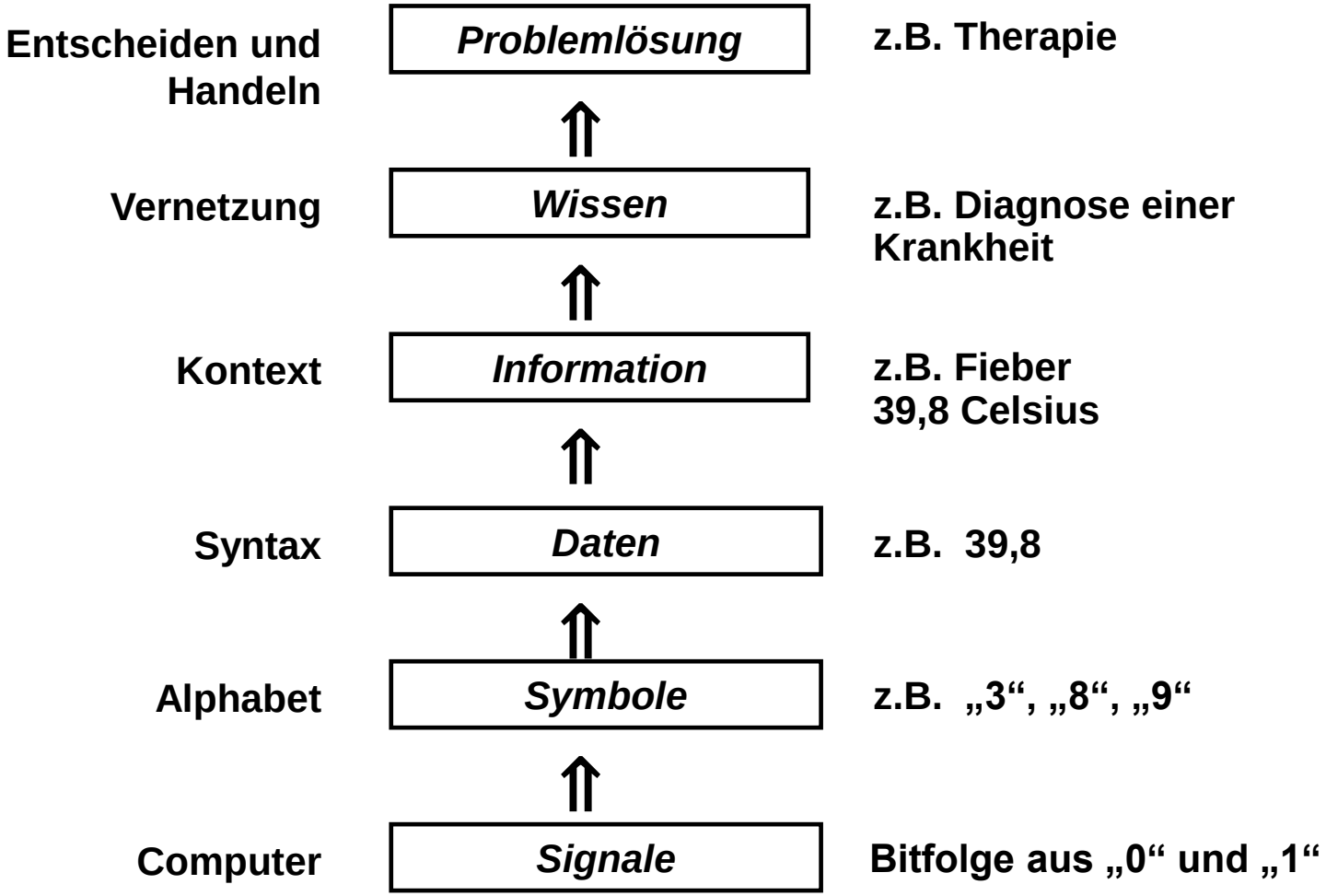
Die Berechnung des Menschen

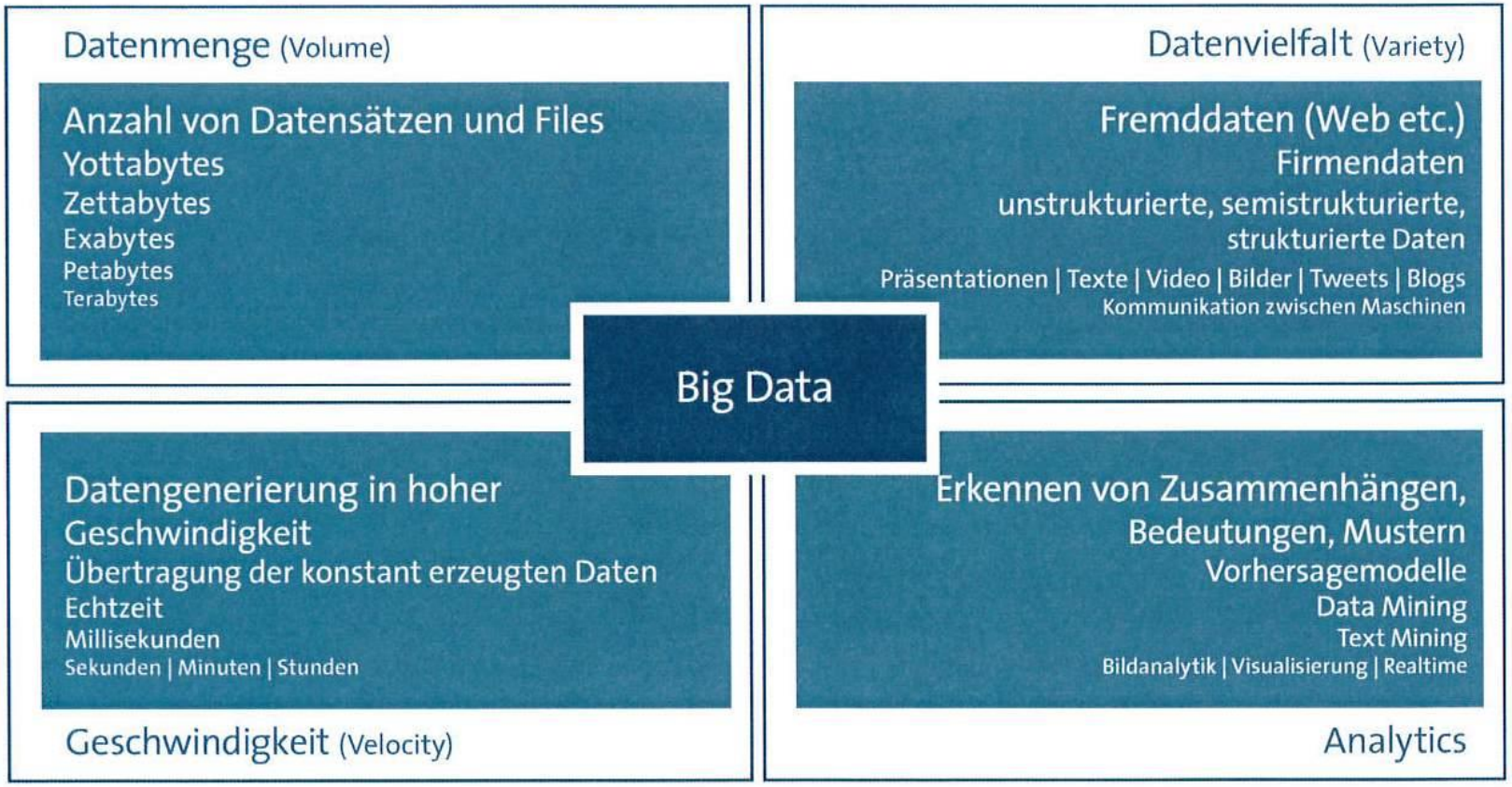
Wissenschaftstheoretische Grundlagen von Big Data in den Life Sciences und im Gesundheitsbereich

Klaus Mainzer
Lehrstuhl für Philosophie und Wissenschaftstheorie
Technische Universität München

