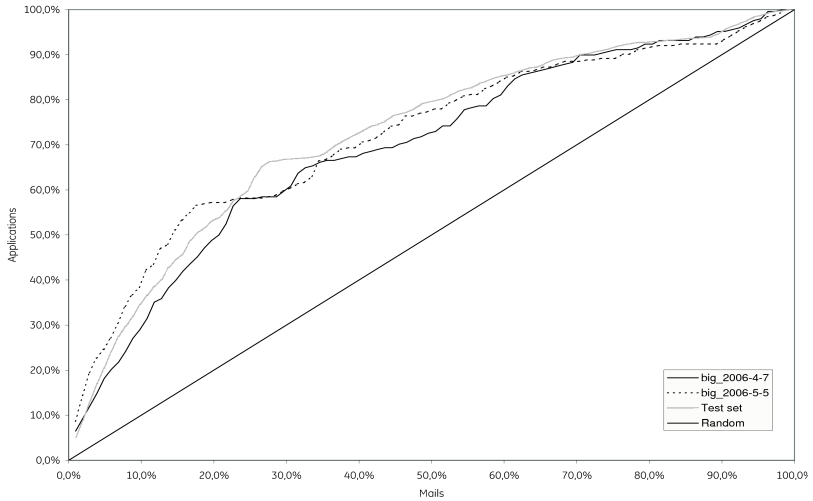
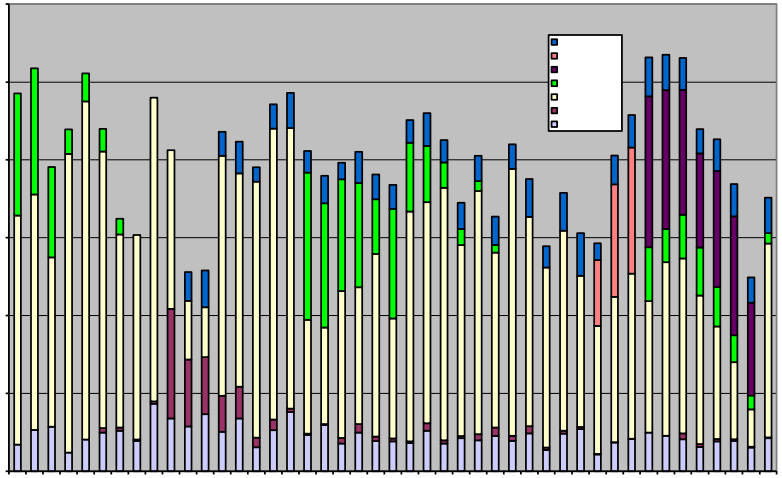
Open Source Data Mining mit WEKA





Dr. Alexander K. Seewald

SEEWALD SOLUTIONS

Was ist WEKA? (1)

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

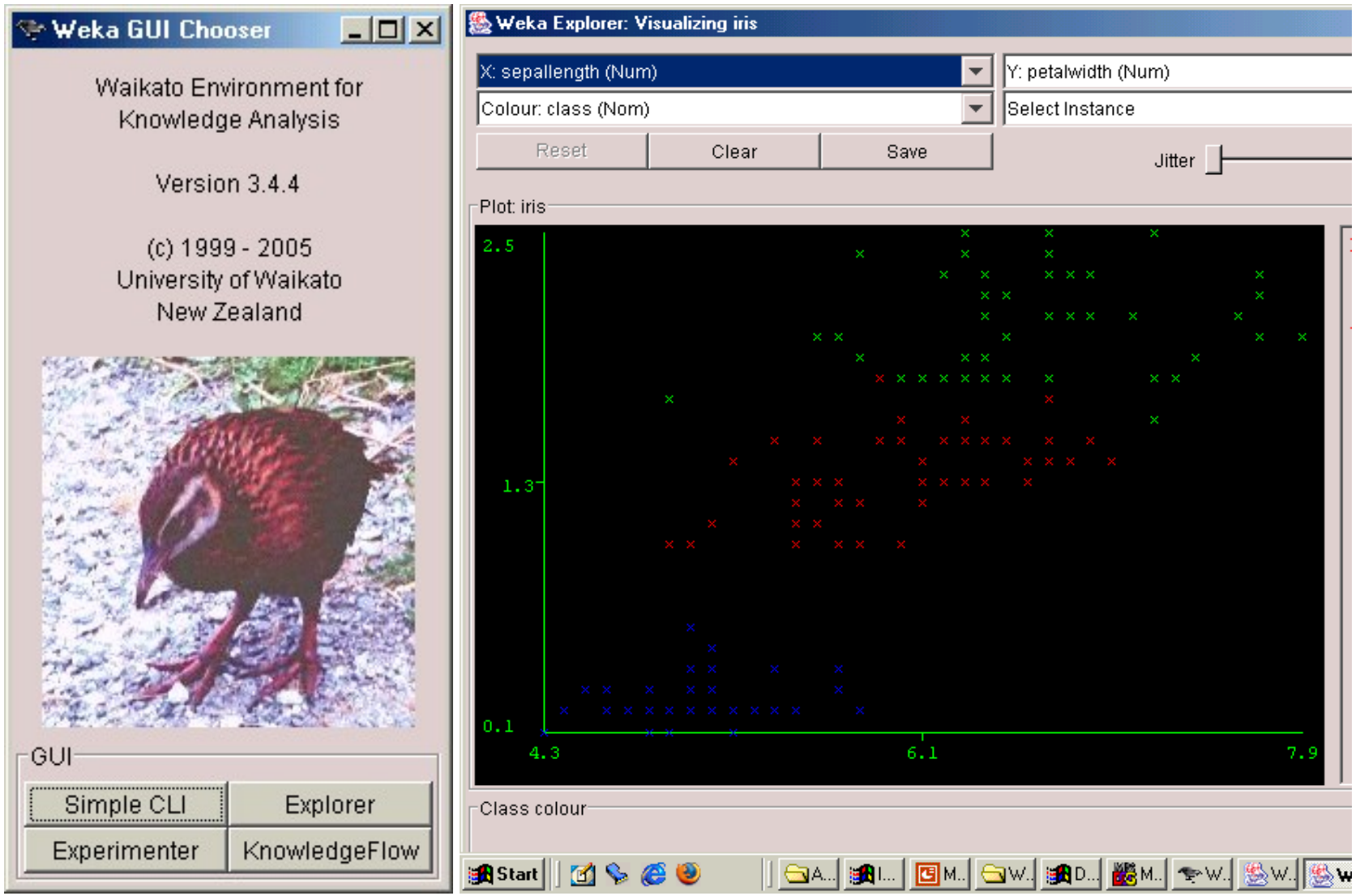
- Benannt nach einem neugierigen flügellosen Vogel, der in Neuseeland heimisch ist und unter Naturschutz steht
- Seit 1999 entwickelt, Open Source (GPL)
- Tausende von Contributers
- Stabilität, Verfügbarkeit und Qualität der Lernalgorithmen weit jenseits von kommerziell verfügbaren Tools

