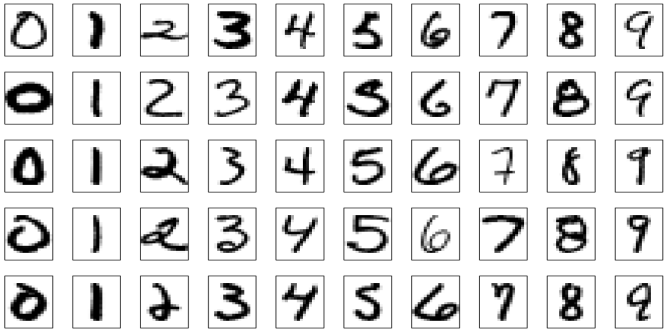
AI Methoden der Datenanalyse



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
- **Why** the digits? - Generating data for AI Methods
- **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
- **Motivation and Outlook**

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/

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 AI Meth.d.DA (509.916)

Jeden Mi 08:45-10:15, HS41; beginnend ab 9.3. = *heute*

Vorlesungsprüfung

29.06.2005 08:45-10:15 schriftlich, HS41. Alle Unterlagen erlaubt!

Laborübung AI Meth.d.DA (561.699)

- Ausgabe von Übungsaufgaben: ab 20.4. in VO & auf LVA-Webseite (s.u.)
- Abgabe der Lösungen in schriftlicher Form jeweils zum angegebenen Termin, ca. im 2-3 Wochen Rhythmus. Arbeit in Kleingruppen (3-4 Personen).
- Präsentation von ausgewählten Arbeiten und gemeinsame Diskussion.

Webseite zur LVA (passwort-geschützt)

http://www.oefai.at/~alexsee/lv/aida - User: **student** Passwort: **aiMda05**

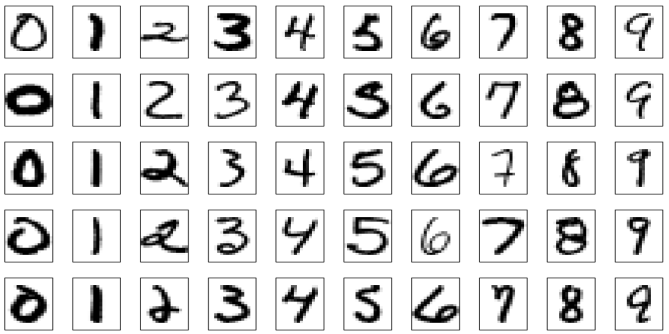
WICHTIG: Mitbelegen an der *Medizinischen Universität*!

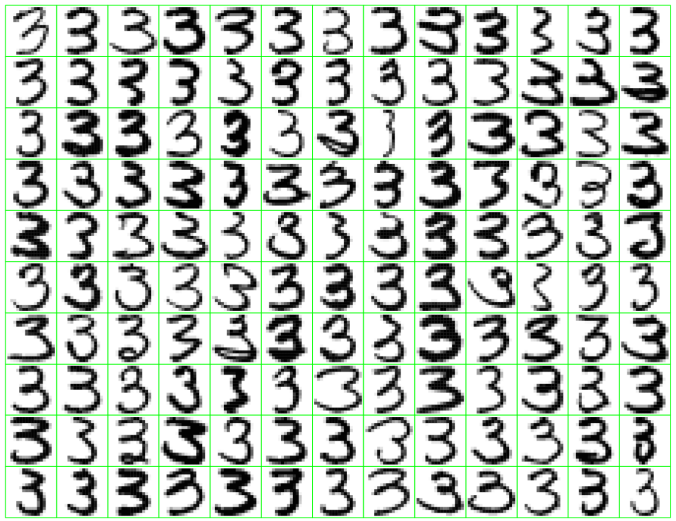
alex@seewald.at / alex.seewald.at

Why the digits?

US Postal Office Digit Dataset

- Optical Character Recognition for ZIP Codes in the 90ies as a learning task
- Scanned >10,000 digits from more than 500 different people.
- Digits were segmented, resized and resampled to 16x16 pixels with numeric gray values. Each pixel is represented by its own attribute.
- Ten classes: {0,1,2..,9}





A lecture hall full of students can generate a similarily large dataset in one minute. This enables us to validate findings on USPS and check whether it is representative.

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

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.
- Ian H. Witten & Eibe Frank (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.
- John W. Tukey (1977). *Exploratory Data Analysis*. Addison-Wesley

...

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)

09.03.2005	Introduction
16.03.2005	Introduction DA/ML, Basic Concepts, WEKA Demo
06.04.2005	Overview of learning systems
13.04.2005	Evaluation Basics
20.04.2005	Classic Data Analysis Methods
27.04.2005	Statistical Graphics
11.05.2005	Preprocessing Methods
18.05.2005	Feature construction, transformation and selection
25.05.2005	Ensemble Methods and Cost-Sensitive Learning
01.06.2005	Multi-View Learning
08.06.2005	Computational Learning Theory
15.06.2005	Learning Task Design & Evaluation
22.06.2005	Applications and Outlook
29.06.2005	Final exam (written)

alex@seewald.at / alex.seewald.at

What are AI Methods of Data Analysis?

