Begriffsdefinitionen (1)

> „Data mining ist defined as the process of discovering patterns in data. The Process must be automatic or (more usually) semiautomatic. The patterns discoverd must be meaningful in that they lead to some advantage, usually an economic one. The data is invariably present in substantial quantities.“

... und den des (maschinellen) Lernens:

> „Things learn when they change their behavior in a way that makes them perform better in the future.“

Witten et al. (2013). *Data Mining*. Morgan Kaufmann.

Begriffsdefinitionen (2)

> „In particular, we define machine learning as a set of methods, that can automatically detect patterns in data and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainity (...).“

Murphy, K. (2012) *Machine Learning - A Probabilistic Perspective*. MIT-Press. (S. 1)

Begriffsdefinitionen (3)

> „The Problem of searching for patterns in data is a fundamental one and has a long and successful history. For instance, the extensive astronomical observations of Tycho Brahe in the 16^{th} centory allowed Johannes Kepler to discover the empirical laws of planetary motion, with in turn provided a springboard for the development of classical mechanics. (...) The field of pattern recognition is concerned with the automatic discovery of regularities to take actions such as classifying the data into different categegories.“

Bishop, C. (2007) *Pattern Recognition and Machine Learning*. MIT-Press. Springer.

Wozu maschinelles Lernen?

- Das indizierte WWW enthält am 8.9.2018 mindestens 4.42 Milliarden Webseiten (www.worldwidewebsize.com)
- Die Sentinel-Satelliten der Europäischen Weltraumorganisation (ESA) sendeten im Jahr 2016 täglich 10 Terabyte an die Bodenstation in Oberpfaffenhofen/DE. (www.dlr.de, 7.9.2018)
- ...

Arten des maschinellen Lernens

- *Überwachtes Lernen* (*unsupervised learning*)

  Ziel: Aufgrund einer Menge von Input-Output-Paaren (Trainingsdaten) eine Abbildung finden, die einem Input-Vektor $\mathbf{x}$ jeweils einen Output y zuordnet.

  - Klassifikation (y ist diskret)
  - Regression (y ist kontinuierlich)

- *Unüberwachtes Lernen* (*unsupervised learning*)

  Ziel: „Interessante Muster“ in den Daten zu finden

  - Clustering
  - Dimensionsreduktion

- *Bestärkendes Lernen* (*reinforcement learning*)

  Ziel: Mit welcher Strategie kann ein virtueller Agent in einer dynamischen Umgebung eine vorgegebene Belohnungsfunktion maximieren?

  - genetische Algorithmen
  - dynamische Programmierung

Diagnose

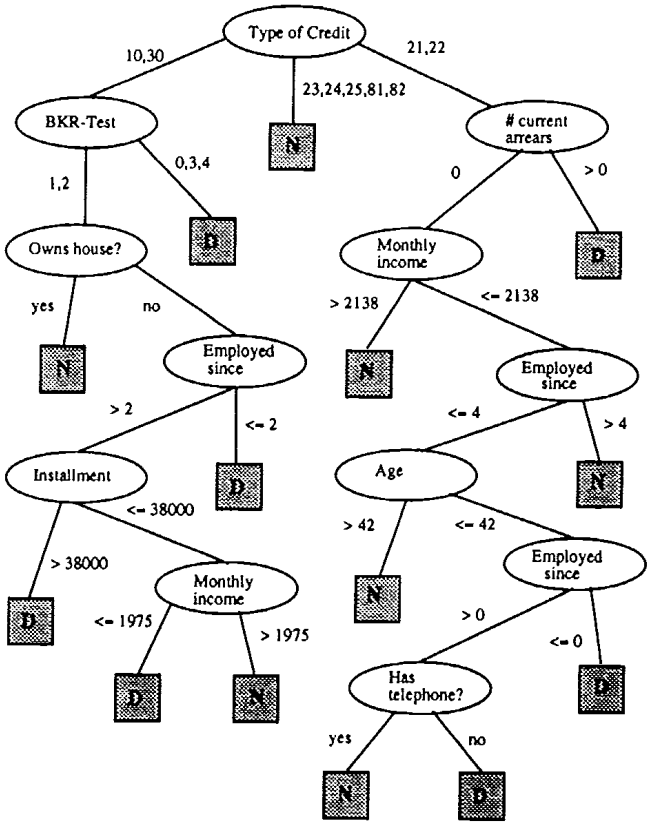
In den 80er Jahren des letzten Jahrhunderts zeigte eine Untersuchung, dass Data Mining-Regeln zur Diagnose von Krankheiten bei Sojabohnenpflanzen eine bessere Erkennungsrate hatten als Regeln, die von Experten aufgestellt wurden.

