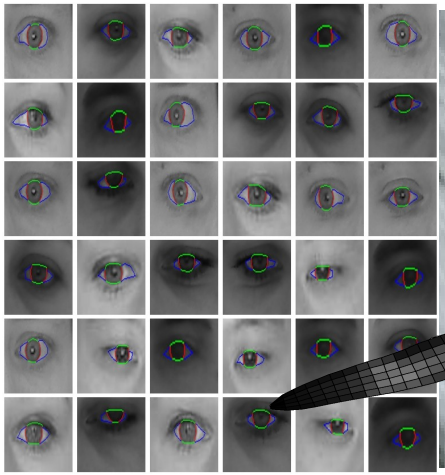
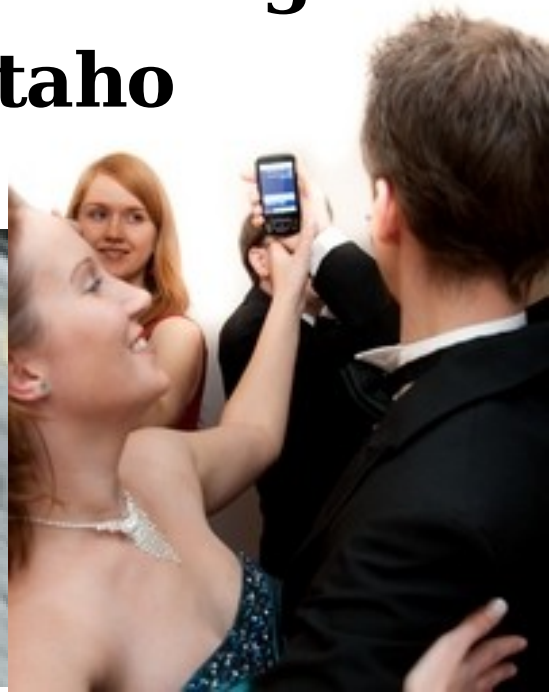
Open Source Data Mining mit WEKA/Pentaho





Dr. Alexander K. Seewald

SEEWALD SOLUTIONS

Was ist WEKA? (1)

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

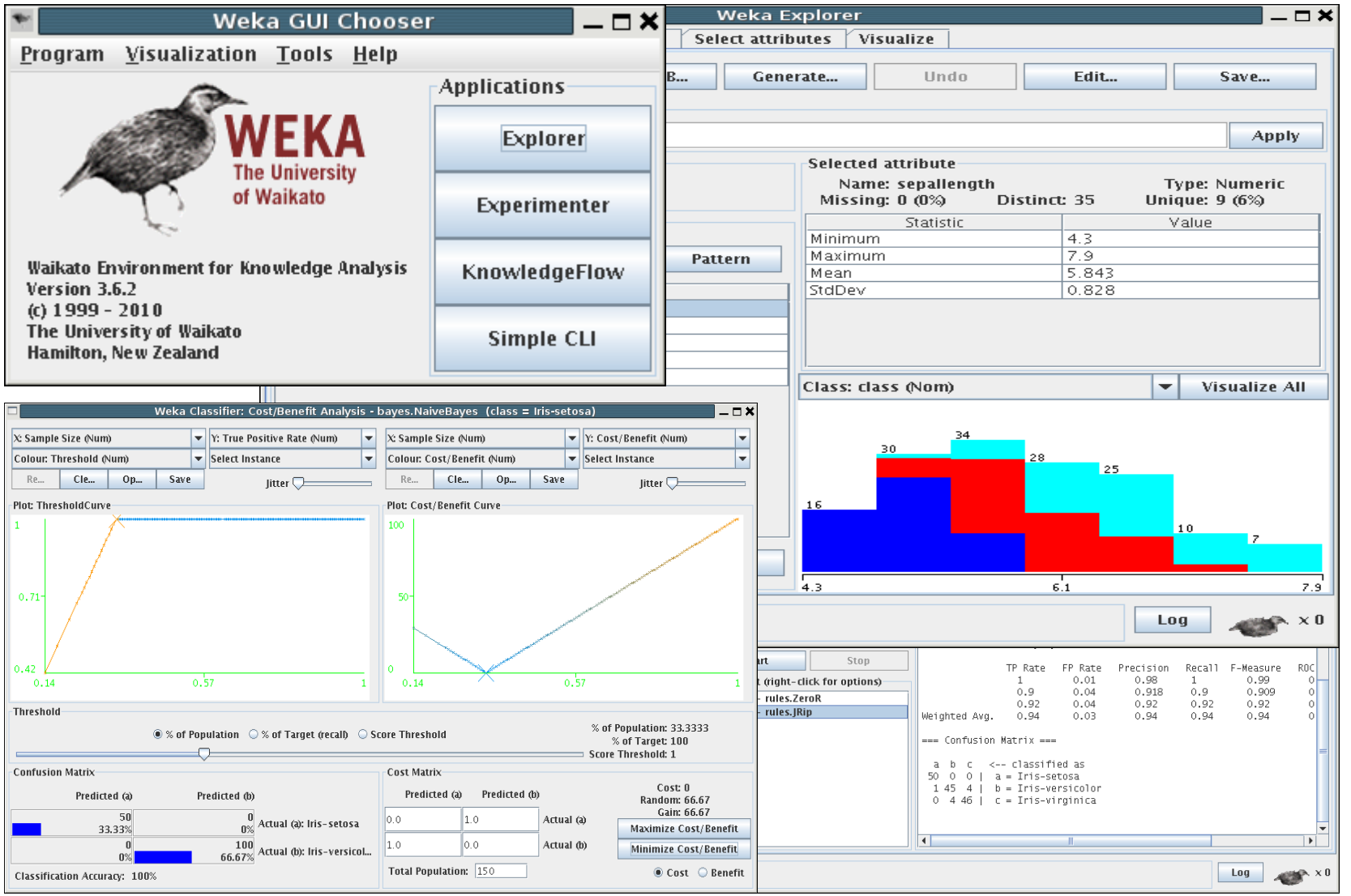
- Benannt nach einem neugierigen flügellosen Vogel, der in Neuseeland heimisch ist und unter Naturschutz steht
- 1000+ Contributers seit 1999, GPL
- Stabilität, Verfügbarkeit und Qualität der Lernalgorithmen weit jenseits von kommerziell verfügbaren Tools
- Sponsor: Pentaho Corp. – in den Top Ten wichtigsten OS Projekten 2007,Infoworld