Data Analysis comes in three varieties

- **Confirmatory** (can we confirm or reject a given hypothesis?)
- **Descriptive** (aims for compact description of data given a model class)
- **Exploratory** (explore data from many viewpoints; aims to find a plausible until a plausible "story" emerges)

For this lecture, we focus on **exploratory DA**. One viewpoint we will specially concentrate on is **Artificial Intelligence**. This viewpoint involves the active research fields of **Machine Learning** and **Data Mining.**

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."

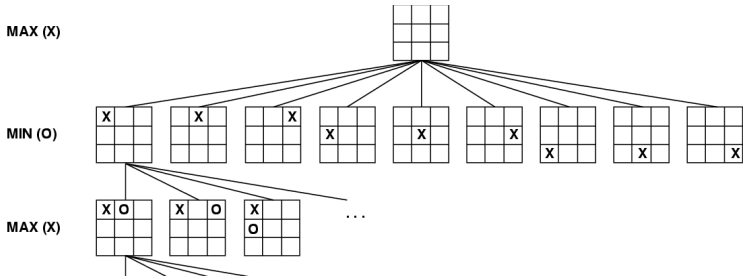
(Fayyad, Piatetsky-Shapiro & Smyth, 1996)

alex@seewald.at / alex.seewald.at

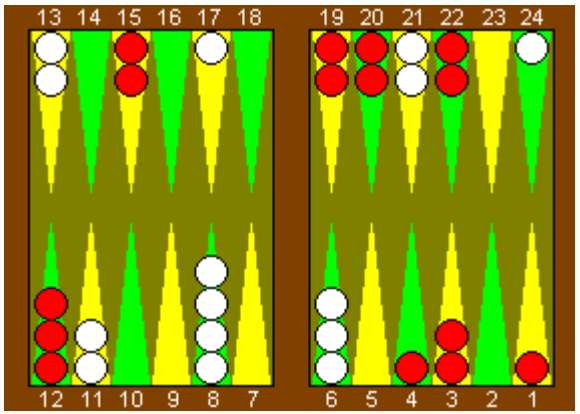
Can Machines Learn from Experience?

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

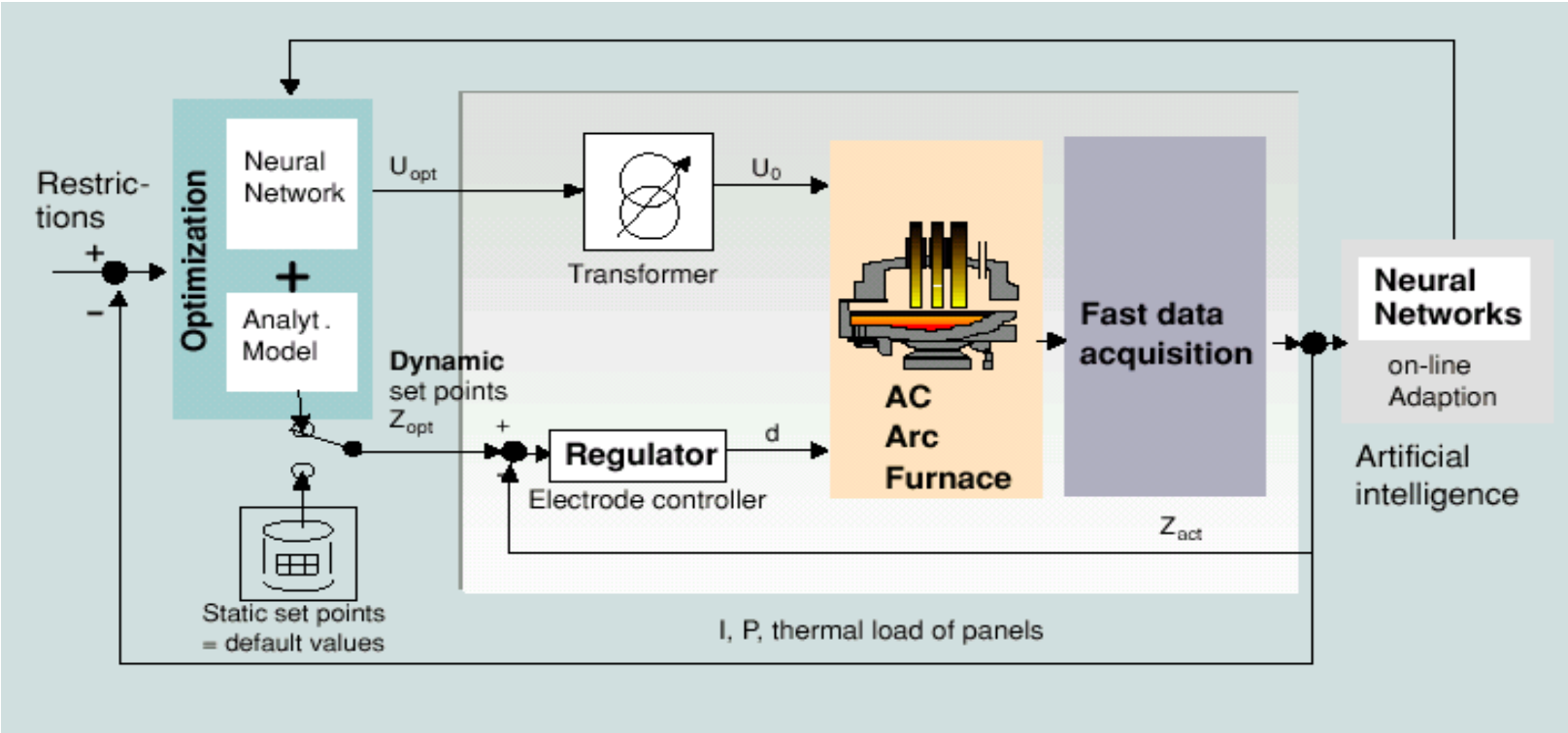
Why? To build autonomous systems...

