

# Data Mining für Technische Anwendungen – Einführung

*PD Dr.-Ing. habil. Sven Tomforde*  
Prof. Dr. Bernhard Sick

Universität Kassel  
Fachbereich Elektrotechnik / Informatik  
Fachgebiet „Intelligent Embedded Systems“

WS 2017/2018



# Agenda

- Erläuterung des Begriffs Data Mining
- Inhalt der Vorlesung
- Organisatorisches
- Industrie- und Forschungsprojekte mit Bezug zu Data Mining im Fachgebiet „Intelligent Embedded Systems“ (IES)
- Sonstiges

# Erläuterung des Begriffs Data Mining

# Daten, Daten und noch mehr Daten – 1

- Europas Teleskopsystem VLBI (Very Long Baseline Interferometry) hat 16 Teleskope, von denen jedes während einer 25-tägigen Beobachtungsphase 1 GB/s an astronomischen Daten produziert. Quelle:  
[http://www.kdnuggets.com/data\\_mining\\_course/x1-intro-to-data-mining-notes.html](http://www.kdnuggets.com/data_mining_course/x1-intro-to-data-mining-notes.html) Stand: 21.10.2014
- France Telecom hat ein Entscheidungsunterstützungssystem, das ca. 30 TB an Daten beinhaltet. Quelle: [http://www.kdnuggets.com/data\\_mining\\_course/x1-intro-to-data-mining-notes.html](http://www.kdnuggets.com/data_mining_course/x1-intro-to-data-mining-notes.html) Stand: 21.10.2014
- In 1999 benötigte Google 1 Monat zur Indexierung von 50 Millionen Seiten. In 2012 benötigt Google dafür weniger als 1 Minute. Quelle:  
<http://www.internetlivestats.com/google-search-statistics/> Stand: 21.10.2014
- UC Berkeley schätzte 2003, dass im Jahr 2002 weltweit ca. 5 EB an neuen Daten produziert wurden (1 EB (Exabyte) =  $10^{18}$  Byte). Quelle:  
[http://www.kdnuggets.com/data\\_mining\\_course/x1-intro-to-data-mining-notes.html](http://www.kdnuggets.com/data_mining_course/x1-intro-to-data-mining-notes.html) Stand: 21.10.2014

# Daten, Daten und noch mehr Daten – 2

- Facebook produziert und analysiert täglich 25 TB an Log-Files zur Untersuchung des Nutzerverhaltens zur Verbesserung der Funktionen. <http://www.golem.de/0910/70585.html> (14. Okt. 2014)
- Facebook hatte im April 2014 rund 300 PB Speicherkapazität. Diese wächst mit rund 0,6 PB pro Tag an. (1PB (Petabyte)= $10^{15}$  Bit)  
<https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/> (21. Okt. 2014)
- CISCO schätzt, dass sich der Internet-Datenverkehr von 2010 bis 2015 vervierfachen wird auf knapp ein ZB pro Jahr (1 ZB (Zettabyte) =  $10^{21}$  Byte).  
Quelle: <http://newsroom.cisco.com/press-release-content;jsessionid=55F03FDC4A055F2B8675599135C4E670?type=webcontent&articleId=324003> Stand: 21.10.2014

# Daten, Daten und noch mehr Daten – 3

Ein neues Schlagwort: **Big Data**

- “Big data” refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future.

Quelle: National Science Foundation (NSF) proposal call

- “Big data” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data—i.e., we don’t define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes).

Quelle: McKinsey Global Institute Report on Big Data 2011



Intelligent  
Embedded Systems

# Daten, Daten und noch mehr Daten – 4



Die Speicher-  
und/oder  
Laufzeitkomplexität  
vieler Algorithmen  
der Datenanalyse ist  
nicht linear von der  
Datenmenge  
abhängig ...

Quelle: Bild: <http://de.desktopwallpaperhd.net/wallpapers-images-surfer-120485.html> Stand: 20.10.2014

# Daten, Daten und noch mehr Daten – 5

Dezimalpräfixe		Unterschied (gerundet)	Binärpräfixe	
Name (Symbol)	Bedeutung <sup>[G 1]</sup>		IEC-Name (IEC-Symbol)	Bedeutung
Kilobyte (kB) <sup>[G 2]</sup>	$10^3$ Byte = 1.000 Byte	2,40 %	Kibibyte (KiB) <sup>[G 3]</sup>	$2^{10}$ Byte = 1.024 Byte
Megabyte (MB)	$10^6$ Byte = 1.000.000 Byte	4,86 %	Mebibyte (MiB)	$2^{20}$ Byte = 1.048.576 Byte
Gigabyte (GB)	$10^9$ Byte = 1.000.000.000 Byte	7,37 %	Gibibyte (GiB)	$2^{30}$ Byte = 1.073.741.824 Byte
Terabyte (TB)	$10^{12}$ Byte = 1.000.000.000.000 Byte	9,95 %	Tebibyte (TiB)	$2^{40}$ Byte = 1.099.511.627.776 Byte
Petabyte (PB)	$10^{15}$ Byte = 1.000.000.000.000.000 Byte	12,6 %	Pebibyte (PiB)	$2^{50}$ Byte = 1.125.899.906.842.624 Byte
Exabyte (EB)	$10^{18}$ Byte = 1.000.000.000.000.000.000 Byte	15,3 %	Exbibyte (EiB)	$2^{60}$ Byte = 1.152.921.504.606.846.976 Byte
Zettabyte (ZB)	$10^{21}$ Byte = 1.000.000.000.000.000.000.000 Byte	18,1 %	Zebibyte (ZiB)	$2^{70}$ Byte = 1.180.591.620.717.411.303.424 Byte
Yottabyte (YB)	$10^{24}$ Byte = 1.000.000.000.000.000.000.000.000 Byte	20,9 %	Yobibyte (YiB)	$2^{80}$ Byte = 1.208.925.819.614.629.174.706.176 Byte

1. ↑ SI-Präfixe sind nur für SI-Einheiten standardisiert; Byte ist keine SI-Einheit  
 2. ↑ wird gelegentlich mit „KB“ abgekürzt  
 3. ↑ wird gelegentlich mit „KB“ abgekürzt, um den Unterschied zu „kB“ zu kennzeichnen (nicht standardisiert)

Quelle: <http://de.wikipedia.org/wiki/Byte> Stand: 20.10.2014

# Zitate

*Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data. A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining.*

[Fayyad, Piatetsky-Shapiro und Smyth 1996]

*Data doubles about every year, but useful information seems to be decreasing. The area of data mining has arisen over the last decade to address this problem.*

[Dunham 2003]

# Wesentliche Aufgabe des Data Mining

**Daten → Wissen**

# Anwendungsbeispiele – 1

## Netflix-Filmempfehlung:

- Kunden sollen auf sie personalisierte Filme bzw. Serien vorgeschlagen bekommen
- Netflix hat rund 40 freie Mitarbeiter zur objektiven Bewertung der Filme und Serien in festen Kategorien
- Erstellung von Datensätzen zu Tagesgewohnheiten (Wochentag und Uhrzeit)
- Aufgabe des Data Mining ist, zur richtigen Zeit das richtige Interessengebiet des Nutzers zu finden und darauf basierend passende Filme oder Serien vorzuschlagen.