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)



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

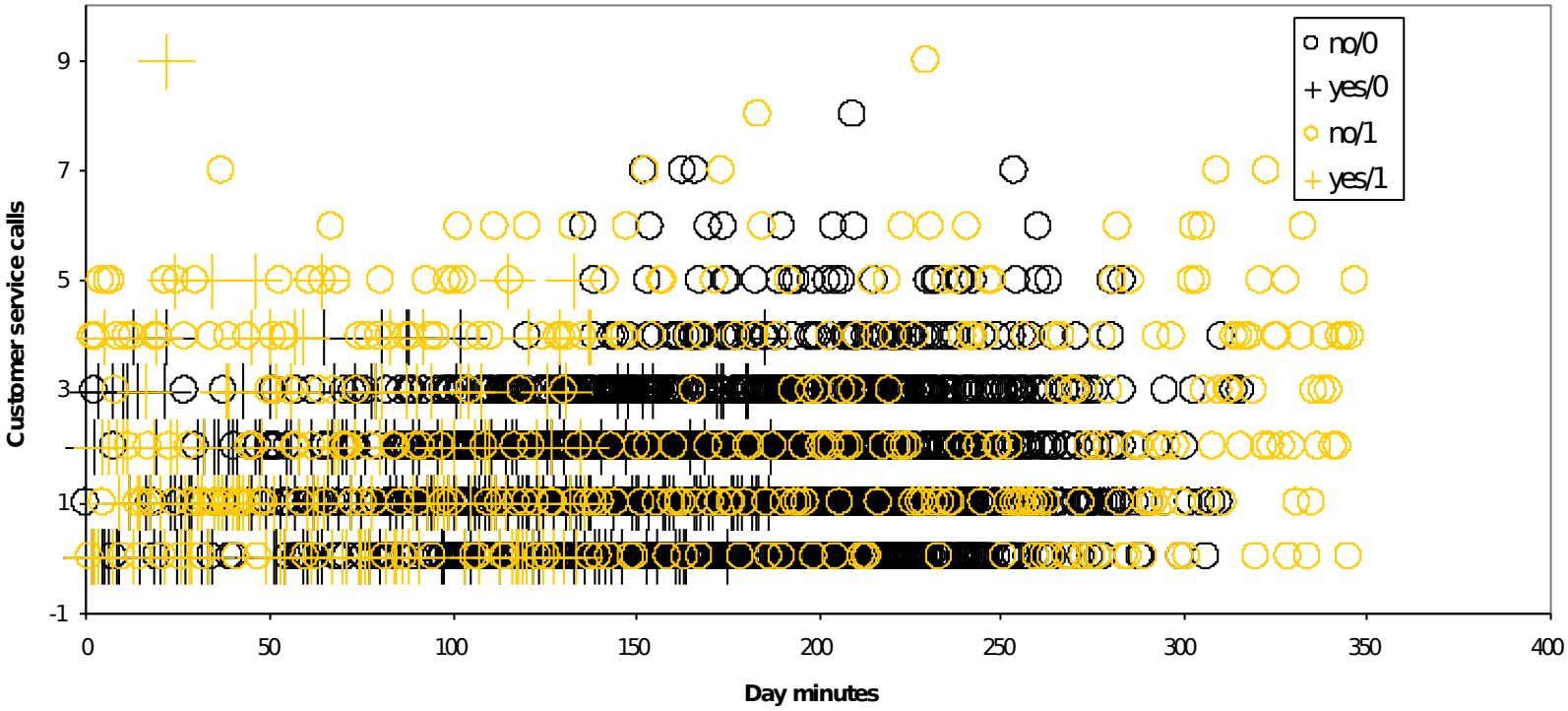
Übersicht

- **Kunden-Abwanderung, Telekoms**
- **Marketing-Effizienz erhöhen, Banken**
- **Validierung Marketingmaßnahmen, Banken**
- **BioMinT – Biologisches Text-Mining**
- **IGO 2 - Image-Mining mit WEKA**
- **Ein Frühwarnsystem für Bot-Netze – WEKA gegen Spam, Bots und Bot-Netze**

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

Kunden-Abwanderung verringern (1)

- **Problem:** Kundenabwanderung/*Churn* = hat Kunde am Monatsende den Provider gewechselt?
- **Lösung:** Modell von „anfälligen Kunden“, damit die Kundenabwanderung reduziert werden kann!



