Maschinelles Lernen
und
Data Mining

Univ.-Lektor Dr.techn. Alexander K. Seewald
Österreichisches Forschungsinstitut
für Artificial Intelligence

alex@seewald.at / alex.seewald.at

Overview

- **Who** are we? - Institutional context
- **How** does this work? - Organisation(al) matters
- **What** else do we teach? - Other lectures on ML & DM
- **Where** to look for additional information? - Resources
- **When** will this all happen? - Preliminary lecture plan
- **Why** is this interesting? - Motivation and Introduction, plus some basic definitions.

alex@seewald.at / alex.seewald.at

Who are we?

http://www.ai.univie.ac.at/

Institut für Medizinische Informatik und Artificial Intelligence (IMKAI) des Zentrums für Hirnforschung der Medizinischen Universität Wien

http://www.oefai.at/oefai

Österreichisches Forschungsinstitut für Artificial Intelligence (ÖFAI)

http://www.oefai.at/oefai/ml/mldm

The Machine Learning and Data Mining Research Group @ ÖFAI

alex@seewald.at / alex.seewald.at

How does this work?

Vorlesungstermine ML & DM (509.014)

Jeden Di 16:15-17:45, IMKAI Seminarraum

Vorlesungsprüfung

1.2.2005 16:15-17:45 schriftlich, IMKAI Seminarraum. Alle Unterlagen erlaubt! Weitere Termine nach Bedarf.

Übung ML & DM (561.340) - Empfohlen zur Vorbereitung auf die Prüfung

- Ausgabe von Übungsaufgaben: ab 19.10. in VO & auf LVA-Webseite (s.u.)
- Abgabe der Lösungen in schriftlicher Form jeweils zum angegebenen Termin
- Gemeinsame Diskussion der Lösungen jeweils zum angegebenen Termin

Webseite zur LVA (passwort-geschützt)

http://www.oefai.at/~alexsee/lv/mldm - User: **student** Passwort: **ml+dm04**

WICHTIG: Mitbelegen an der *Medizinischen Universität*!

http://www.meduniwien.ac.at/index.php?id=81

alex@seewald.at / alex.seewald.at

What else do we teach?

Lectures on ML & DM @ IMKAI

Wintersemester

509.014 Maschinelles Lernen und Data Mining, VO 2h

561.340 Maschinelles Lernen und Data Mining, UE 1h

Sommersemester

509.916 AI-Methoden der Datenanalyse, VO 2h

561.699 AI-Methoden der Datenanalyse, LU 1h

Jederzeit: Praktika, Diplomarbeiten, Diss., Projektmitarbeit...

alex@seewald.at / alex.seewald.at

Where to look for additional information?

Recommended books

- T. Hastie et al (2001). *The Elements of Statistical Learning*. Springer Verlag.
- Tom M. Mitchell (1997). *Machine Learning.* McGraw-Hill
- Ian H. Witten & Eibe Frank (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

...

Starting points on the Web:

- The European Network of Excellence in Machine Learning (MLNet), with bibliography, links to machine learning courses, data, software, ...
  http://www.mlnet.org

- The European Network of Excellence on Knowledge Discovery (KDNet)
  http://www.kdnet.org/

- David Aha's extensive list of machine learning resources
  http://home.earthlink.net/~dwaha/research/machine-learning.html

alex@seewald.at / alex.seewald.at

When will this all happen?

Lecture plan (preliminary)

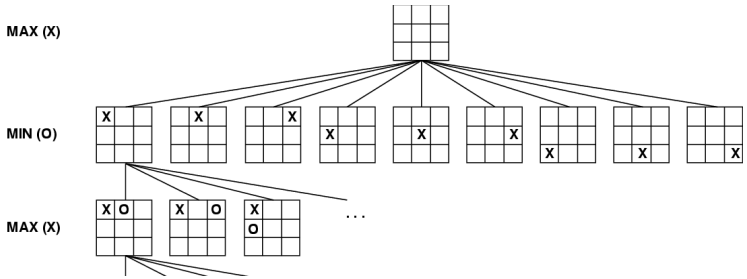
05.10.2004	Introduction
12.10.2004	Inductive Concept Learning: Basic Concepts
19.10.2004	Decision Tree Learning + Ausgabe Übungsfragen 1
02.11.2004	Evaluation Basics & Overfitting Avoidance + **Gruppeneinteilung**
09.11.2004	Instance-Based Learning & Bayesian Methods
16.11.2004	Simple Learning of Classification Rules & Ripper
17.11.2004	**Deadline Übungsfragen 1**
23.11.2004	Relational Learning and ILP **+ Besprechung ÜF 1 + Ausgabe ÜF 2**
30.11.2004	Computational Learning Theory
07.12.2004	Elements of Statistical Learning
14.12.2004	Ranking, unsupervised learning, Clustering and Ensembles
11.01.2005	Applications of ML & DM
18.01.2005	Reinforcement Learning
19.01.2005	**Deadline Übungsfragen 2**
25.01.2005	Construction of learning problems & NFL theorem **+ Bespr. ÜF2**
01.02.2005	Final exam (written)