vom Diaporthe-Pilz befallene Sojabohne (*stem canker*)

http://bulletin.ipm.illinois.edu/pastpest/articles/200218d.html (31.12.2015)

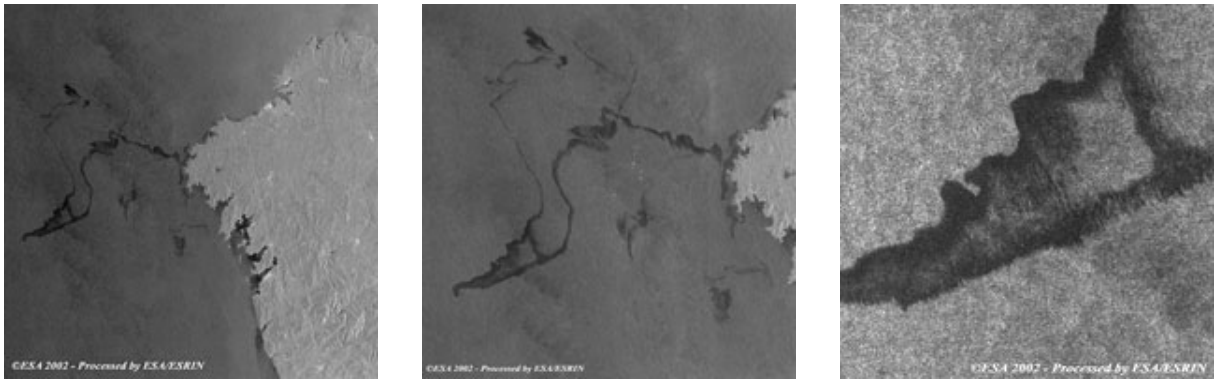
Kreditvergabe

In einer Fallstudie für Privatkredite (Feelder et al., 1995) wurden historische Kundendaten einer niederländischen Bank verwendet, um vorherzusagen, ob ein Kreditnehmer seinen Raten zurückzahlen wird. (D=Defaulter=säumiger Schuldner, N=Non Defaulter)



Analyse von Satellitenbildern

2002 bricht der mit 77 000 Tonnen Schweröl beladene Tanker „Prestige“ vor der Nordwestküste Spaniens entzwei und verursacht eine Naturkatastrophe.



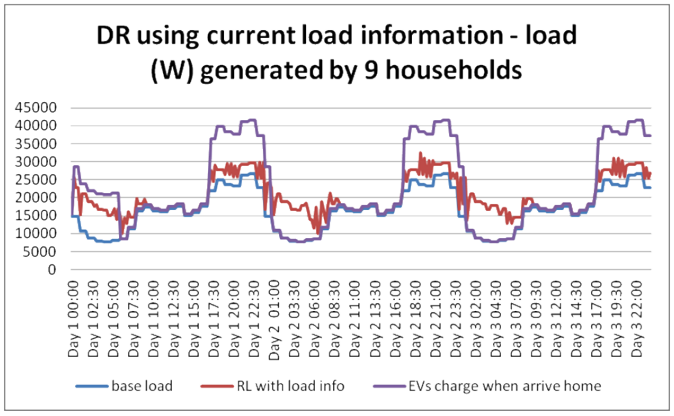
https://earth.esa.int/web/earth-watching/natural-disasters/oil-slicks/content/-/asset_publisher (30.12.2015)

Die schwarzen Flecken an der Küste im Bild links sind natürlichen Ursprungs.

Kann ein Computerprogramm den Unterschied zwischen „guten“ und „schlechten“ Flecken erkennen?

Optimierung von Angebot und Nachfrage in Stromnetzen

Durch „intelligente“ elektrische Geräte und Stromnetze lassen sich Schwankungen in der Stromnachfrage glätten. (DR = Demand Response, EV = Electric Vehicle)



https://www.tcd.ie/futurecities/research/energy/residential-demand.php (30.12.2015)

Marketing

http://content.time.com/time/magazine/0,9263,7601110321,00.html (30.12.2015)

Ethische Aspekte

Informationen über Geschlecht oder Rasse können für medizinische Fragen von Bedeutung sein. Wie verhält es sich jedoch bei der Vergabe eines Kredits oder bei der Auswahl eines neuen Mitarbeiters?

Selbst wenn solche Indentifikationsmerkmale aus den Daten entfernt werden, könnte beispielsweise die Postleitzahl Auskunft über den sozialen Status einer Person geben. Darüber hinaus ist bekannt, dass 85% der Amerikaner durch ihr Geschlecht, die Postleitzahl ihres Wohnorts, und ihr vollständiges Geburtsdatum identifiziert werden können. (Witten, Frank & Hall, S. 33).