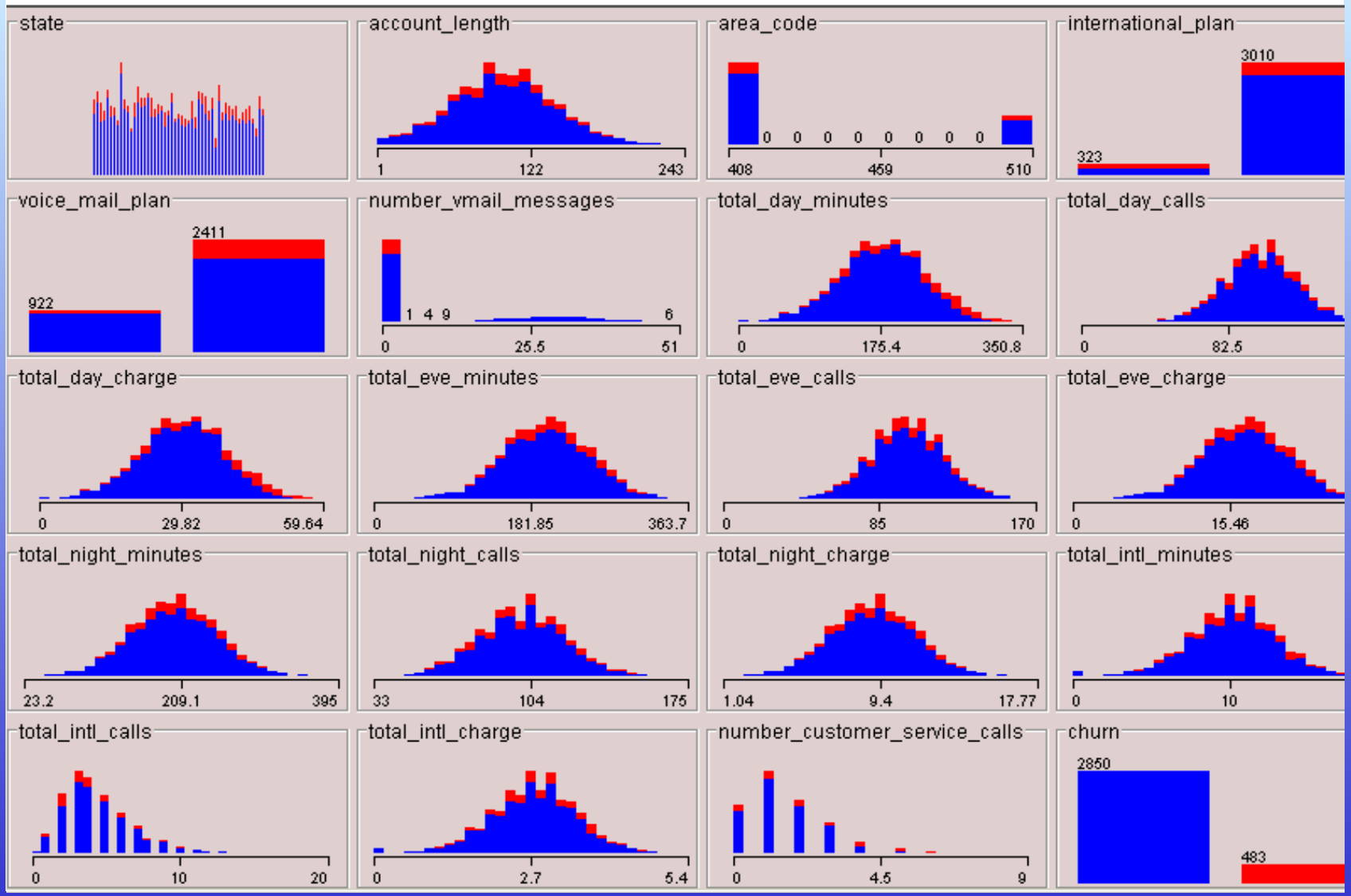
alex@seewald.at
http://www.seewald.at

Kunden-Abwanderung verringern (2)

Field name	Description
State	Two-letter abbreviation of US state
Acc. Len	Account length
Area	Area code
Int. Plan	Customer has international plan?
Voice Mail	Customer has Voice Mail?
VMail mins.	Minutes of Voice Mail
Day mins.	Telephone minutes during the day
Day calls	Number of telephone calls during the day
Day charge	Incurred charges for day telephony
Eve mins.	Telephone minutes during the evening
Eve calls	Number of telephone calls during the evening
Eve charge	Incurred charges for telephony during the evening
Night mins.	Telephone minutes at night (22:00-08:00)
Night calls	Number of telephone calls at night
Night charge	Incurred charges for telephony at night
Intl. mins.	Telephone minutes for internation calls
Intl. Calls	Number of international calls, last month
Intl. Charge	Incurred charges for international telephony
Serv. Calls	Number of service calls last month
True Churn	*Has customer switched operator at the end of month? = 1*
Pred. Churn	*predict true churn as a function of the variables (except true churn) here!*

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

Kunden-Abwanderung verringern (3)



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

Regeln für Churn (1)

(total_day_minutes >= 245) and (total_eve_minutes >= 225.2) and (voice_mail_plan = no) and (total_night_minutes >= 170.6) => churn=True.

(total_day_minutes >= 236.9) and (total_night_minutes >= 230.6) and (voice_mail_plan = no) and (total_eve_minutes >= 197.7) => churn=True.

(total_day_minutes >= 223.3) and (total_day_minutes >= 264.8) and (voice_mail_plan = no) and (total_eve_minutes >= 188) and (total_night_minutes >= 132.9) => churn=True.

(total_day_minutes >= 222.3) and (total_day_minutes >= 286.2) and (voice_mail_plan = no) and (total_eve_minutes >= 150.8) => churn=True.

(total_day_minutes >= 221.9) and (total_eve_minutes >= 261.6) and (voice_mail_plan = no) => churn=True.

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

Regeln für Churn (2)

(number_customer_service_calls >= 4) and (total_day_minutes <= 160) and (total_eve_minutes <= 233.2) and (total_night_minutes <= 254.9) => churn=True.

(number_customer_service_calls >= 4) and (total_day_minutes <= 182.1) and (total_eve_minutes <= 190.7) and (total_night_minutes <= 285) => churn=True.

(number_customer_service_calls >= 4) and (total_day_minutes <= 135.9) and (account_length >= 72) => churn=True.

(number_customer_service_calls >= 4) and (total_eve_minutes <= 135) => churn=True.

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

Regeln für Churn (3)

(international_plan = yes) and (total_intl_minutes >= 13.2) => churn=True.

(international_plan = yes) and (total_intl_calls <= 2) => churn=True.

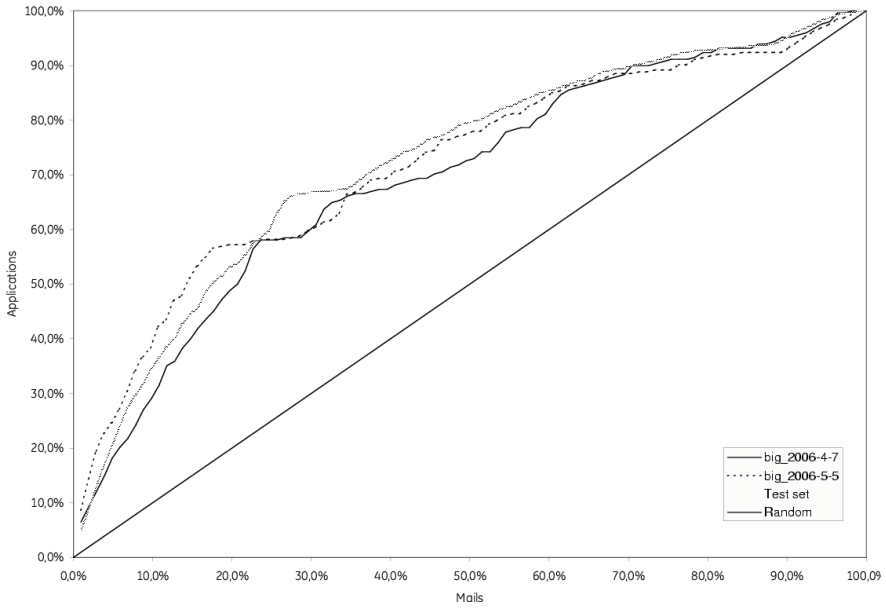
=> churn=False.