alex@seewald.at / alex.seewald.at

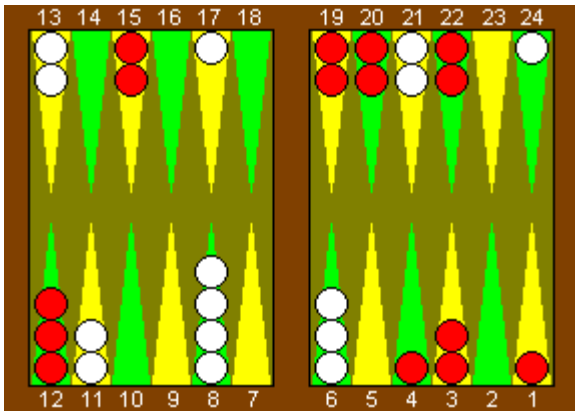
Can machines learn from Experience?

Yes, of course!

- MENACE (D.Michie, 1961): Tic-Tac-Toe by matchboxes
  - Lookup table for each position
  - Feedback after each game
  - Implemented as 308 matchboxes and beads in nine colors.



- TD-Gammon (G.Tesauro, 1995): Backgammon
  - Started with no knowledge on backgammon except basic rules
  - Learned by playing against itself
  - Result: Excellent play at grandmaster level. In some cases, has even changed expert's judgement on best move.