Quelle: [www.wired.com](http://www.wired.com) Stand: 13. Oktober 2014

# Anwendungsbeispiele – 2

## Kreditrisiko:

- Bei der Vergabe von Krediten sollen Anträge automatisch bewertet und einer von drei Klassen zugeordnet werden: Genehmigung, zusätzliche Rückfrage oder Ablehnung.
- Beispielsweise sollen vor allem Personen mit mittlerem, gesichertem Einkommen einen Kredit erhalten, Personen ohne geregeltes Einkommen und ohne Vermögen dagegen nicht. Auch Personen mit hohem Vermögen und/oder hohem Einkommen sind als Kreditnehmer eher uninteressant.
- Aufgabe des Data Mining ist, auf der Basis historischer Daten einer Bank einen geeigneten Klassifikator für Anträge zu entwickeln.

# Anwendungsbeispiele – 3

## Marketing:

- Die Kündigungsrate von Handybesitzern (ca. 25% bis 30% pro Jahr) ist vielen Netzbetreibern zu hoch.
- Kunden sollen, ihrem „Wert“ für das Unternehmen entsprechend, rechtzeitig ein geeignetes, neues Angebot bekommen.
- Aufgabe des Data Mining ist, auf der Basis von allgemeinen Kundeninformationen und unter Berücksichtigung des Telefonierverhaltens des jeweiligen Kunden ein geeignetes (den Gewinn maximierendes) Angebot zu wählen.

# Anwendungsbeispiele – 4

## E-Commerce:

- Fast 80% der Kunden eines Online-Buchhändlers entscheiden sich noch an dem Tag für einen Kauf, an dem sie ein Produkt zum ersten Mal sehen.
- Kunden sollen personalisierte (d. h. auf sie zugeschnittene) Angebote erhalten, um sie auf bestimmte Produkte aufmerksam zu machen.
- Aufgabe des Data Mining ist, Interessensgebiete des Kunden zu erkennen und Kunden mit ähnlichem Kaufverhalten in einer Datenbank zu finden, um die von ihnen häufig gekauften Bücher anzubieten.

# Anwendungsbeispiele – 5

Opera | Amazon.de Empfehlungen | www.amazon.de/gp/yourstore/recu/ref=sv\_ys\_1

amazon.de

Marcus Amazon Angebote Gutscheine Verkäufer Hilfe

Kategorien Suche Alle

Main Amazon Ihre persönliche Seite Ihre Empfehlungen Verlassen Sie Ihre Empfehlungen Gutscheine Mein Profil Mehr dazu

Mein Amazon.de – Unsere Empfehlungen für Sie

Diese Empfehlungen basieren auf den von Ihnen gekauften Artikeln und weiteren Informationen.

Anzeigen Alle | Dauerreihenfolge | In Kürze

**1.**  **Reiseadapter, 1xPS20, Travel Adapter Schutzkontakt/USA, Japan**  
von **Amazon.de** (13. August 2007)  
Durchschnittliche Kundenbewertung: ★★★★★ (12)  
Auf Lager  
**Preis: EUR 4,25**  
EUR ab EUR 1,50

Gehen Sie zu [Amazon.de](#) Diesen Artikel bewerten  
Diesen Artikel haben wir empfohlen, weil Sie [Gummischwamm-Membran WSA-Schuhstop-Schuh-Spitze](#), gekauft haben. (Bitte ändern)

**2.**  **Universal WSA Fullcover für WSA-IP Spülkasten**  
von **Amazon.de** (13. August 2007)  
Durchschnittliche Kundenbewertung: ★★★★★ (12)  
Auf Lager  
**Preis: EUR 13,69**  
EUR ab EUR 11,00

Gehen Sie zu [Amazon.de](#) Diesen Artikel bewerten  
Diesen Artikel haben wir empfohlen, weil Sie [Gummischwamm-Membran WSA-Schuhstop-Schuh-Spitze](#), gekauft haben. (Bitte ändern)

**3.**  **Gummischwamm-Membranen Membrane für WSA oder Sanitop Spülkasten / Fullcover**  
von **Amazon.de** (13. August 2007)  
Durchschnittliche Kundenbewertung: ★★★★★ (12)  
Auf Lager  
**Preis: EUR 6,80**

Gehen Sie zu [Amazon.de](#) Diesen Artikel bewerten  
Diesen Artikel haben wir empfohlen, weil Sie [Gummischwamm-Membran WSA-Schuhstop-Schuh-Spitze](#), gekauft haben. (Bitte ändern)

**4.**  **Colord Waterstop 230 20001008 Schuhzement Gummizeder 75 ml (9)**  
Colom (5. Juni 2013)  
Durchschnittliche Kundenbewertung: ★★★★★ (12)  
Auf Lager  
**Preis: EUR 5,30 - EUR 10,39**

Gehen Sie zu [Amazon.de](#) Diesen Artikel bewerten  
Diesen Artikel haben wir empfohlen, weil Sie [Colord Mischzement 71429999010 Schuhzement Trans...](#), gekauft haben. (Bitte ändern)

**5.**  **Spülkastendichtungen, 25285**  
von **Santop-Wingertsdorf** (3. April 2008)  
Durchschnittliche Kundenbewertung: ★★★★★ (12)  
Auf Lager  
**Preis: EUR 5,23**  
EUR ab EUR 4,36

Anzeigen von Ihren Empfehlungen | In den Wunschzettel | Auf meinen Wunschzettel



Intelligent  
Embedded  
Systems

# Anwendungsbeispiele – 6

Kolibree ist eine seit Sommer 2014 erhältliche elektrische Zahnbürste. Sie erfasst die Bewegungen des Nutzers, bewertet die Gründlichkeit des Zähneputzens und überträgt die ermittelten Daten via Bluetooth auf ein Smartphone.

So könnten beispielsweise Eltern das Zahnpflegeverhalten ihrer Kinder überwachen.



Quellen: [http://www.focus.de/digital/multimedia/ces-2014-kolibree-clevere-zahnbuerste\\_id\\_3525781.html](http://www.focus.de/digital/multimedia/ces-2014-kolibree-clevere-zahnbuerste_id_3525781.html) Stand: (20.10.2014)

Bild: <http://www.kolibree.com/en/> Stand: 20.10.2014

# Anwendungsbeispiele – 7

**Schrei-Analyse für Babys**

Foto: Reer

**Warum weint mein Baby?** Diese Frage verzweifelter Eltern will mit digitaler Technik der Handheld Why-Cry des Herstellers Reer beantwortet. Das Gerät analysiert Frequenz, Intensität sowie Intervall des Schreiens und erkennt eine der Kategorien Hunger, Langeweile, Unwohlsein/Schmerz, Müdigkeit oder Stress/Kolik. Der spanische Arzt Manuel Pardos Algás hatte dazu das Weinen von 500 Babys aller Nationalitäten studiert und ein wissenschaftliches Modell entwickelt. pk

## Signalverarbeitung

(Bericht von 2004;  
inzwischen später bei  
Neckermann online für  
€ 29.– erhältlich.)

# Anwendungsbeispiele – 8

## TouchID auf dem iPhone

Mit dem eigenem Fingerabdruck das Smartphone entsperren, (In-)App-Käufe verifizieren und vertrauliche Apps starten. Dies ist mit dem in den Home-Button integrierten Fingerabdrucksensor ab dem iPhone 5S möglich. Merkmale und Eigenschaften des Fingers werden dabei gespeichert und bei jedem Aufruf analysiert. Stimmt der Fingerabdruck überein, wird die Freigabe erteilt. Dabei sollen auch subepidermale (tiefere) Hautschichten analysiert werden, so dass bspw. ein künstlicher Fingerabdruck nicht als Verifikationsmöglichkeit in Frage kommt.



