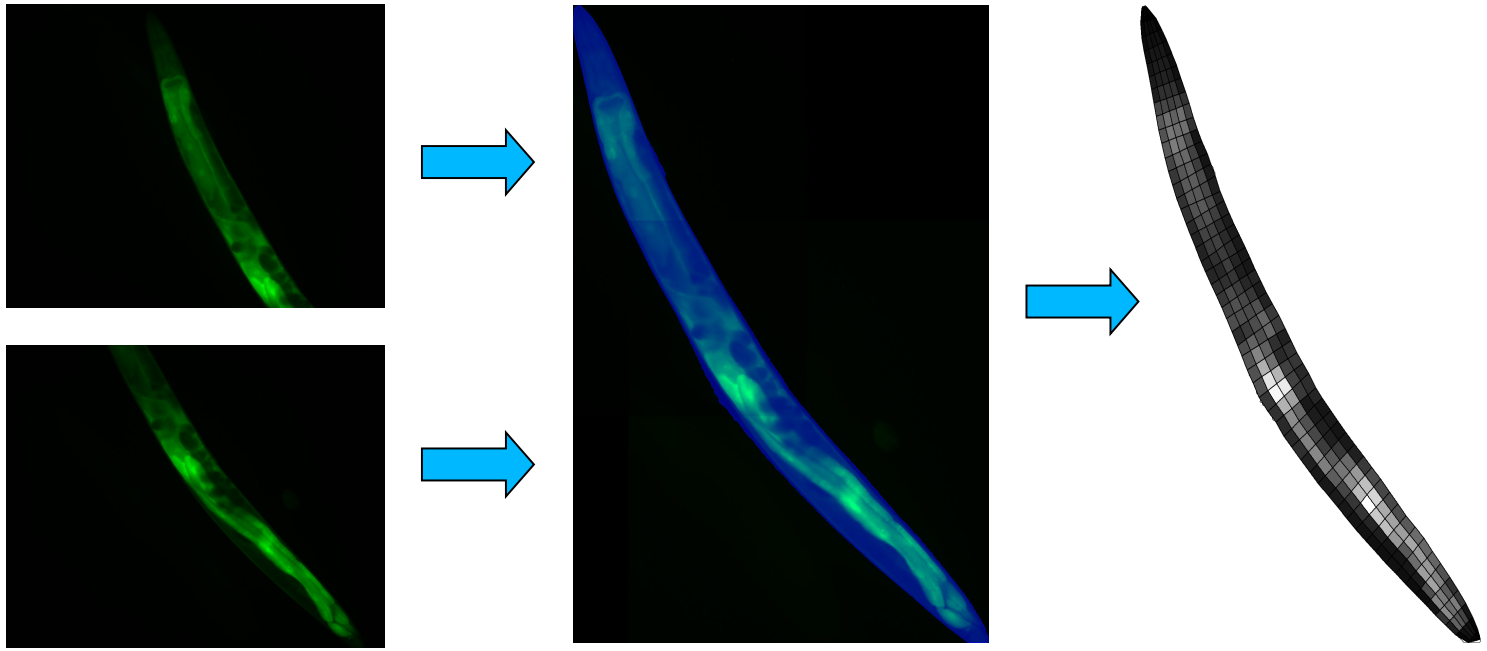


Image processing done via ImageIJ & WEKA

Reduces workload by 80% (paper pending)