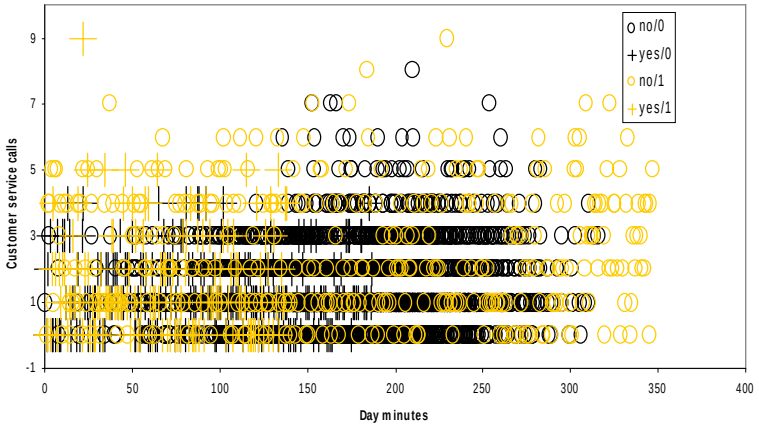
alex@seewald.at / alex.seewald.at

Why? Example: Customer Churn

St.	Acc. Len	Area	Int. Plan	Voice Mail	VMail mins.	Day mins.	Day calls	Day charg	Eve mins.	Eve calls	Eve char	Night mins.	Night calls	Night charge	Intl. mins.	Intl. Calls	Intl. Charge	Serv, Calls
KS	128	415	no	yes	25	265,1	110	45,07	197,4	99	16,8	244,7	91	11,01	10	3	2,7	1
OH	107	415	no	yes	26	161,6	123	27,47	195,5	103	16,6	254,4	103	11,45	13,7	3	3,7	1
NJ	137	415	no	no	0	243,4	114	41,38	121,2	110	10,3	162,6	104	7,32	12,2	5	3,29	0
OH	84	408	yes	no	0	299,4	71	50,9	61,9	88	5,26	196,9	89	8,86	6,6	7	1,78	2
OK	75	415	yes	no	0	166,7	113	28,34	148,3	122	12,6	186,9	121	8,41	10,1	3	2,73	3
AL	118	510	yes	no	0	223,4	98	37,98	220,6	101	18,8	203,9	118	9,18	6,3	6	1,7	0
MA	121	510	no	yes	24	218,2	88	37,09	348,5	108	29,6	212,6	118	9,57	7,5	7	2,03	3
MO	147	415	yes	no	0	157	79	26,69	103,1	94	8,76	211,8	96	9,53	7,1	6	1,92	0
LA	117	408	no	no	0	184,5	97	31,37	351,6	80	29,9	215,8	90	9,71	8,7	4	2,35	1
WV	141	415	yes	yes	37	258,6	84	43,96	222	111	18,9	326,4	97	14,69	11,2	5	3,02	0
IN	65	415	no	no	0	129,1	137	21,95	228,5	83	19,4	208,8	111	9,4	12,7	6	3,43	4
RI	74	415	no	no	0	187,7	127	31,91	163,4	148	13,9	196	94	8,82	9,1	5	2,46	0
IA	168	408	no	no	0	128,8	96	21,9	104,9	71	8,92	141,1	128	6,35	11,2	2	3,02	1
MT	95	510	no	no	0	156,6	88	26,62	247,6	75	21,1	192,3	115	8,65	12,3	5	3,32	3
IA	62	415	no	no	0	120,7	70	20,52	307,2	76	26,1	203	99	9,14	13,1	6	3,54	4
NY	161	415	no	no	0	332,9	67	56,59	317,8	97	27	160,6	128	7,23	5,4	9	1,46	4
ID	85	408	no	yes	27	196,4	139	33,39	280,9	90	23,9	89,3	75	4,02	13,8	4	3,73	1
VT	93	510	no	no	0	190,7	114	32,42	218,2	111	18,6	129,6	121	5,83	8,1	3	2,19	3
VA	76	510	no	yes	33	189,7	66	32,25	212,8	65	18,1	165,7	108	7,46	10	5	2,7	1
TX	73	415	no	no	0	224,4	90	38,15	159,5	88	13,6	192,8	74	8,68	13	2	3,51	1
FL	147	415	no	no	0	155,1	117	26,37	239,7	93	20,4	208,8	133	9,4	10,6	4	2,86	0
CO	77	408	no	no	0	62,4	89	10,61	169,9	121	14,4	209,6	64	9,43	5,7	6	1,54	5
AZ	130	415	no	no	0	183	112	31,11	72,9	99	6,2	181,8	78	8,18	9,5	19	2,57	0
SC	111	415	no	no	0	110,4	103	18,77	137,3	102	11,7	189,6	105	8,53	7,7	6	2,08	2
VA	132	510	no	no	0	81,1	86	13,79	245,2	72	20,8	237	115	10,67	10,3	2	2,78	0
NE	174	415	no	no	0	124,3	76	21,13	277,1	112	23,6	250,7	115	11,28	15,5	5	4,19	3
WY	57	408	no	yes	39	213	115	36,21	191,1	112	16,2	182,7	115	8,22	9,5	3	2,57	0
MT	54	408	no	no	0	134,3	73	22,83	155,5	100	13,2	102,1	68	4,59	14,7	4	3,97	3
MO	20	415	no	no	0	190	109	32,3	258,2	84	22	181,5	102	8,17	6,3	6	1,7	0
HI	49	510	no	no	0	119,3	117	20,28	215,1	109	18,3	178,7	90	8,04	11,1	1	3	1
IL	142	415	no	no	0	84,8	95	14,42	136,7	63	11,6	250,5	148	11,27	14,2	6	3,83	2
NH	75	510	no	no	0	226,1	105	38,44	201,5	107	17,1	246,2	98	11,08	10,3	5	2,78	1
LA	172	408	no	no	0	212	121	36,04	31,2	115	2,65	293,3	78	13,2	12,6	10	3,4	3
AZ	12	408	no	no	0	249,6	118	42,43	252,4	119	21,5	280,2	90	12,61	11,8	3	3,19	1