Quelle: Bild und Inhalt <http://www.macwelt.de/news/Technik-im-iPhone-5S-Touch-ID-erklaert-8202455.html> (13. Oktober 2014).

# Anwendungsbeispiele – 9

## Weitere Beispiele in den Bereichen:

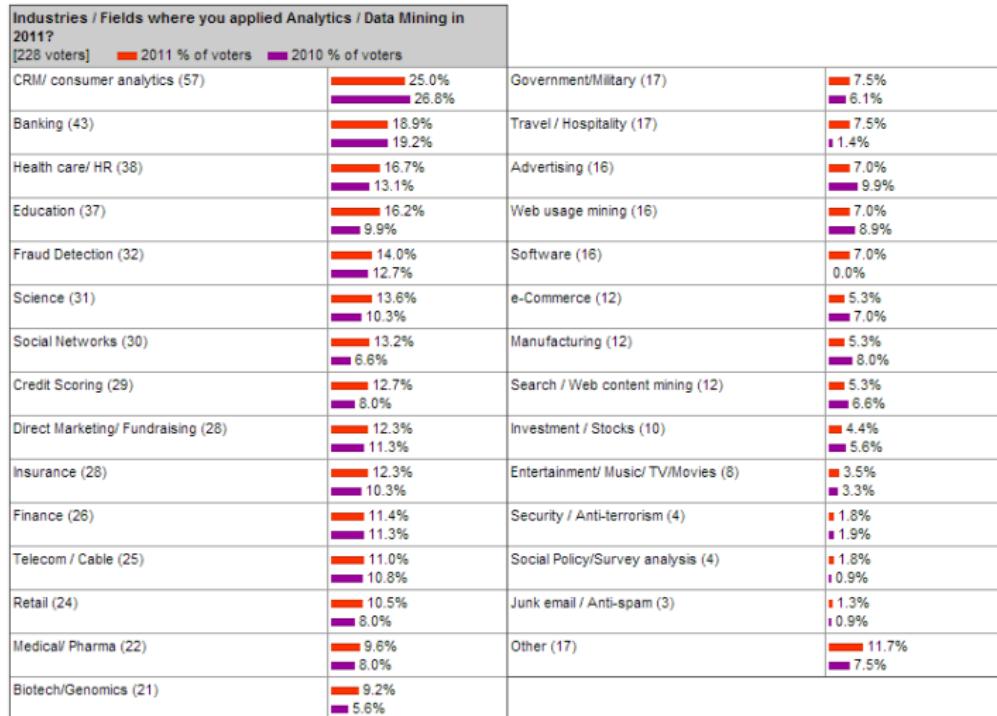
- Bildverarbeitung (z. B. Image Retrieval)
- Rechnernetze (z. B. Intrusion Detection)
- Medizin (z. B. Erkennung von Krankheiten in genetischer Information)
- Chemie (z. B. Drug Design)
- ...

# Anwendungsbeispiele – 10

**Allgemein wird Data Mining in Bereichen benötigt,**

- die wissensbasierte Entscheidungen erfordern,
- in denen konventionelle (z. B. statistische) Methoden nicht zu optimalen Ergebnissen führen,
- wo Daten verfügbar und in ausreichender Menge vorhanden sind.

# Anwendungsbeispiele – 11



Quelle: <http://www.kdnuggets.com/polls/2011/industries-applied-anaytcs-data-mining.html> (13. Oktober 2014)

# Was ist Data Mining?

## Definition des Begriffs (vorläufig):

*Data mining is the process of automatically extracting valid, novel, potentially useful, and ultimately comprehensible information from large databases.*

[Fayyad, Piatetsky-Shapiro, Smyth und Uthurusamy 1996]

*Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.*

[Hand, Mannila, Smyth 2001]

# Was ist Data Mining nicht?



Copyright © 2000 United Feature Syndicate, Inc.  
Redistribution in whole or in part prohibited

Torture the data until they confess!!!

# Charakteristika des Data Mining – 1

Data Mining ist ein vielschichtiger Prozess:



[Embrechts, Szymanski und Sternickel 2004]

# Charakteristika des Data Mining – 2

Data Mining erfordert vielseitige Fähigkeiten:

*Data mining is an interdisciplinary science ranging from the domain area and statistics to information processing, database systems, machine learning, artificial intelligence and soft computing.*

[Embrechts, Szymanski und Sternickel 2004]

# Durch welche Gebiete ist Data Mining beeinflusst?

- Mustererkennung / Statistik / Zeitreihenanalyse
- Artificial Intelligence / Expertensysteme
- Soft Computing / Machine Learning
- Datenbanken
- High Performance Computing / paralleles bzw. verteiltes Rechnen
- Datenvisualisierung
- ...

# Wichtigste Verwandte Gebiete

- Statistik: z. B. Testen von Hypothesen (eher „mathematischer“ als Data Mining)
- Machine Learning: z. B. Verbesserung der Performanz lernender Agenten (eher „heuristischer“ als Data Mining)

Data Mining: integriert verschiedene Ansätze, auch Aufbereitung und Bereinigung von Daten, Speicherung von Daten und effizienter Zugriff sowie Visualisierung von Ergebnissen

# Beispiele für Aufgaben von Algorithmen des Data Mining

- Klassifikation: Bestimmung einer Klassenzugehörigkeit mit Hilfe von Beispieldaten mit bekannter Klassenzugehörigkeit
- Clustering: Finden von „natürlichen“ Gruppen von Daten in Beispieldaten ohne bekannte Klassenzugehörigkeit
- Assoziationen: Finden von Gemeinsamkeiten von Daten
- Ausreißererkennung: Finden von Anomalien in Daten
- ...

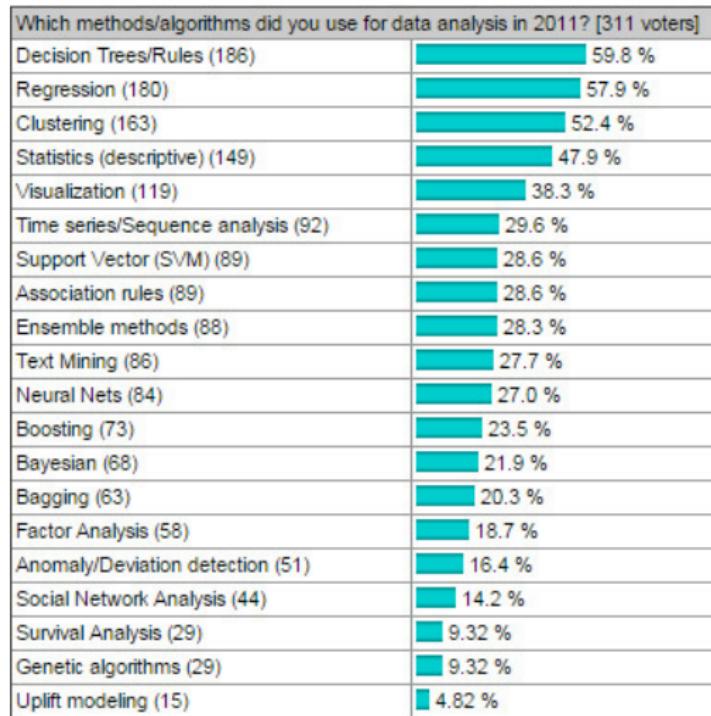
**Ausgabe eines Data Mining Prozesses ist meist nicht Teilmenge einer Datenbank, sondern Ergebnis einer Analyse des Datenbankinhalts!**