- 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..



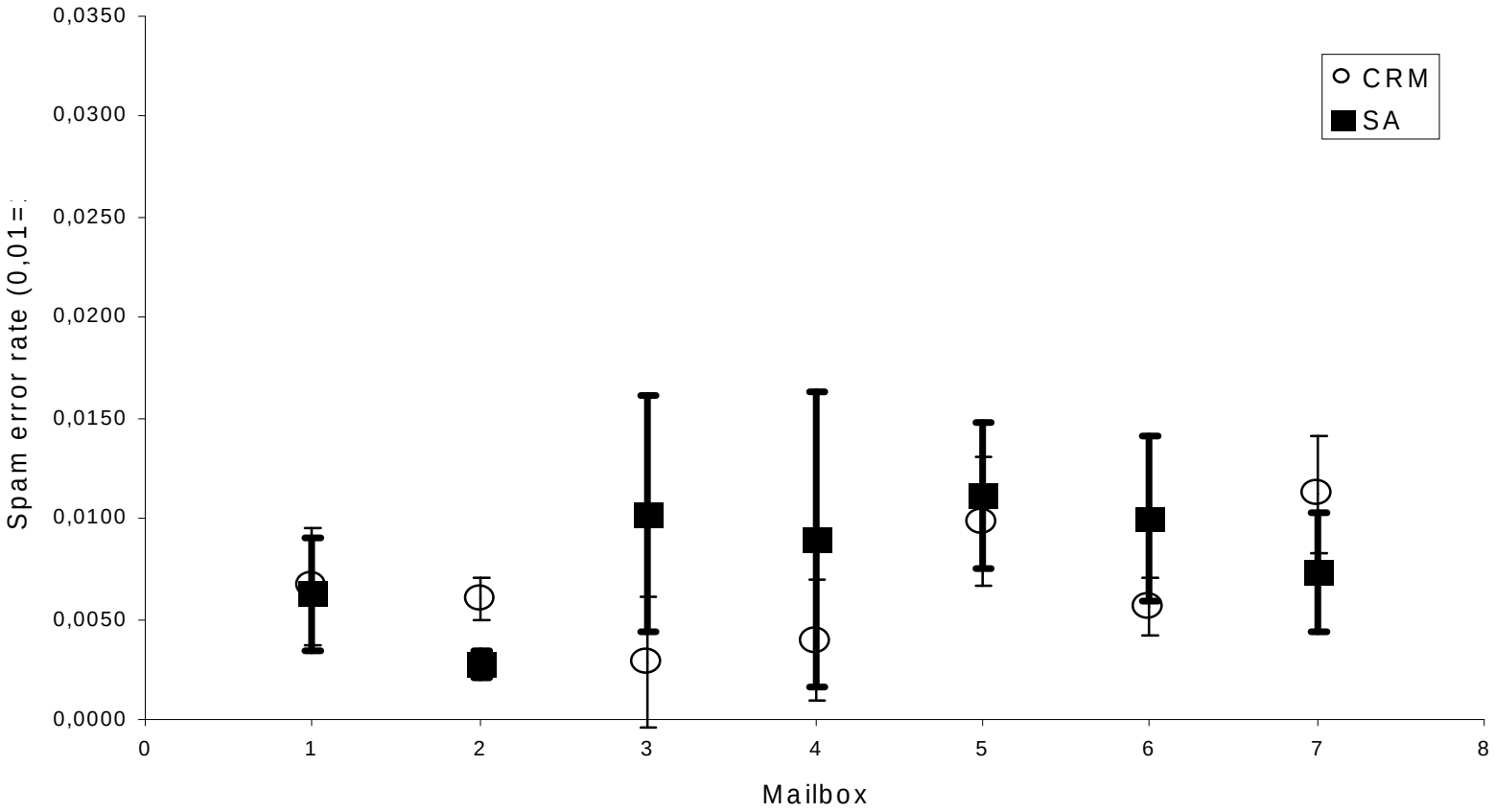
- Optimization of melting process with neural network and analytical model: Steel production +6,0%; Energy consumption -3,1%

alex@seewald.at / alex.seewald.at

..or improving Spam filters

Problem: 95% of mails are spams ~ 300 spams/day/users

...solved! Better than BrightMail, GMail, SpamAss. 3.0.2



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