Erstellt in ein paar Minuten mit WEKA mittels rules.JRip: Beschreibung drei potentieller Zielgruppen für Churn-Reduktion gefunden.

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.

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

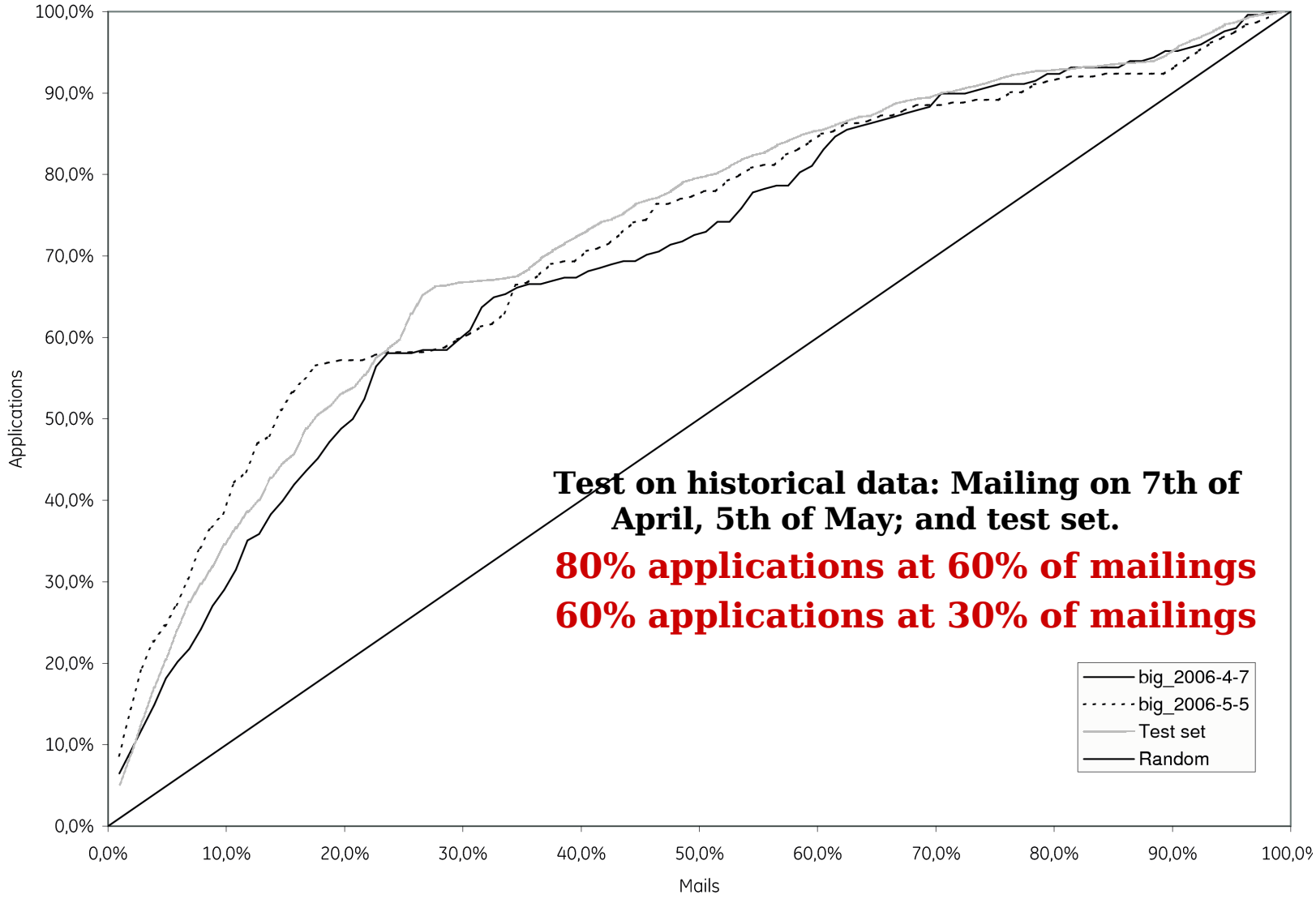
• totalPastAppl	No. of applications for CL F2F last 180 days
• Dependant_partner	Spouse with / without income
• No_accounts	Total number of past contracts
• Worst_payment	Worst paying_score on all contracts
• No_deferrals_not_liquidated	number of deferrals on all active contracts
• Industry	"Hauptbranche"
• 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)

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

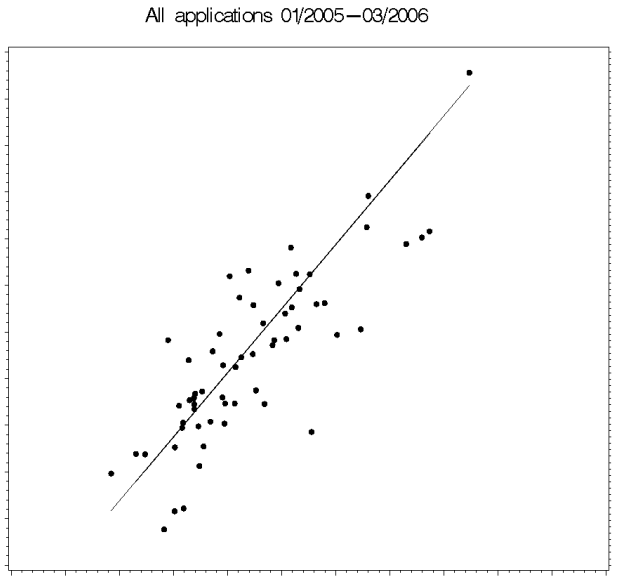
Marketing-Effizienz erhöhen (4)