# Inhalt der Vorlesung

# Inhalt der Vorlesung

- Grundlagen des Data Mining
- Datenvorverarbeitung und -aufbereitung
- Merkmalsselektion und Hauptkomponentenanalyse
- Grundlegende Algorithmen für Clustering
- Grundlegende Algorithmen für Klassifikation
- Radiale-Basisfunktionen-Netze
- Support Vector Machines
- Generative probabilistische Klassifikatoren
- Bayes-Netze
- Ensembletechniken
- Ausblick auf weitere Verfahren

# Techniken des Data Mining in Anwendungen

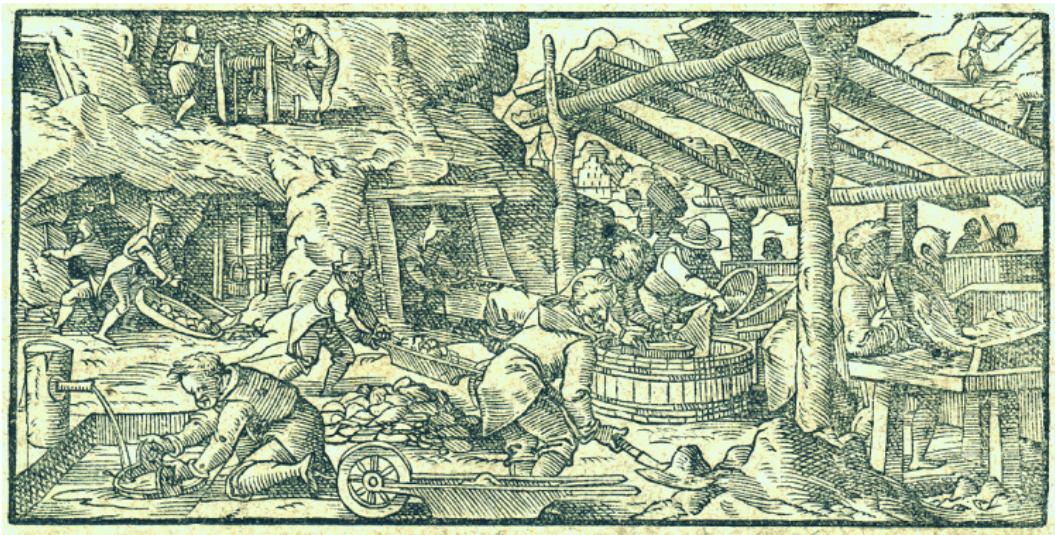


# Warnung – 1

*In practice, a large portion of the applications effort can go into properly formulating the problem (asking the right question) rather than optimizing the algorithmic details of a particular method.*

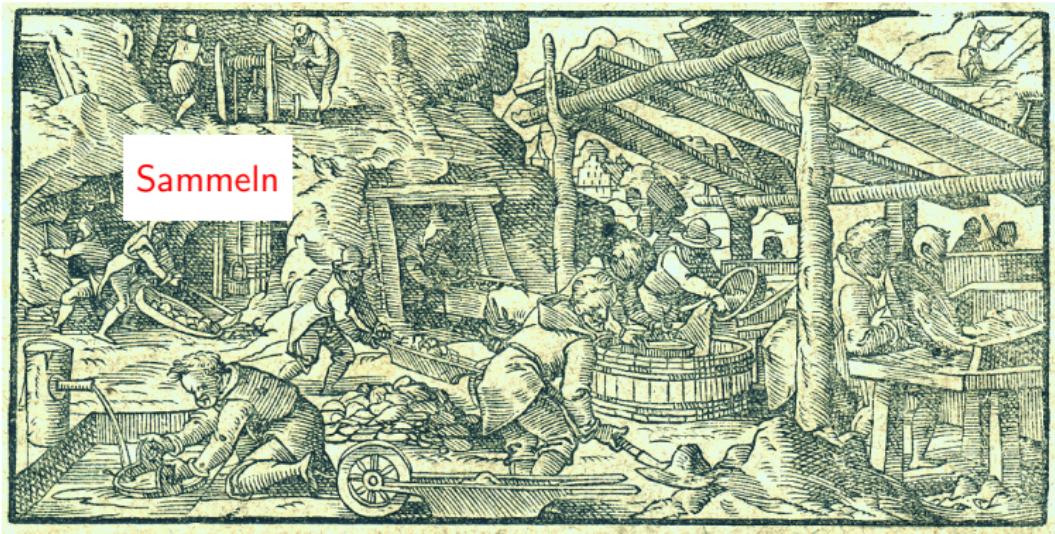
[Fayyad, Piatetsky-Shapiro und Smyth 1996]

# Warnung – 2



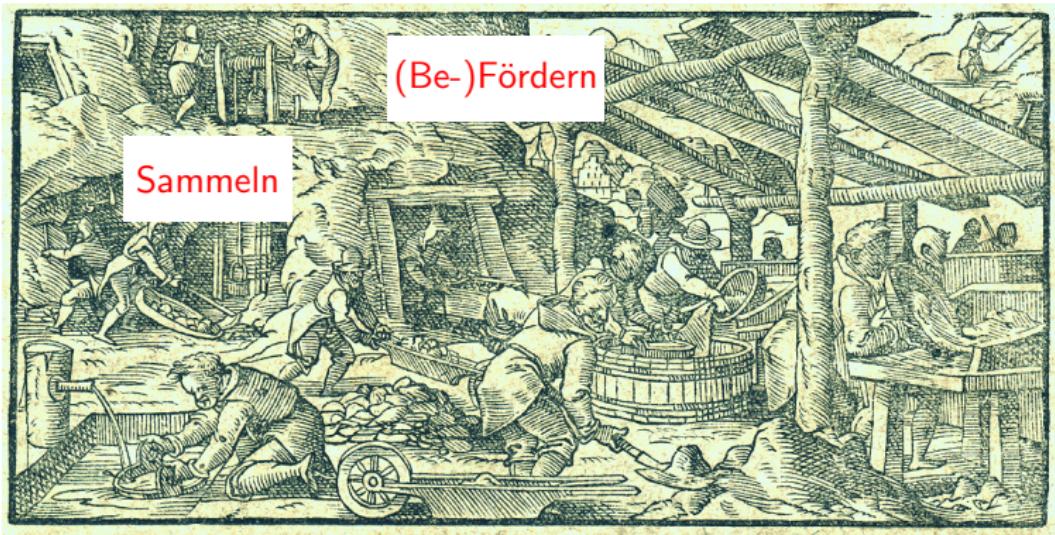
[Schedel 1493]

# Warnung – 2



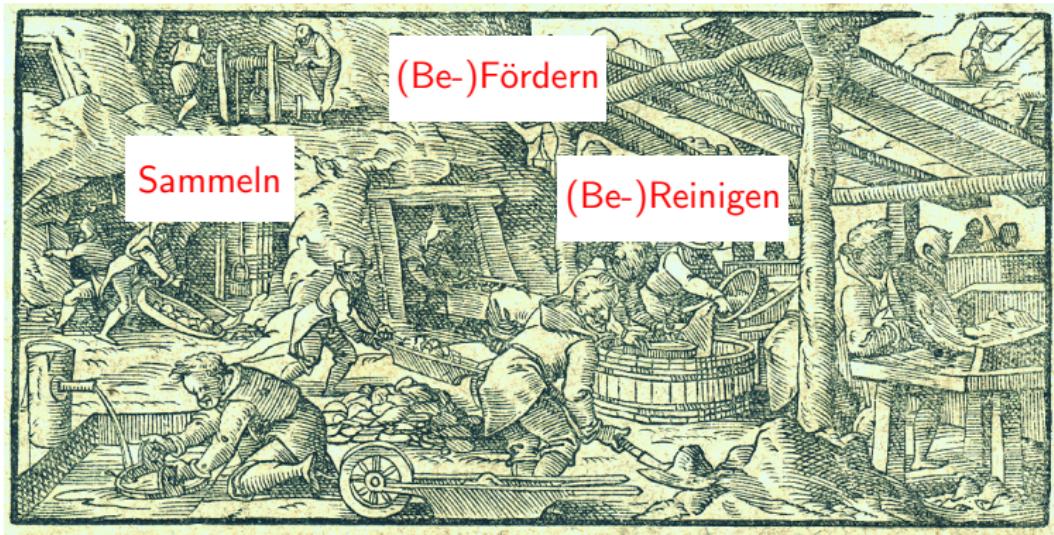
[Schedel 1493]