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

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

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

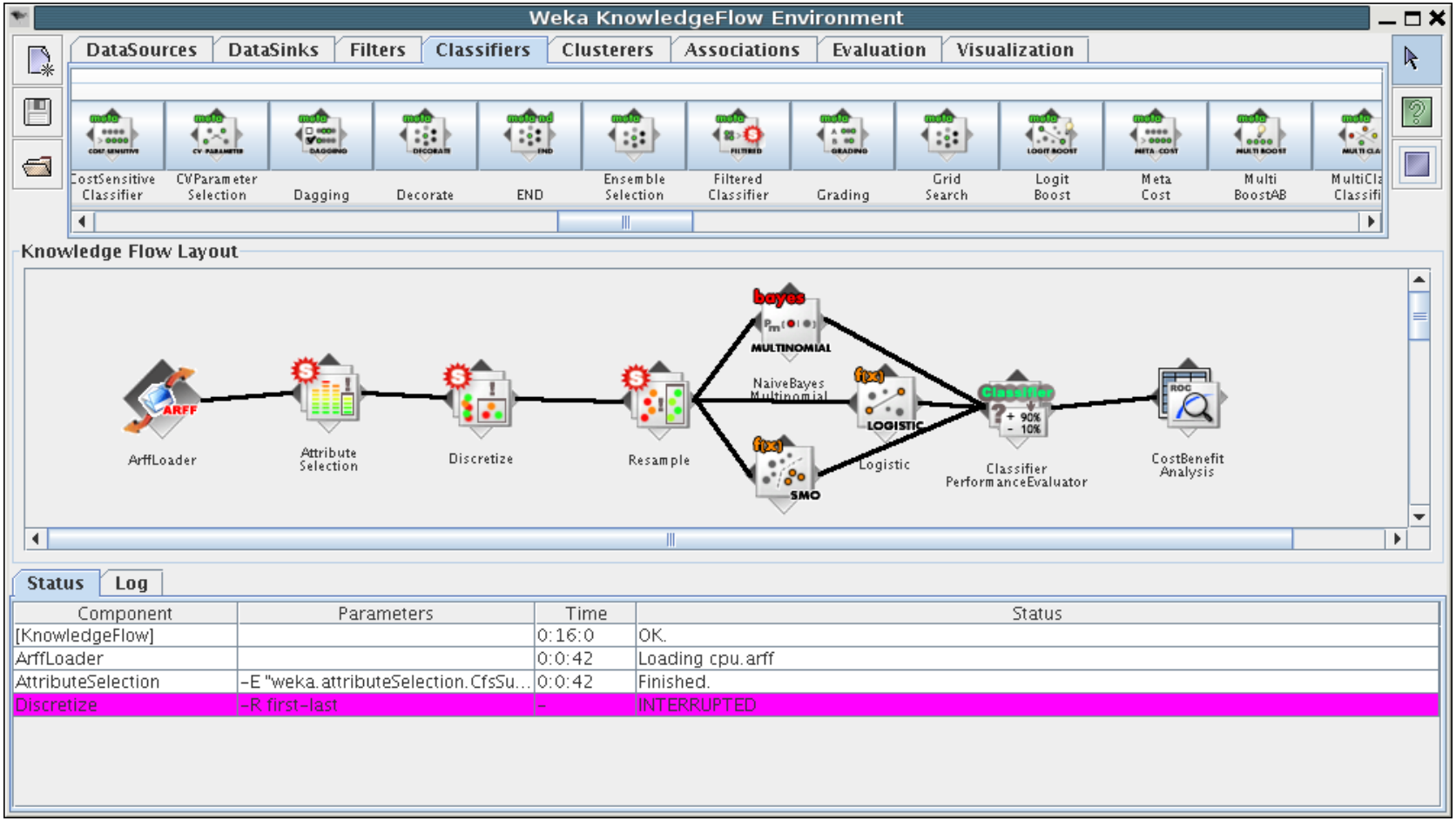
Was ist WEKA? (2)