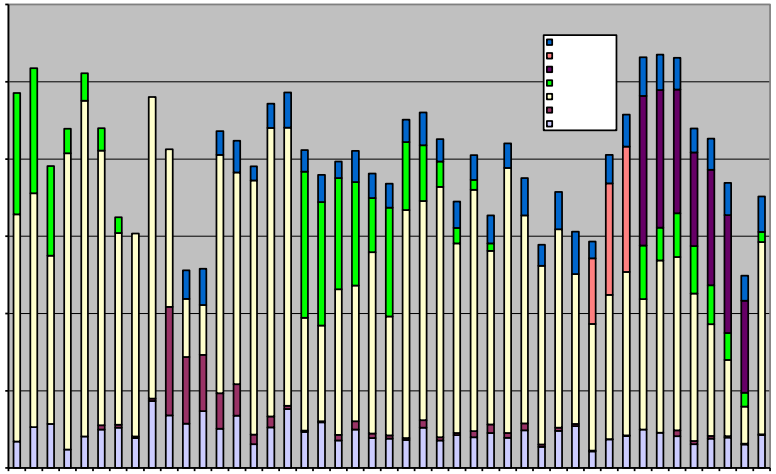


alex@seewald.at
http://www.seewald.at

Validierung Marketing-Maßnahmen (1)

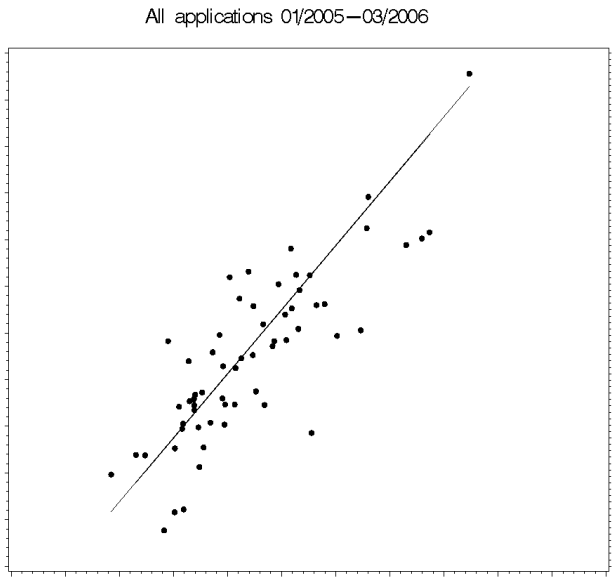
- **Problem:** Uneinheitliches Reporting – keine definitive Effektgröße pro Marketing-Maßnahme
- **Lösung:** Input/Output-Analyse aller Marketing-Maßnahmen über ein Jahr





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 linearily proportional to the number of mails sent out.

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

Validierung Marketing-Maßnahmen (3)

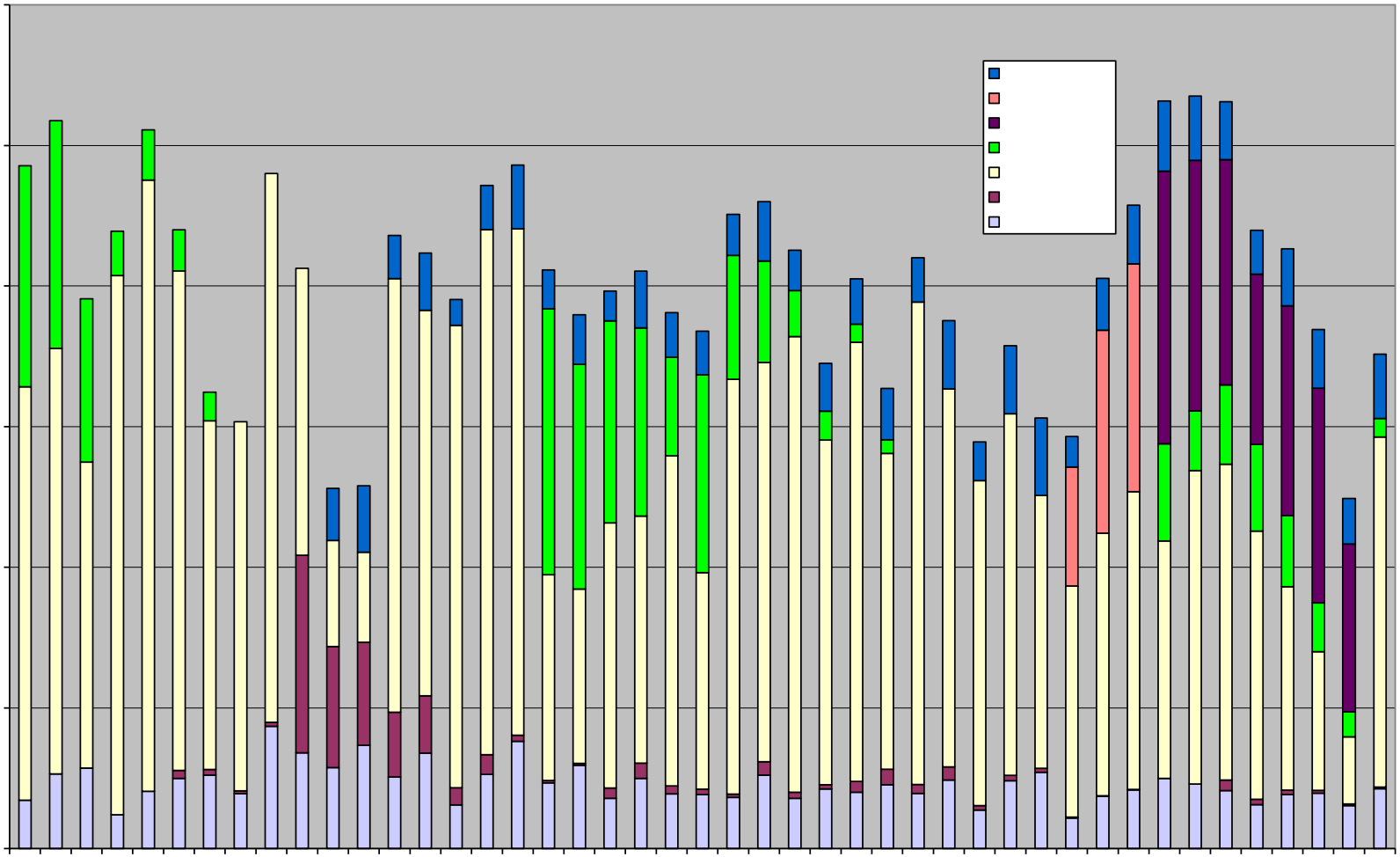
Extending the Mailing Response Model

- 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)