# Warnung – 2



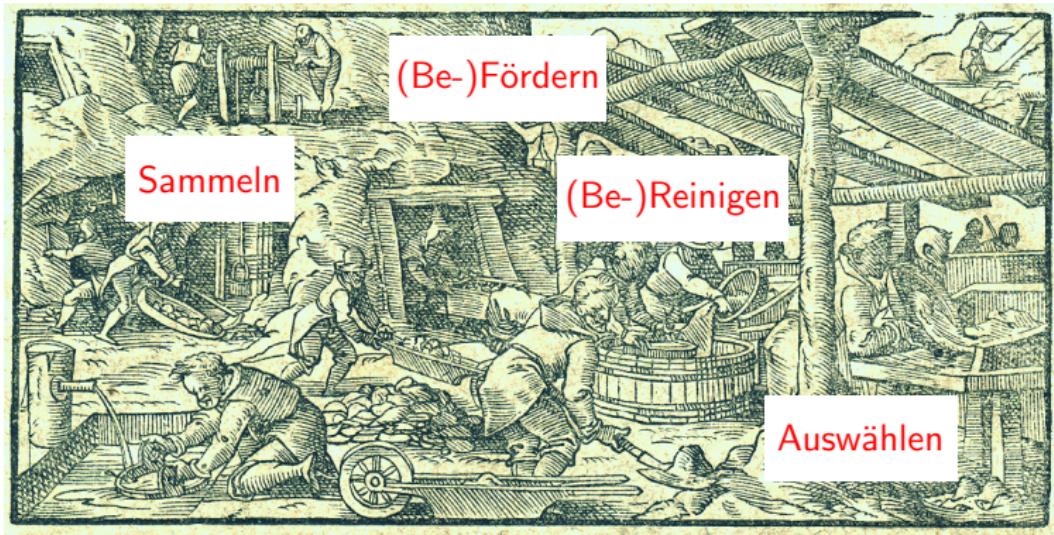
[Schedel 1493]

# Warnung – 2



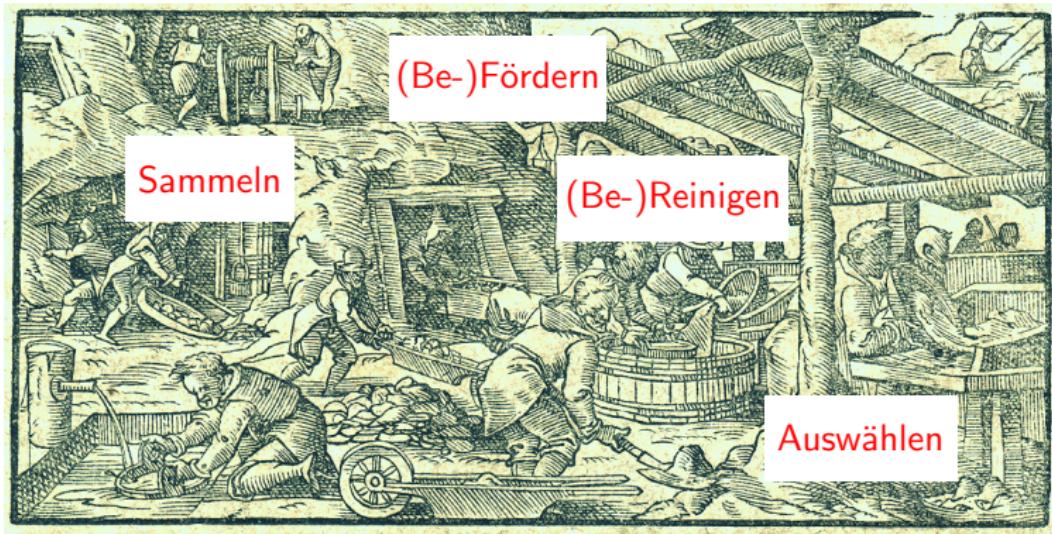
[Schedel 1493]

# Warnung – 2



[Schedel 1493]

# Warnung – 2



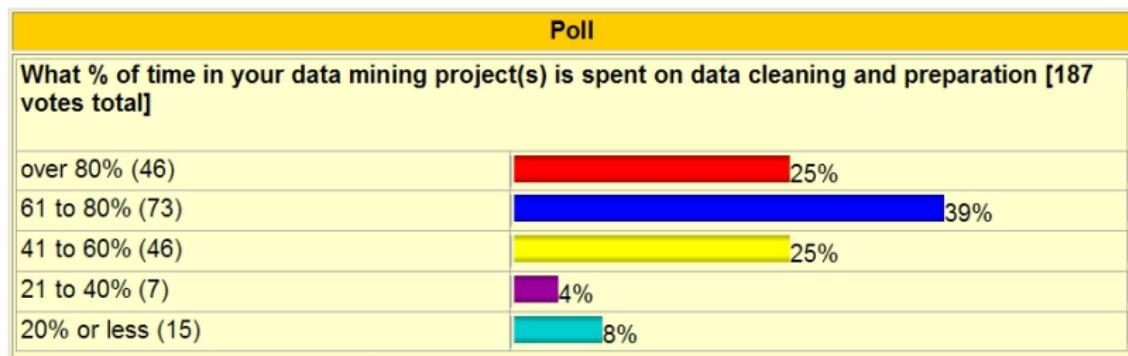
[Schedel 1493]

... und **dann** erst Weiterverarbeiten!!!



Intelligent  
Embedded  
Systems

# Zeitaufwand beim Data Mining



Oct.2003 [Quelle: KDnuggets]

# Nicht Inhalt der Vorlesung – 1



# Nicht Inhalt der Vorlesung – 2

- paralleles oder verteiltes Data Mining
- Datenbankaspekte
- Datenvisualisierung und Darstellung der Ergebnisse des Data Mining
- Webmining, Textmining, Techniken für kategorische Daten (z. B. Assoziationsregeln)
- Inhalt der Vorlesungen Soft Computing oder Pattern Recognition (leichte Überschneidungen)
- Datenschutz- und Ethikaspekte

# Organisatorisches

# Organisatorisches

- Namen und Adressen
- Webseiten
- Termine
- Übungen
- Prüfungen
- Kombination mit anderen Vorlesungen
- Praktika / Diplomarbeiten / Seminar

# Namen und Adressen

## Vorlesung:

- **Dozent:** PD Dr.-Ing. habil. Sven Tomforde
- **Adresse:** Fachgebiet IES, WA, Zimmer 0306 (Erdgeschoß)
- **Sprechstunde:** mit Voranmeldung per email (feste Zeit für WS 17/18 noch nicht festgelegt)
- **e-mail:** stomforde@uni-kassel.de

## Übung:

- Janosch Henze, MSc, janosch.henze@uni-kassel.de
- Jens Schreiber, MSc, jens.schreiber@uni-kassel.de

# Webseiten

- Vorlesungsunterlagen (Folien) in Moodle, Passwort "DhieIESK!"
- Webseite Fachgebiet [www.ies-research.de](http://www.ies-research.de)

# Termine

- Vorlesung:: Mi, 8:30 – 10:00 Uhr
- Übung: Do, 8:30 – 10:00 Uhr
- Übungen begleiten die Vorlesung und vertiefen die dortigen Inhalte.  
Teilnahme ist nicht verpflichtend - aber eindeutig empfohlen!
- Übungen und Vorlesungen im Labor des Fachgebiets, Raum 0303c

# Übungen

- Rechnerübungen
- im Labor des Fachgebiets (Raum 0303c)
- mit Jupyter Notebooks
- erster Termin ist Donnerstag, 26.10.