Details & GPL v3 code: http://elegans.seewald.at/

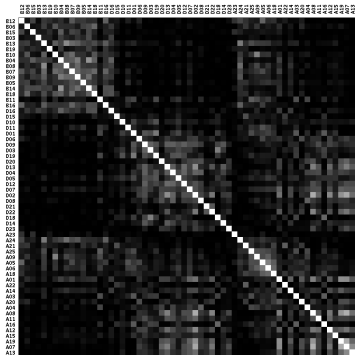
alex@seewald.at
http://www.seewald.at

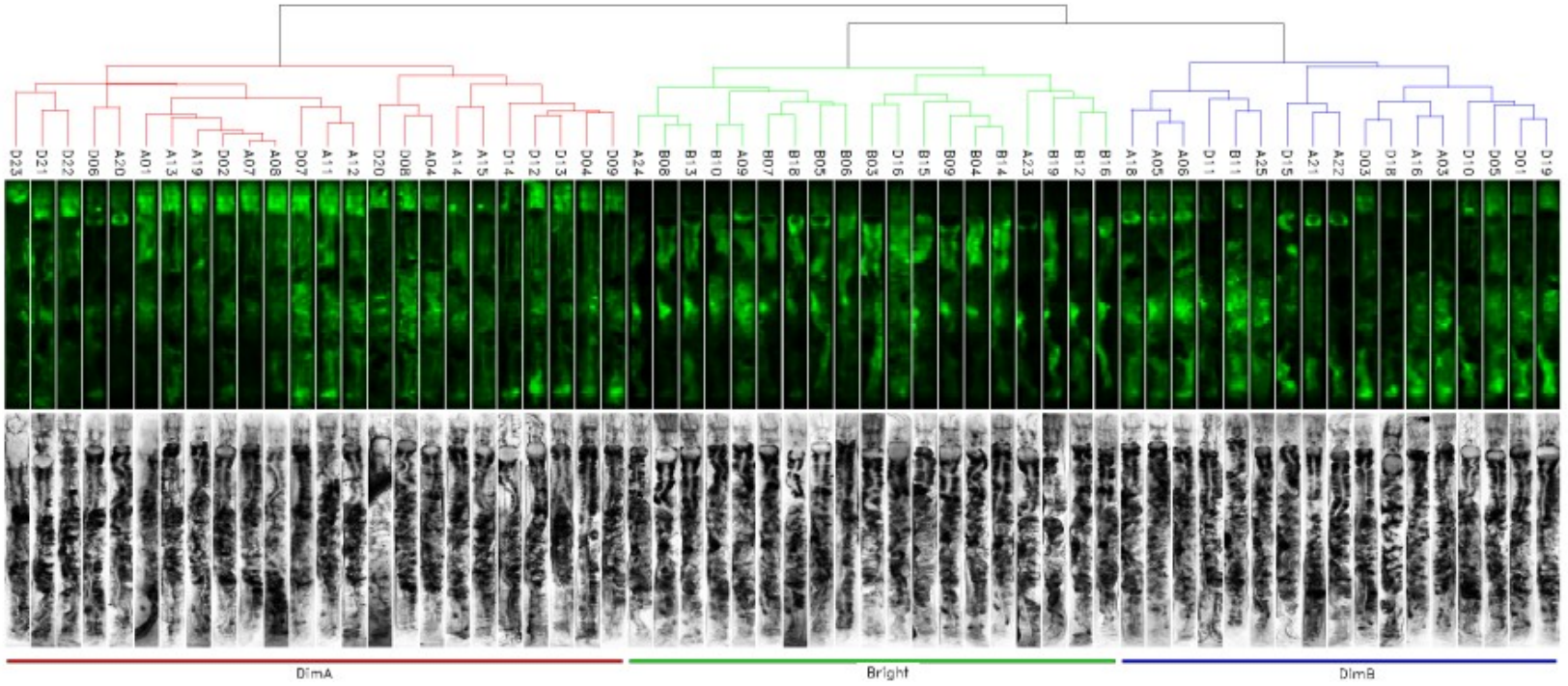
Watching C. Elegans Think (3)

Some interesting results:

Bright worms live longer than dim worms.

Even when discounting brightness, bright worms show distinct expression patterns.





SEEWALD SOLUTIONS

alex@seewald.at
http://www.seewald.at

Vielen Dank für die Aufmerksamkeit!

Für Fragen stehe ich jederzeit gerne zu Ihrer Verfügung.

SEEWALD SOLUTIONS

alex@seewald.at
http://www.seewald.at