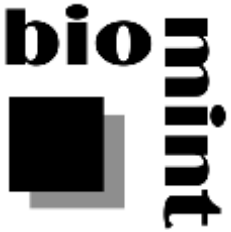
alex@seewald.at
http://www.seewald.at

BioMinT: Biological Text Mining

Research project funded by the EU (2003 – 2005)

- 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
- Adapted to the needs of biological researchers in general and specifically for SwissProt / PRINTS annotation.

Useful metaphor: In-silico research/curator assistant



biomint.pharmadm.com

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)**

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

BioMinT: Species from MEDLINE (3)

Predict the species of an organism from MEDLINE publ.

- 19.0% Baseline (most common class *Human*)
- 26.5% Rule based on single word *Fungal* (OneR, 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

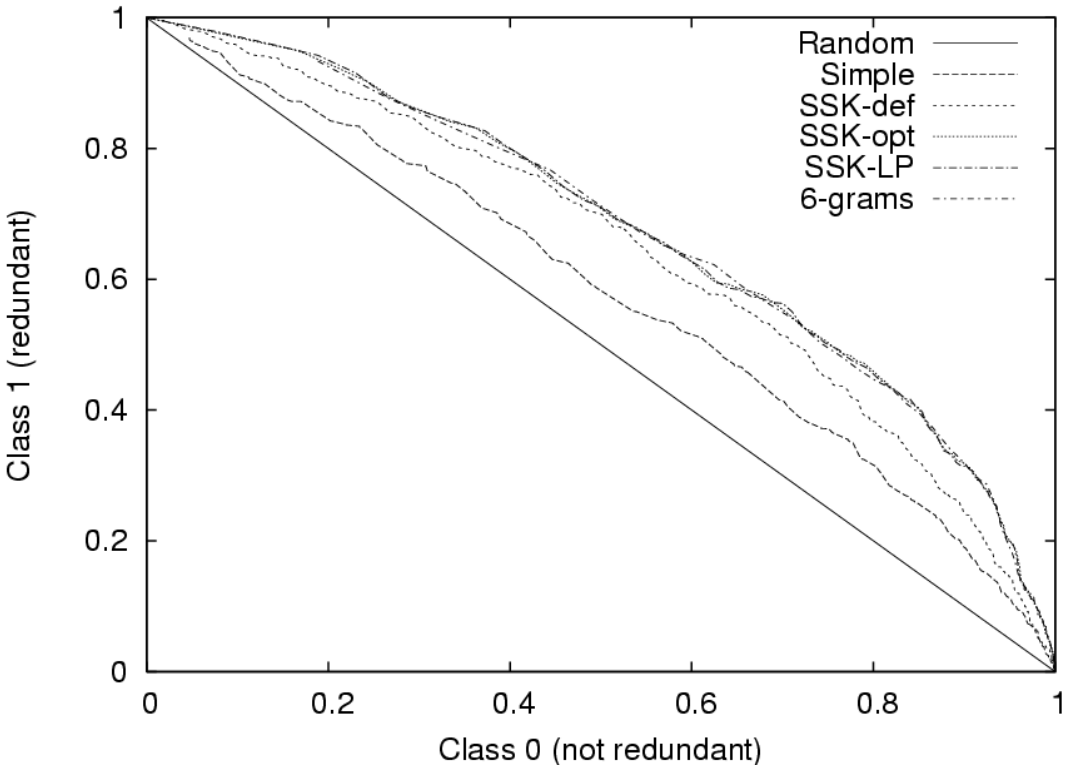
- Expert: + precision, – recall; — F-Measure
- JRip: – precision, + recall; ++ F-measure

SEEWALD SOLUTIONS
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.

We contributed the SSK & SSK-LP algorithms to WEKA after concluding these experiments,which resulted in the journal publication [Seewald& Kleedorfer, 2007].



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

Watching C. Elegans Think (1)