# Prüfungen

Prüfung / Alternativen gemäß Modulplan:

- schriftliche Prüfung (120 min), vorlesungsfreie Zeit
- mündliche Prüfung (ca. 25 min.), Semesterende

Festlegung muss jetzt erfolgen: **mündliche Prüfung**.

Prüfungsbonus:

- Semesteraufgabe, begleitend zu “normalen” Übungsaufgaben
- Ziel: Data Mining Prozess eigenständig ausführen
- Am Ende: Wettbewerb der Gruppen
- Bonus: Erfolgreiche Teilnahme → 3 Minuten Möglichkeit zur Erläuterung in der Prüfung

# Kombination mit anderen Vorlesungen

Welche Themen passen zu Data Mining?

- Signalverarbeitung, Bildverarbeitung
- Datenbanken
- Statistik, Zeitreihenanalyse, Datenanalyse
- Soft-Computing
- ...

# Lehre am Fachgebiet – 1

## Vorlesungen und Praktika (Bachelor):

- Einführung in C
- Stochastik in der technischen Anwendung/Grundlagen der Stochastik
- Intelligente technische Systeme
- Praktikum Intelligente eingebettete Systeme
- Soft Computing
- Echtzeitsysteme
- Praktikum Intelligente humanoide Roboter

## Vorlesungen und Praktika (Master):

- Pattern Recognition
- Temporal and Spatial Data Mining
- Organic Computing

# Lehre am Fachgebiet – 2

Das Anwendungsgebiet *Embedded Intelligence* beschäftigt sich mit Grundlagen des Maschinellen Lernens und der Signalverarbeitung in technischen Anwendungen (z. B. Robotik, Grafiktablets, Smartphones).

- Intelligente Technische Systeme (Sick): Basis des Anwendungsgebiets (empfohlen)
- Soft Computing (Sick/Tomforde)
- Computational Intelligence in der Automatisierung (Kroll, FB 15 Maschinenbau)
- Echtzeitsysteme (Sick)
- Signalverarbeitung mit Mikroprozessoren I (Börcsök)
- Digitale Systeme (Zipf)
- Autonome mobile Roboter (Geihs)
- Grundlagen der Regelungstechnik (Stursberg)

# Lehre am Fachgebiet – 3

- Data Mining für Technische Anwendungen (Tomforde/Sick)
- Knowledge Discovery (Stumme)
- Praktikum Java Code-Camp Context Awareness I (David)
- Praktikum Intelligente Eingebettete Systeme (Sick)
- Praktikum Kooperative verteilte Robotersysteme (Geihs)
- Praktikum Intelligente humanoide Roboter (Sick)

# Projekte, Seminare, Bachelorarbeiten

- **Projekte:** jederzeit, algorithmen- oder anwendungsorientiert (z. B. Roboterkooperation, Analyse von Graphiktablettdaten, Machine Learning, Deep Belief Networks, ...) wichtig: Aushänge/Web, Gespräch mit Mitarbeitern
- **Seminare:** auch in diesem Semester, Vorbesprechung siehe Aushänge
- **Bachelorarbeiten:** auch ab sofort, Algorithmen- oder Anwendungs-orientiert

Fragen ...

... zur Organisation?

# Projekte mit Bezug zu Data Mining im Fachgebiet IES

# Definition IES – 1

## Definition „Eingebettetes System“ nach Wikipedia:

Der Ausdruck *eingebettetes System* (*embedded system*) bezeichnet einen elektronischen Rechner oder auch Computer, der in einen technischen Kontext eingebunden (eingebettet) ist.

Dabei hat der Rechner entweder die Aufgabe, das System, in das er eingebettet ist, zu steuern, zu regeln oder zu überwachen. Oder der Rechner ist für eine Form der Daten- bzw. Signalverarbeitung zuständig.

Eingebettete Systeme verrichten – weitestgehend unsichtbar für den Benutzer – den Dienst in einer Vielzahl von Anwendungsbereichen.

# Definition IES – 2

## Anwendungsbereiche eingebetteter Systeme:

- Geräte der Medizintechnik
- Waschmaschine
- Flugzeuge
- Kraftfahrzeuge
- Kühlschränke
- Fernseher
- Roboter
- Mobiltelefone
- ...

# Definition IES – 3

## **Definition „Intelligentes System“:**

Ein Computersystem kann als *intelligent* bezeichnet werden, wenn es dazu fähig ist, seine eigene Leistung zu verbessern oder mindestens ein akzeptables Leistungsniveau unter Einfluss auftretender Ungewissheiten aufrechtzuerhalten.

# Definition IES – 4

## Anwendungsbereiche intelligenter Systeme:

- Steuerungs- und Regelungssysteme
- Prozessüberwachungssysteme (z. B. Systeme zur Qualitätskontrolle)
- Prozessoptimierungssysteme
- Datenbanken – Knowledge Discovery & Data Mining
- Bildbearbeitung / Bildverarbeitung
- „intelligente“ Suchmaschinen
- Fehlersuche in Software
- „intelligente“(teil)-autonome Robotersysteme
- Empfehlungen auf Webseiten
- Notbremsassistsysteme in PKWs
- u. v. m.

# Definition IES – 5

Das Fachgebiet *Intelligente Eingebettete Systeme (intelligent embedded systems, IES)* liegt an der Schnittstelle beider Bereiche.

## Schwerpunkte in der Lehre:

- Brückenschlag zwischen Technischer Informatik und Computational Intelligence

## Schwerpunkte in der Forschung:

- Autonomic and Organic Computing, Technical Data Analytics, Anwendungen z. B. im Bereich Automobil/Verkehr, Energie, Biometrie

# Data Mining Projekte in der Arbeitsgruppe IES

- Vorhersage von Kundenverhalten
- Intrusion Detection
- Unterschriftenverifikation

# Vorhersage von Kundenverhalten – 1

Eigene Untersungen bei Einführung der neuen Mercedes E-Klasse um 2002



- potenzielle Kunden sollten direkt angeschrieben werden und einen Prospekt erhalten (Direct-Mailing)
- unter allen deutschen Haushalten waren die vielversprechendsten für ein solches Anschreiben auszuwählen
- ein Modell (Prädiktor) sollte dazu auf der Basis mikro-geographischer Daten das Kundenverhalten (Kauf / kein Kauf) prognostizieren

# Vorhersage von Kundenverhalten – 2

Beispiele für mikro-geographische Daten (Merkmale):

Feature name	Explanation
ORTSGRKL	size of the city
STATUS	status of the residents w.r.t. education and income
ALTERSTD	average age of the heads of the households
RISIKO	credit risk information
P_DICHTE	density of cars
MERCEDES	proportion of MERCEDES-brand cars
PROHH	average purchasing power per household
FLUKT	fluctuation in the micro-geographical unit

insgesamt 47 mögliche Merkmale

Nebenbedingung: Minimierung der Zahl der für die Prognose erforderlichen Merkmale!