SEEWALD SOLUTIONS

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

Was ist WEKA? (3)



SEEWALD SOLUTIONS

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

Übersicht

- **Best Moves – Dancing Guide***
- **Ein Frühwarnsystem für Bot-Netze°**
- **Low-cost Eyetracking***
- **Watching C. elegans Think°**
- **deFlicker***

** = Video, ° = kompletter Code unter GPL verfügbar*

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

Best Moves – Dancing Guide (1)

- Video

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

Best Moves – Dancing Guide (2)

- „Berechnet“ den dazupassenden Tanz zur gegebenen (Ball-) Musik
- Seit November für Android, seit Mitte April auch für iPhone

- Analysiert Takt- und Audiosignal-Features (Onset detector histogram, MFCC) direkt am Handy – keine Netzwerkverbindung notwendig
- Online Feedback möglich
- Training mit WEKA basierend auf ca. 600 Testliedern

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

Ein Frühwarnsystem für Bot-Netze (1)

Forschungsprojekt im Bereich IT Security

- Komplementär zum klassischen Spamfiltern
- Vorbeugende Identifizierung und Früherkennung der Ursache von Spam – Bots bzw. Bot-Netze

Vorgehensweise

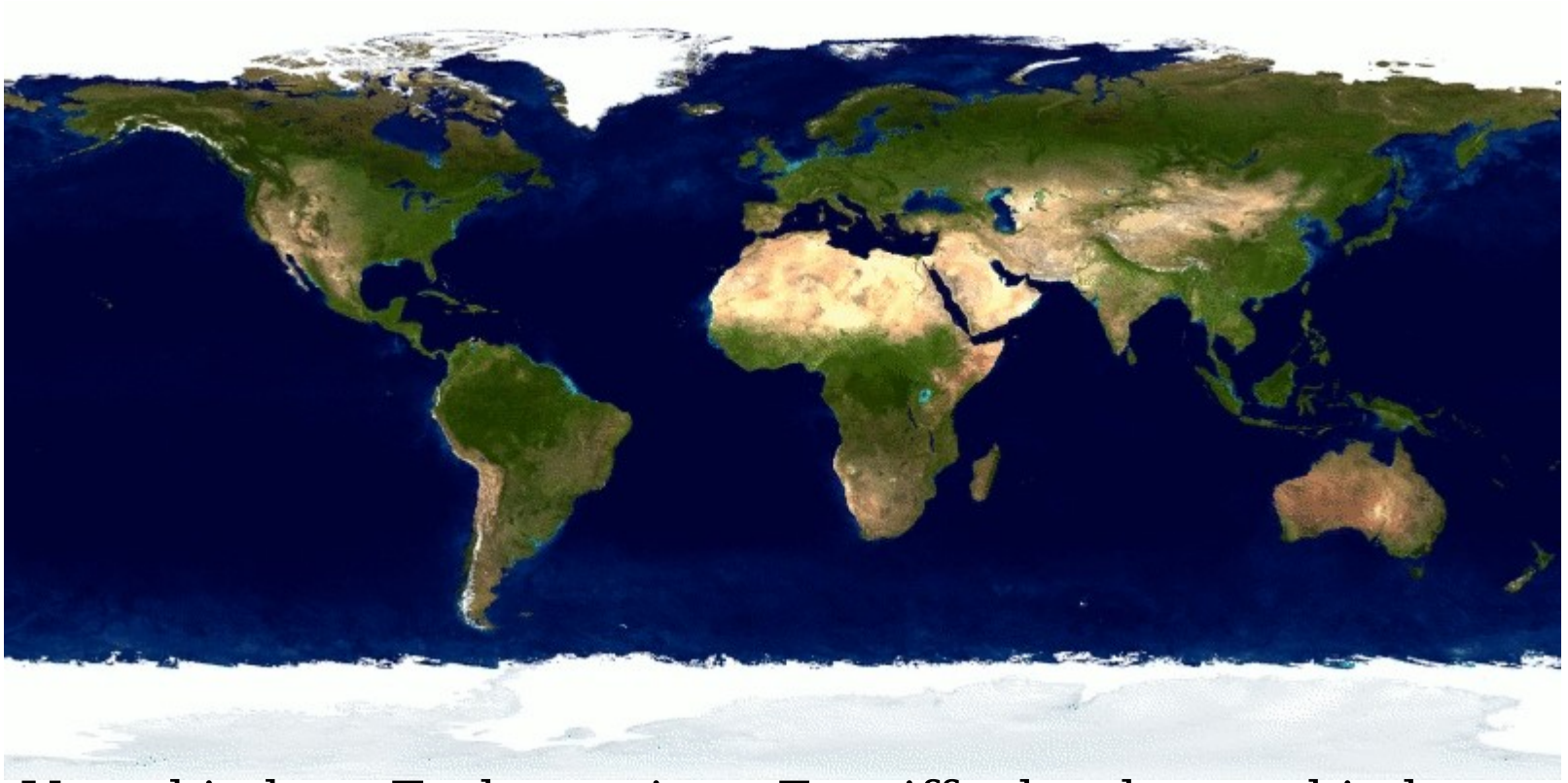
- Referenzdaten zu bekannten Bots- und Bot-Netzen
- Trainieren von Lernmodellen zur Erkennung von TCP/IP-Traffic eines bestimmten Bots
- Validierung und Test

Basiert vollständig auf Open-Source Software; WEKA wird für alle Lernmodelle & spezifische Vorverarbeitung verwendet.

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

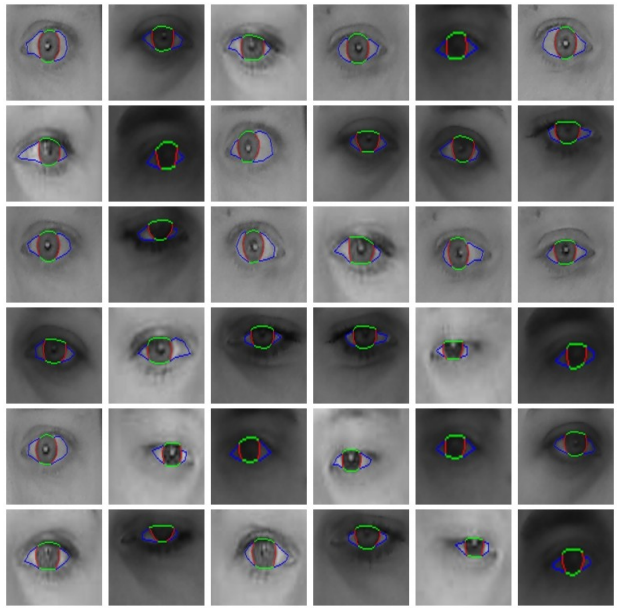
Low-cost Eyetracking

Warum?