Basic research project in Systems Neuroscience

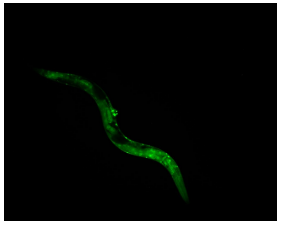
Four Objectives

- 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

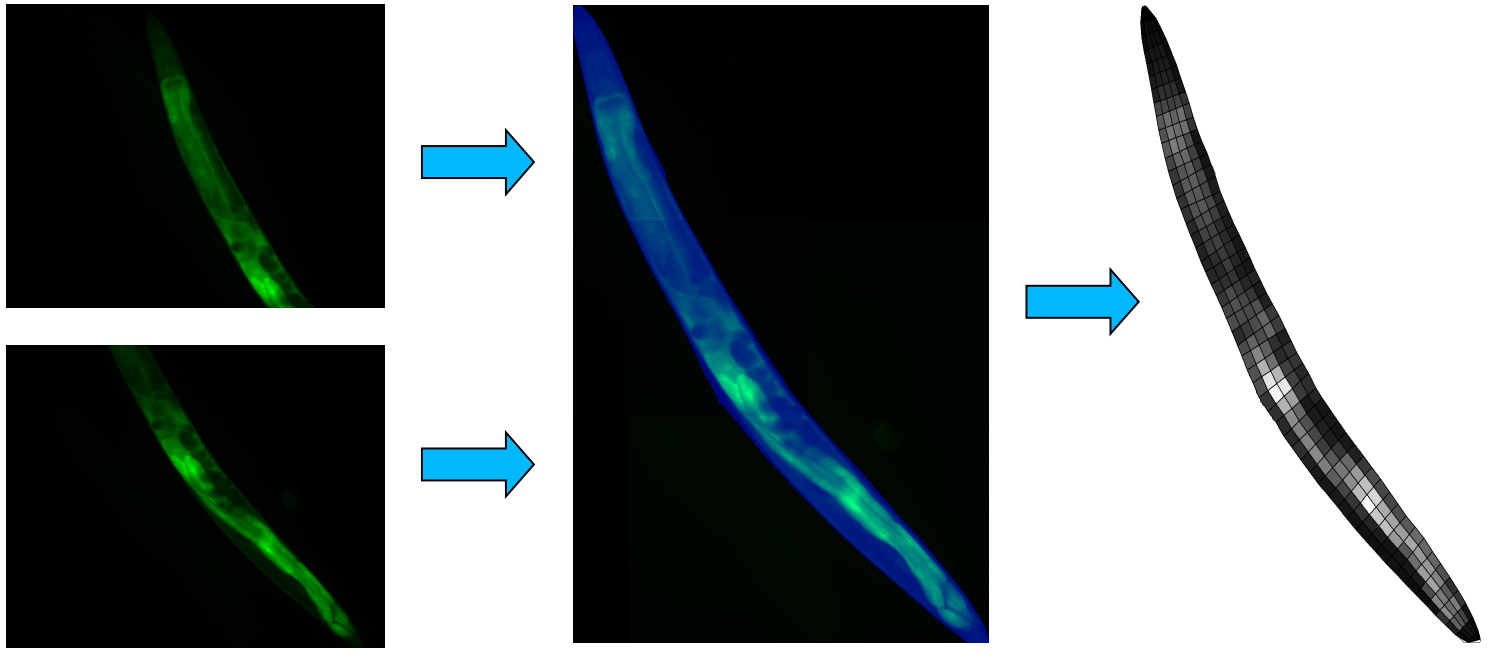
Might be feasible to simulate



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)



Reduces workload by 80% for tagging worm images, fully automated system seems feasible (eng. obj.)

Development time about 3PM, ongoing collaboration

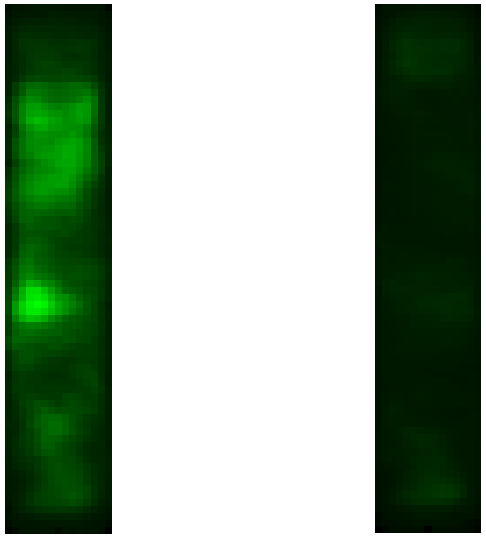
alex@seewald.at
http://www.seewald.at

Watching C. Elegans Think (3)

Some interesting results...

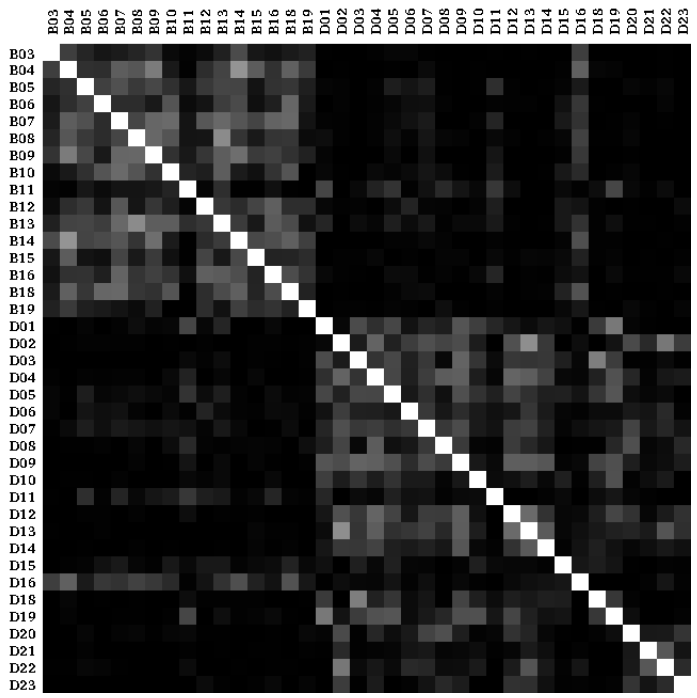
Known: Bright worms live longer than dim worms.