OK	57	408	no	yes	25	176,8	94	30,06	195	75	16,6	213,5	116	9,61	8,3	4	2,24	0
GA	72	415	no	yes	37	220	80	37,4	217,3	102	18,5	152,8	71	6,88	14,7	6	3,97	3
AK	36	408	no	yes	30	146,3	128	24,87	162,5	80	13,8	129,3	109	5,82	14,5	6	3,92	0
MA	78	415	no	no	0	130,8	64	22,24	223,7	116	19	227,8	108	10,25	10	5	2,7	1
AK	136	415	yes	yes	33	203,9	106	34,66	187,6	99	16	101,7	107	4,58	10,5	6	2,84	3
NJ	149	408	no	no	0	140,4	94	23,87	271,8	92	23,1	188,3	108	8,47	11,1	9	3	1
GA	98	408	no	no	0	126,3	102	21,47	166,8	85	14,2	187,8	135	8,45	9,4	2	2,54	3
MD	135	408	yes	yes	41	173,1	85	29,43	203,9	107	17,3	122,2	78	5,5	14,6	15	3,94	0
AR	34	510	no	no	0	124,8	82	21,22	282,2	98	24	311,5	78	14,02	10	4	2,7	2
ID	160	415	no	no	0	85,8	77	14,59	165,3	110	14,1	178,5	92	8,03	9,2	4	2,48	3
WI	64	510	no	no	0	154	67	26,18	225,8	118	19,2	265,3	86	11,94	3,5	3	0,95	1
OR	59	408	no	yes	28	120,9	97	20,55	213	92	18,1	163,1	116	7,34	8,5	5	2,3	2
MI	65	415	no	no	0	211,3	120	35,92	162,6	122	13,8	134,7	118	6,06	13,2	5	3,56	3
DE	142	408	no	no	0	187	133	31,79	134,6	74	11,4	242,2	127	10,9	7,4	5	2	2
ID	119	415	no	no	0	159,1	114	27,05	231,3	117	19,7	143,2	91	6,44	8,8	3	2,38	5
WY	97	415	no	yes	24	133,2	135	22,64	217,2	58	18,5	70,6	79	3,18	11	3	2,97	1
IA	52	408	no	no	0	191,9	108	32,62	269,8	96	22,9	236,8	87	10,66	7,8	5	2,11	3
IN	60	408	no	no	0	220,6	57	37,5	211,1	115	17,9	249	129	11,21	6,8	3	1,84	1
VA	10	408	no	no	0	186,1	112	31,64	190,2	66	16,2	282,8	57	12,73	11,4	6	3,08	2
UT	96	415	no	no	0	160,2	117	27,23	267,5	67	22,7	228,5	68	10,28	9,3	5	2,51	2
WY	87	415	no	no	0	151	83	25,67	219,7	116	18,7	203,9	127	9,18	9,7	3	2,62	5
IN	81	408	no	no	0	175,5	67	29,84	249,3	85	21,2	270,2	98	12,16	10,2	3	2,75	1
CO	141	415	no	no	0	126,9	98	21,57	180	62	15,3	140,8	128	6,34	8	2	2,16	1
CO	121	408	no	yes	30	198,4	129	33,73	75,3	77	6,4	181,2	77	8,15	5,8	3	1,57	3
WI	68	415	no	no	0	148,8	70	25,3	246,5	164	21	129,8	103	5,84	12,1	3	3,27	3
OK	125	408	no	no	0	229,3	103	38,98	177,4	126	15,1	189,3	95	8,52	12	8	3,24	1
ID	174	408	no	no	0	192,1	97	32,66	169,9	94	14,4	166,6	54	7,5	11,4	4	3,08	1
CA	116	415	no	yes	34	268,6	83	45,66	178,2	142	15,2	166,3	106	7,48	11,6	3	3,13	2
MN	74	510	no	yes	33	193,7	91	32,93	246,1	96	20,9	138	92	6,21	14,6	3	3,94	2
SD	149	408	no	yes	28	180,7	92	30,72	187,8	64	16	265,5	53	11,95	12,6	3	3,4	3
NC	38	408	no	no	0	131,2	98	22,3	162,9	97	13,9	159	106	7,15	8,2	6	2,21	2
WA	40	415	no	yes	41	148,1	74	25,18	169,5	88	14,4	214,1	102	9,63	6,2	5	1,67	2
WY	43	415	yes	no	0	251,5	105	42,76	212,8	104	18,1	157,8	67	7,1	9,3	4	2,51	0
MN	113	408	yes	no	0	125,2	93	21,28	206,4	119	17,5	129,3	139	5,82	8,3	8	2,24	0
UT	126	408	no	no	0	211,6	70	35,97	216,9	80	18,4	153,5	60	6,91	7,8	1	2,11	1
TX	150	510	no	no	0	178,9	101	30,41	169,1	110	14,4	148,6	100	6,69	13,8	3	3,73	4
NJ	138	408	no	no	0	241,8	93	41,11	170,5	83	14,5	295,3	104	13,29	11,8	7	3,19	3
MN	162	510	no	yes	46	224,9	97	38,23	188,2	84	16	254,6	61	11,46	12,1	2	3,27	0
NM	147	510	no	no	0	248,6	83	42,26	148,9	85	12,7	172,5	109	7,76	8	4	2,16	3