- teure Eye-Tracking Systeme mit Spezialhardware
- keine Anwendung in gewohnter Umgebung möglich
- keine Analyse von reinem Videomaterial möglich

Entwicklung eines Low-cost Eyetrackers

- Off-the-shelf USB-Kameras, integrierte Notebook-Kameras, Handycams, ...
- Analyse Videomaterial möglich
- Segmentierung der Augen durch in WEKA trainiertes Modell (drei Klassen: Haut, Sclera, Iris/Pupille)



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

Watching C. Elegans Think (1)

Basic research project in Systems Neuroscience

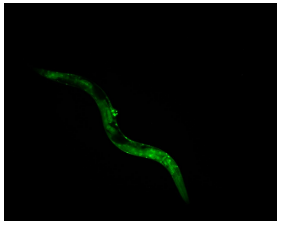
Four Objectives

- 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)

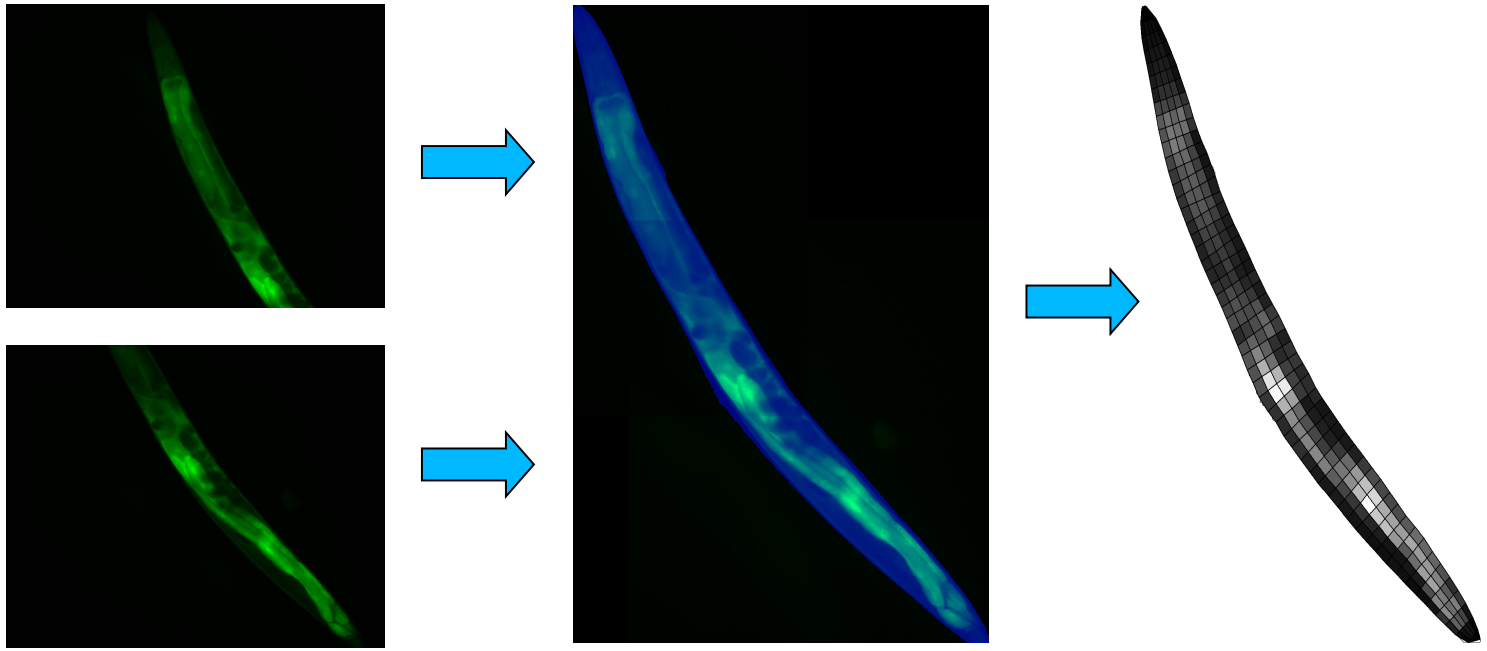


Image processing done via ImageIJ & WEKA