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



Avg. Bright

Avg. Dim



SEEWALD SOLUTIONS

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

Image-Mining mit WEKA

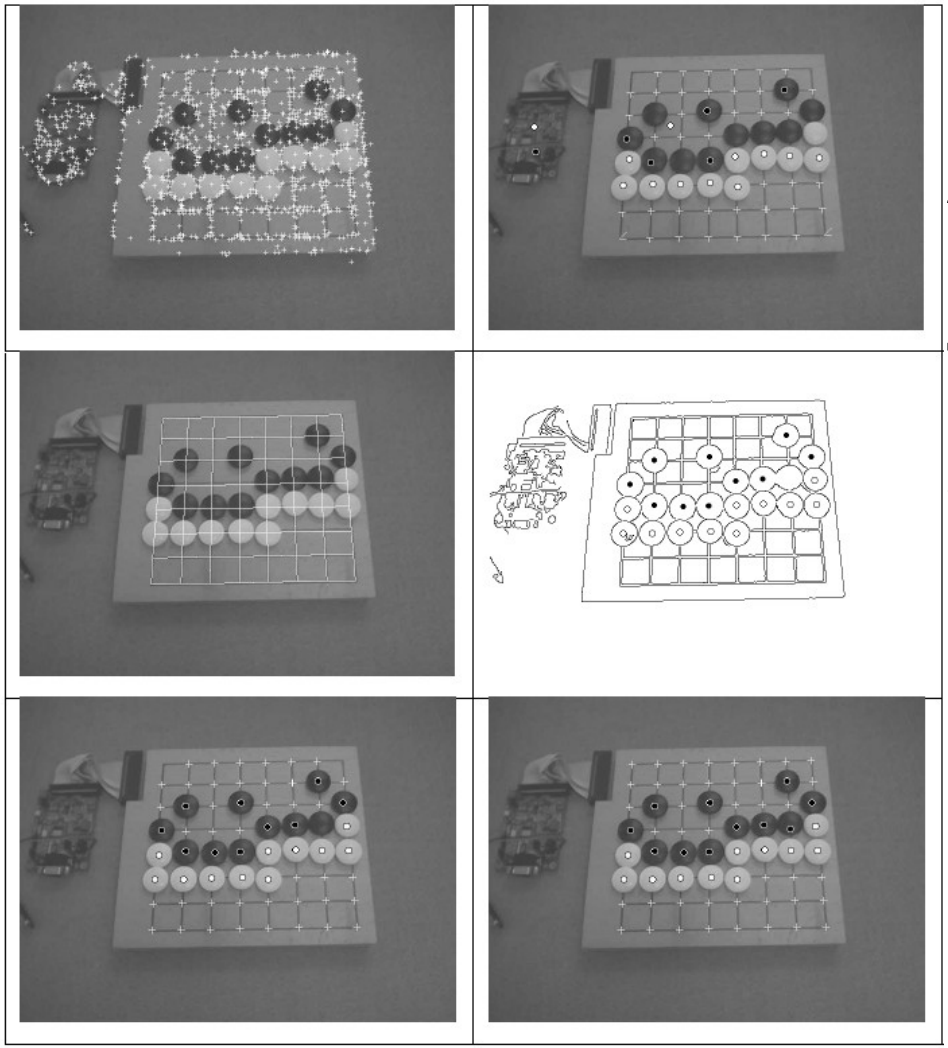


Figure 1: Steps 1-6 with sample images after each step, left-to-right, top-to-bottom.

1.SIFT keypoint detection

2. Keypoint classification

3.Estimate board position

4.Canny stone detection

5.Keypoint stone class.

6.Last-ditch classification

98.4% accurate, 4/6 steps use WEKA

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

Ein Frühwarnsystem für Bot-Netze (1)

Aktuelles Forschungsprojekt im Bereich IT Security

- Komplementär zum klassischen Spamfiltern
- Vorbeugende Identifizierung und Früherkennung der Ursache von Spam – Bots bzw. Bot-Netze
- Gefördert von IPA als NetIdee 2007-Projekt

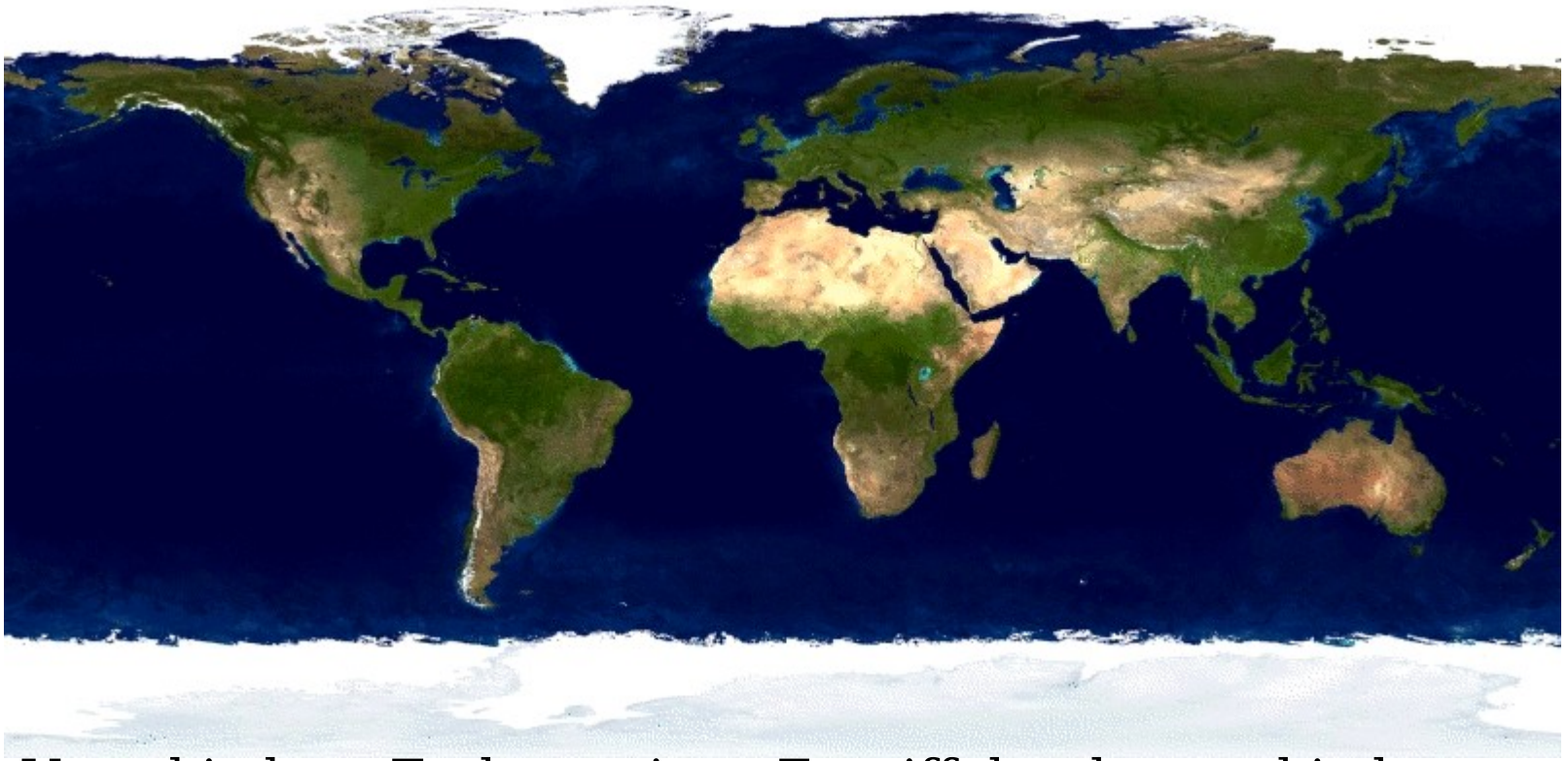
Vorgehensweise

- 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.

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

Ein Frühwarnsystem für Bot-Netze (2)



Verschiedene Farben zeigen Zugriff durch verschiedene Spambots an. 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

Vielen Dank für die Aufmerksamkeit!

Für Fragen stehe ich jederzeit gerne zu Ihrer Verfügung.

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