# Vorhersage von Kundenverhalten – 3

Voraussetzungen:

- 18 000 Datensätze aus früheren Kampagnen  
(47 Merkmale + Kaufentscheidung ja / nein)

Lösung:

- Neuronales Netz (Radiales-Basisfunktionen-Netz) zur Prognose
- Evolutionärer Algorithmus zur Selektion der geeigneten Merkmale und zur Optimierung der Architektur des Netzes

# Vorhersage von Kundenverhalten – 4

Ergebnisse:

Modellparadigma	Klassifikationsgüte
Zufallsauswahl	1
CHAID	1.5
C5	1.6
Mehrlagiges Perzeptron	1.8
Radiales-Basisfunktionen-Netz (manuell optimiert)	2.6
Radiales-Basisfunktionen-Netz (optimiert mit Evolutionärem Algorithmus)	3.45

# Intrusion Detection – 1

IT-Sicherheit – aktuell eines der wichtigsten Themen!!!



Quelle: <http://cdn8.howtogeek.com/wp-content/uploads/2012/08/windows-firewall-prompt.png> Stand: 21.10.2014



# Intrusion Detection – 2

## Intrusion (Angriff)

- ... ist eine böswillige Verletzung der (impliziten oder expliziten) Sicherheitspolitik durch eine nicht autorisierte Person.

## Intrusion Detection (Angriffserkennung)

- ... beschäftigt sich mit der Entwicklung von Methoden zur automatischen Erkennung von Angriffen auf Rechnersysteme.

## Intrusion Detection System (IDS)

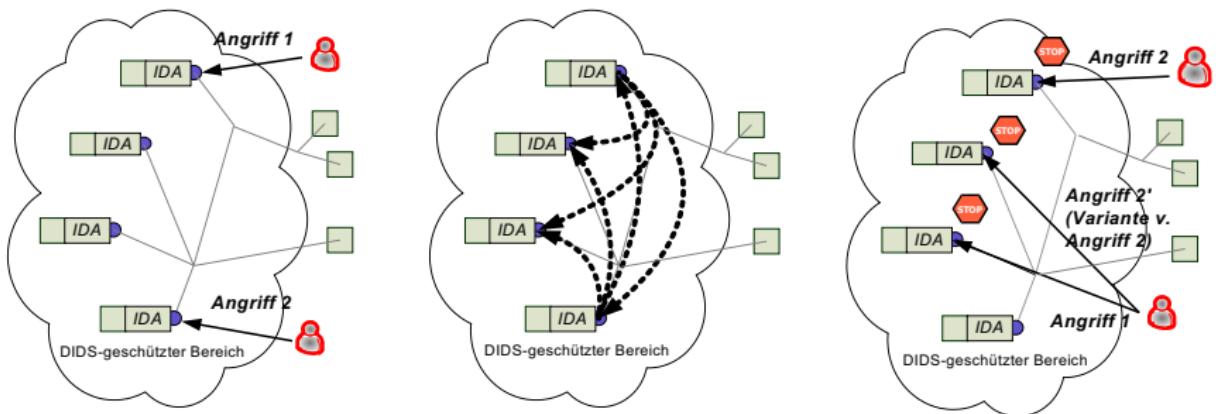
- ... soll Rechnersysteme und Netzinfrastrukturen mit dem Ziel der Erkennung von Einbrüchen und Systemmissbrauch überwachen. Dabei sollte ein IDS möglichst schnell, zuverlässig und mit möglichst wenig menschlichem Eingreifen arbeiten.

# Intrusion Detection – 3

1. Messung verdächtiger Daten  
(z.B. potentielle Angriffe)

2. Austausch von Information, Wissen und Erfahrung unter Berücksichtigung des erworbenen Vertrauens in andere ID-Agenten:  
 • aktuelle Situation (Angriffe, Angreifer),  
 • erlernte Regeln (Erkennung neuer Angriffe),  
 • Reputation (Vertrauen in weitere ID-Agenten)

3. Sensibilisierung des gesamten DIDS-geschützten Bereiches bezüglich der aktuellen Bedrohungssituation und bezüglich neuartiger Angriffe und Angriffsvarianten



# Intrusion Detection – 4

Untersuchungen aus 2003:

- Netzbasierte Missbrauchserkennung (noch nicht: hostbasierte Erkennung, Anomalieerkennung)
- Angriffstypen: z. B. Nmap, Ipsweep, Portsweep, Guest, Dict, Back, Warezclient, Satan
- Erkennung rein auf der Basis von statistischen Informationen aus TCP- und IP-Headern (137 mögliche Merkmale)
- Tests mit Benchmarkdaten der *Defense Advanced Research Projects Agency* (10 GB Daten (tcpdump), 38 Angriffstypen, zwischen ca. 100 und 100000 Beispiele je nach Angriffstyp)
- Merkmalsselektion unter anderem mit Evolutionären Algorithmen

# Intrusion Detection – 5

Attack Type	Back	Dict	Nmap	Portsweep
<b>Radial Basis Function Network (RBF)</b>				
E vali in %	0.42%	0.00%	0.00%	1.39%
FA vali in %	0.20%	0.00%	0.00%	1.61%
MA vali in %	0.64%	0.00%	0.00%	1.15%
<b>Multilayer Perceptron (MLP)</b>				
E vali in %	1.86%	0.54%	0.00%	1.41%
FA vali in %	2.53%	0.86%	0.00%	1.68%
MA vali in %	1.14%	0.17%	0.00%	1.10%
<b>NEFCLASS (NC)</b>				
E vali in %	0.62%	2.54%	0.00%	1.39%
FA vali in %	0.80%	4.57%	0.00%	1.66%
MA vali in %	0.43%	0.17%	0.00%	1.10%
<b>Decision Tree (DT)</b>				
E vali in %	0.73%	1.23%	0.00%	1.41%
FA vali in %	0.26%	1.00%	0.00%	1.68%
MA vali in %	1.21%	1.50%	0.00%	1.10%

# Intrusion Detection – 6

Attack Type	Back	Dict	Nmap	Portsweep
<b>Classifying Fuzzy-k-means (CFKM)</b>				
E vali in %	0.83%	1.23%	0.61%	1.92%
FA vali in %	1.07%	2.14%	1.14%	1.66%
MA vali in %	0.57%	0.17%	0.00%	2.20%
<b>Support Vector Machine (SVM)</b>				
E vali in %	0.52%	0.15%	<b>0.00%</b>	1.40%
FA vali in %	0.87%	0.29%	0.00%	1.00%
MA vali in %	0.14%	0.00%	0.00%	1.85%
<b>Nearest Neighbor (kNN)</b>				
E vali in %	0.55%	0.38%	<b>0.00%</b>	1.43%
FA vali in %	0.47%	0.14%	0.00%	1.00%
MA vali in %	0.64%	0.67%	0.00%	1.90%

E: Error, FA: False Alarms, MA: Missing Alarms, vali: validation data

# Beispiel: Unterschriftenverifikation – 1

Im alltäglichen Leben wird Sicherheit immer wichtiger:

- **Zugangskontrolle:** Sicherheitsbereiche, sicherheitskritische Dienste, Rechenanlagen, ...
- **Amtliche Personenkontrolle:** Grenzkontrollen, Polizeifahndungen, ...
- **Elektronischer Zahlungsverkehr:** Banküberweisungen, Kreditkartenzahlungen, ...
- ...