Reduces workload by 80%, paper upcoming

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

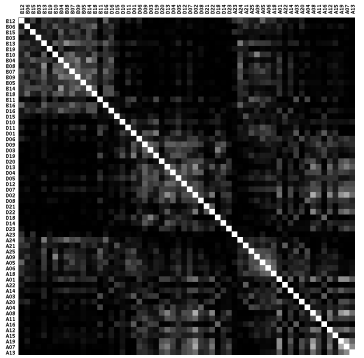
alex@seewald.at
http://www.seewald.at

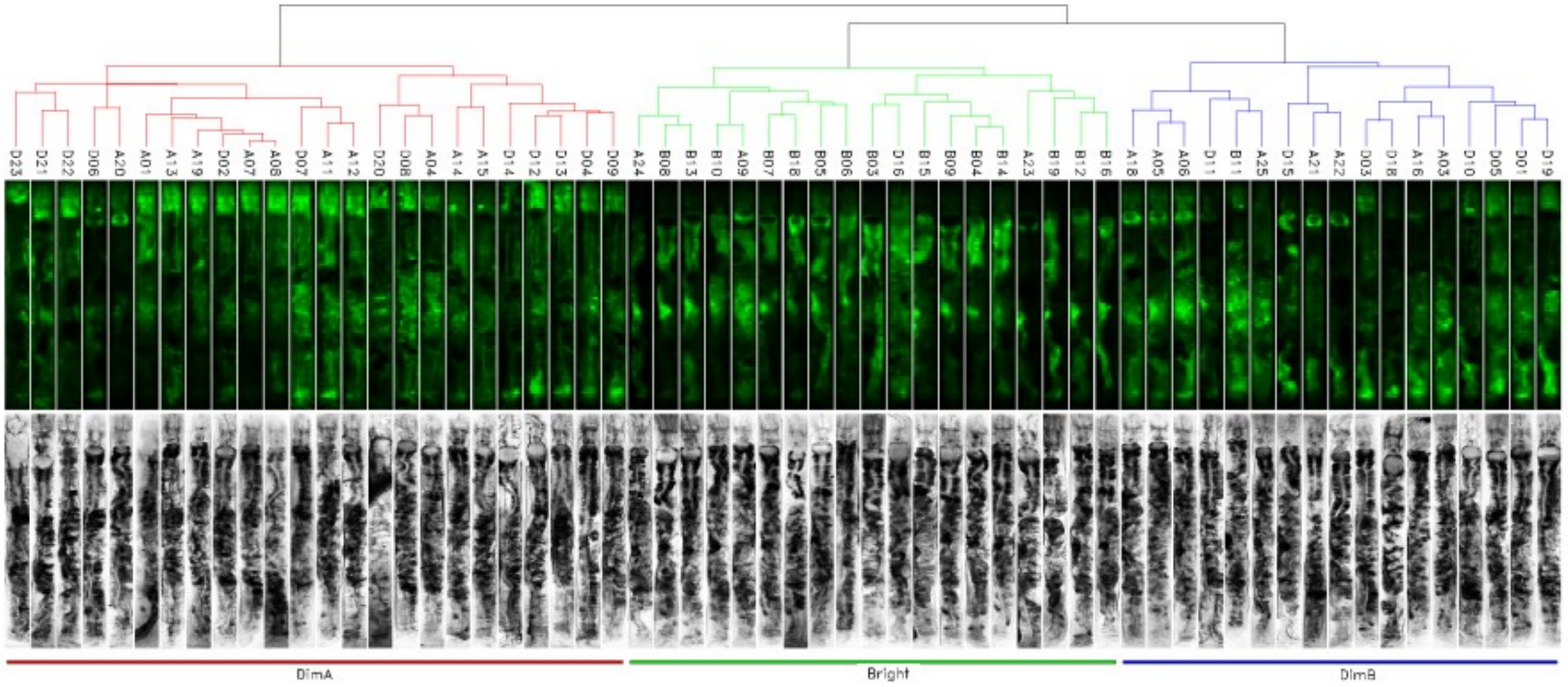
Watching C. Elegans Think (3)

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

deFlicker

Warum?

- Weit verbreitete Scheinwerfer flackern auf hoher Frequenz.
- Nachtaufnahmen mit (>150fps) Hochgeschwindigkeitskameras flackern deshalb stark.

Entwicklung eines Prototypen

- Entfernung des Flackerns in Echtzeit (720p / 50-60Hz)
- Eingesetzt bei UEFA Test-spielen, Olympia 2010 (Vancouver), derzeit ausgestellt bei NAB 2010 (Las Vegas)

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