Given: A set of customers with state, area code, telephone number, and time/cost information for calls in one month; plus churn = have they switched to another provider by the end of the month?

Create a useful model of customer churn, so it can be reduced significantly!



alex@seewald.at / alex.seewald.at

Customer Churn: Learned Rules

(total_day_minutes >= 245) and (total_eve_minutes >= 225.2) and (voice_mail_plan = no) and (total_night_minutes >= 170.6) => churn=True. (64.0/0.0)

(number_customer_service_calls >= 4) and (total_day_minutes <= 160) and (total_eve_minutes <= 233.2) and (total_night_minutes <= 254.9) => churn=True. (69.0/0.0)

(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. (52.0/1.0)

(international_plan = yes) and (total_intl_minutes >= 13.2) => churn=True. (54.0/0.0)

(total_day_minutes >= 221.9) and (total_eve_minutes >= 261.6) and (voice_mail_plan = no) => churn=True. (41.0/7.0)

(international_plan = yes) and (total_intl_calls <= 2) => churn=True. (50.0/0.0)

(total_day_minutes >= 222.3) and (total_day_minutes >= 286.2) and (voice_mail_plan = no) and (total_eve_minutes >= 150.8) => churn=True. (17.0/2.0)

(number_customer_service_calls >= 4) and (total_day_minutes <= 182.1) and (total_eve_minutes <= 190.7) and (total_night_minutes <= 285) => churn=True. (22.0/0.0)

(number_customer_service_calls >= 4) and (total_day_minutes <= 135.9) and (account_length >= 72) => churn=True. (14.0/0.0)

(total_day_minutes >= 236.9) and (total_night_minutes >= 230.6) and (voice_mail_plan = no) and (total_eve_minutes >= 197.7) => churn=True. (12.0/1.0)

(number_customer_service_calls >= 4) and (total_eve_minutes <= 135) => churn=True. (12.0/4.0)

=> churn=False. (2926.0/91.0)

alex@seewald.at / alex.seewald.at

Why? So we can build better vehicles...

- Red Team - At mile 7.4, on switchbacks in a mountainous section, vehicle went off course, got caught on a berm and rubber on the front wheels caught fire, which was quickly extinguished.

- Vehicle 9 - The Golem Group - At mile 5.2, while going up a steep hill, vehicle stopped on the road, in gear and with engine running, but without enough throttle to climb the hill.

- Team ENSCO - Vehicle moved out smartly, but, at mile 0.2, when making its first 90-degree turn, the vehicle flipped.