Personen müssen sich authentifizieren!

# Beispiel: Unterschriftenverifikation – 2

## Arten der Authentifizierung:

- „**What you have**“ – Besitz

Beispiele: Schlüssel, SmartCards, ...

- „**What you know**“ – Wissen

Beispiele: Passwörter, PINs, ...

- „**What you are**“ – Biometrische Charakteristika

Beispiele:

- ▶ physiologische Kennzeichen: z. B. Gesicht, Iris, Retina, Ohr, Fingerabdruck, Handgeometrie, Handflächenabdruck
- ▶ verhaltensbasierte Kennzeichen: z. B. Sprache, Gestik, Gang,

## Unterschrift

# Beispiel: Unterschriftenverifikation – 3

# Beispiel: Unterschriftenverifikation – 3

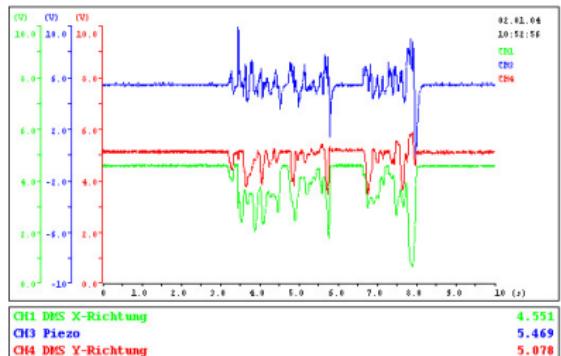


# Beispiel: Unterschriftenverifikation – 3



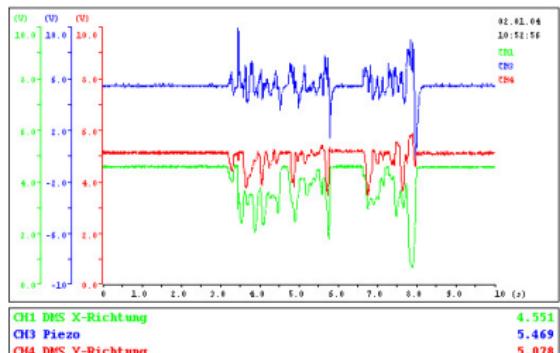
A handwritten signature in black ink, crossed out with a large red X.

# Beispiel: Unterschriftenverifikation – 3



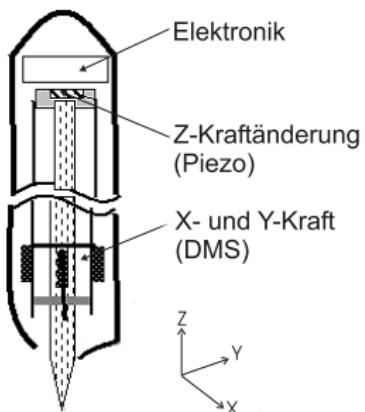
Zeitreihen von Kräften und  
Kraftänderungen

# Beispiel: Unterschriftenverifikation – 3

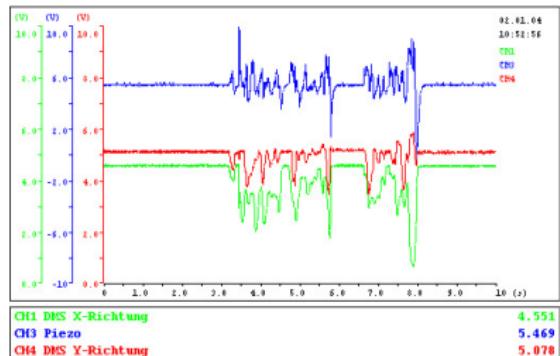


Zeitreihen von Kräften und Kraftänderungen

Instrumentierter Stift:

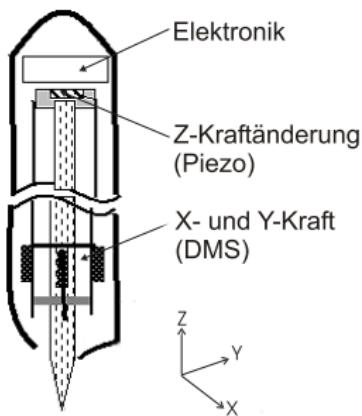


# Beispiel: Unterschriftenverifikation – 3



Zeitreihen von Kräften und Kraftänderungen

Instrumentierter Stift:



Vorteile im Vergleich zu Bild:

- höherer Informationsgehalt
- schwerer zu fälschen

# Beispiel: Unterschriftenverifikation – 5

Eigene Untersuchungen in 2009:

- Pro Person etwa 12 Unterschriften für etwa 100 Personen
- Entwicklung von speziellen Methoden zur Messung der Ähnlichkeit von Zeitreihen, die in Support Vector Machines integriert werden.

# Beispiel: Unterschriftenverifikation – 6

Paradigm	EER	FAR	FRR	$\sigma_{\text{FAR}}^2$	$\sigma_{\text{FRR}}^2$
SVM-CONV	9.78%	1.42%	21.40%	0.04	4.69
SVM-EUCLID	12.13%	2.15%	33.97%	0.34	8.56
SVM-HMM	12.39%	1.96%	60.43%	0.19	12.20
SVM-DTW	1.36%	0.47%	4.63%	0.12	1.35
SVM-LCSS	0.60%	0.04%	2.77%	0.00	0.84

EER: Equal Error Rate, FAR: False Acceptance Rate, FRR: False Rejection Rate,  $\sigma$ : Noun

# Kooperationen

Aktuell:

- BMW, Daimler, EAM/ENM, energycast, Fraunhofer IWES, IAV, SMA

Früher:

- B. Braun, HUK Coburg, DaimlerChrysler, BKH Taufkirchen, Bosch, Micro-Epsilon, T-Systems, Wacker / Siltronic, crealytics, SMA, ...

# Sonstiges

# Aktuelle Forschungsarbeiten am Fachgebiet IES



## Beispiele für zukünftige Aktivitäten:

- **Grundlagen:** Methoden für aktives Lernen, kollaboratives Lernen, Anomalieerkennung und Selbst-Organisation in (verteilten) technischen Systemen
- **Anwendungen:** Automobil/Verkehr, Energiesysteme, Biometrie, Angriffs- und Betrugserkennung

# Literatur zur Vorlesung

- M. H. Dunham: Data Mining – Introductory and Advanced Topics
- I. H. Witten, E. Frank: Data Mining – Praktische Werkzeuge und Techniken für das maschinelle Lernen
- R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification
- C. M. Bishop: Pattern Recognition and Machine Learning
- M. Kantardzic: Data Mining – Concepts, Models, Methods, and Algorithms
- T. Hastie: The Elements of Statistical Learning – Data Mining, Inference, and Prediction
- M. Berthold, D. J. Hand: Intelligent Data Analysis – An Introduction
- L. Sachs: Angewandte Statistik – Anwendung statistischer Methoden
- D. Hand, H. Mannila, P. Smyth: Principles of Data Mining

... weitere in den einzelnen Kapiteln.

Empfohlen wird die regelmäßige Teilnahme an der Vorlesung, Notizen auf Folien und Übung – die Software ist auch zu Hause installierbar.

# Externe Webseiten zur Vorlesung

- <http://rapid-i.com/>  
*RapidMiner*
- <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>  
*LibSVM: a Library for Support Vector Machines*

... weitere in den einzelnen Kapiteln.

# Ende