alex@seewald.at / alex.seewald.at

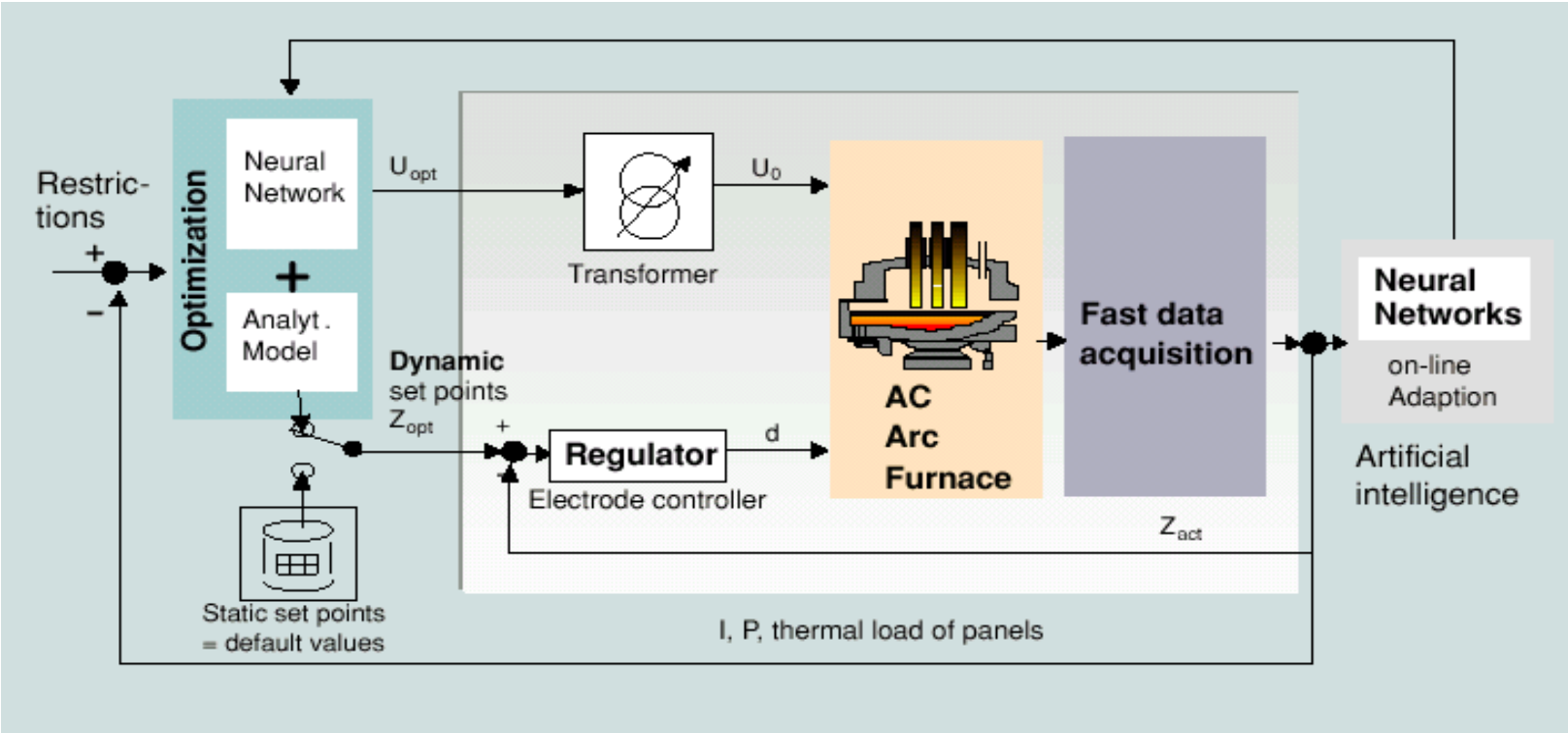
... or other robustly autonomous systems...

- Autopilot for one-person sailing
- Race-proven with many state-of-the-art AI and ML components.
- Human jargon like *gust*, *close-hauled*, *luff* as background knowledge, e.g.:

  If you are sailing close-hauled and there is a gust of wind then steer the boat a bit windward.

alex@seewald.at / alex.seewald.at

... save costs in industrial production..



- Optimization of melting process with neural network and analytical model: Steel production +6,0%; Energy consumption -3,1%

alex@seewald.at / alex.seewald.at

...or because we need a better Spam-Filter

Problem

- Spam : Nonspam = 17 : 1; 300 spams/day/user

Solution: State-of-the-Art System, adapted to our problem via ML

- Spam : Nonspam = 1 : 25
- Combined bayesian & human filter, trained on user's good/bad mails
- Deletes ~99.5% of incoming spam
- Few nonspam mails deleted (est. <1:600,000)
- Low maintenance (stable for >2 months)
- Currently tested by eight of my colleagues + myself: well received
- Compares well to expensive commercial systems which claim 1:2,000,000 chance of nonspam mail deleted.
- However, per-user models are not sufficiently scalable in practice, so much work remains to be done.

alex@seewald.at / alex.seewald.at

What is ML & DM?

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)

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)

alex@seewald.at / alex.seewald.at

ÖFAI Projects

Automated sleep staging (SIESTA, EU project)

- Sleep staging from EEG data; Spin-Off company: *The Siesta Group*

A Meta-Learning Assistant for Providing User Support in Machine Learning and Data Mining (METAL, ESPRIT-LTR EU project)

Biological Textmining (BioMinT, QLRI EU project)

Automated Quality Control for Industrial Printing (MONOTONE)

Meta-level learning for hybrid spamfilters

Commercial projects

- The Use of Machine Learning Methods for Quality Prediction in Steel Casting (+ Data Mining Library) for VÖEST-Alpine.
- Risk analysis for an Austrian insurance company.
- Sales forecasting for a large Austrian supermarket chain.
- Discovering Inefficiencies in the supply chain of an international firm.

alex@seewald.at / alex.seewald.at