Von Daten und Information zu Wissen

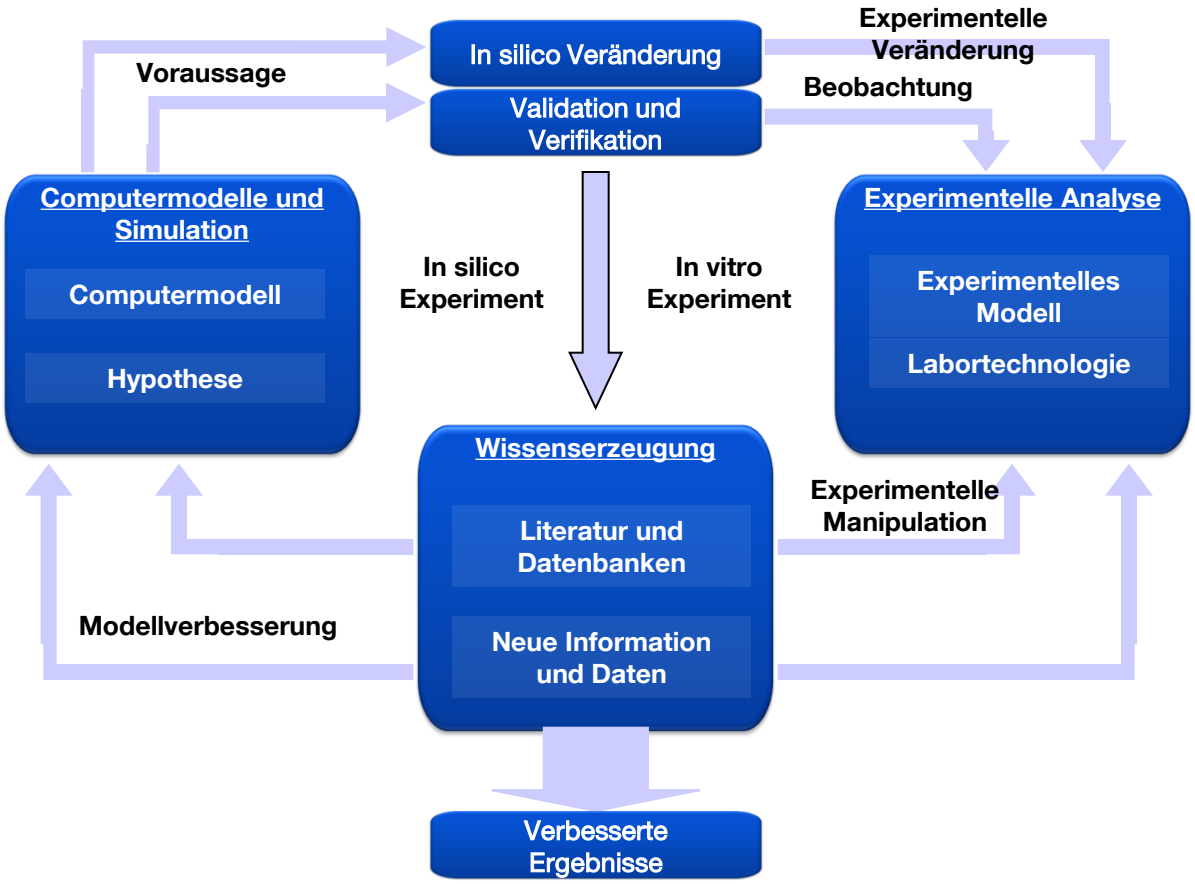


Big Data im Gesundheitssystem



Big Data Technologie **ist notwendig, um die** ***exponentiell wachsenden Datenmengen im Gesundheitssystem*** **(klinisch, epidemologisch, bildgebend, molekularbiologisch, ökonomisch etc.) zu bewältigen (z.B. 20 Terabytes pro Patientenakte für 2015).**

Big Data in den Life Sciences

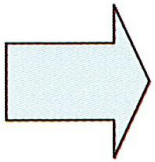


In den *Life Sciences* werden *Computermodelle* und *Simulationen* ("*in silico Experimente*") und *Laborexperimente* ("*in vitro* und *in vivo Experimente*") verbunden, um *neues Wissen* zu erzeugen, mit dem *bessere Modelle* und *neue Experimente* bestimmt werden können.

Wachsende *Anhäufung* von *biologischen Daten* führen zu *Computermodellen* von *Zellen*, *Organen* und *Organismen* mit *komplexen Netzwerken* für *Stoffwechselprozesse*, *Signalübertragung* und *genetische Regulation*.

Berechnung von Patientenprofilen

HIV-Genom
…aagtagggg
ggnaantaatag
aagcncgattag
atacaggagcag
atgatacagtatt
ngaagaaataaa
tttaccaggaaga
tggacaccaaaa
atgataggggga
attggaggttttat
caaagtaa…

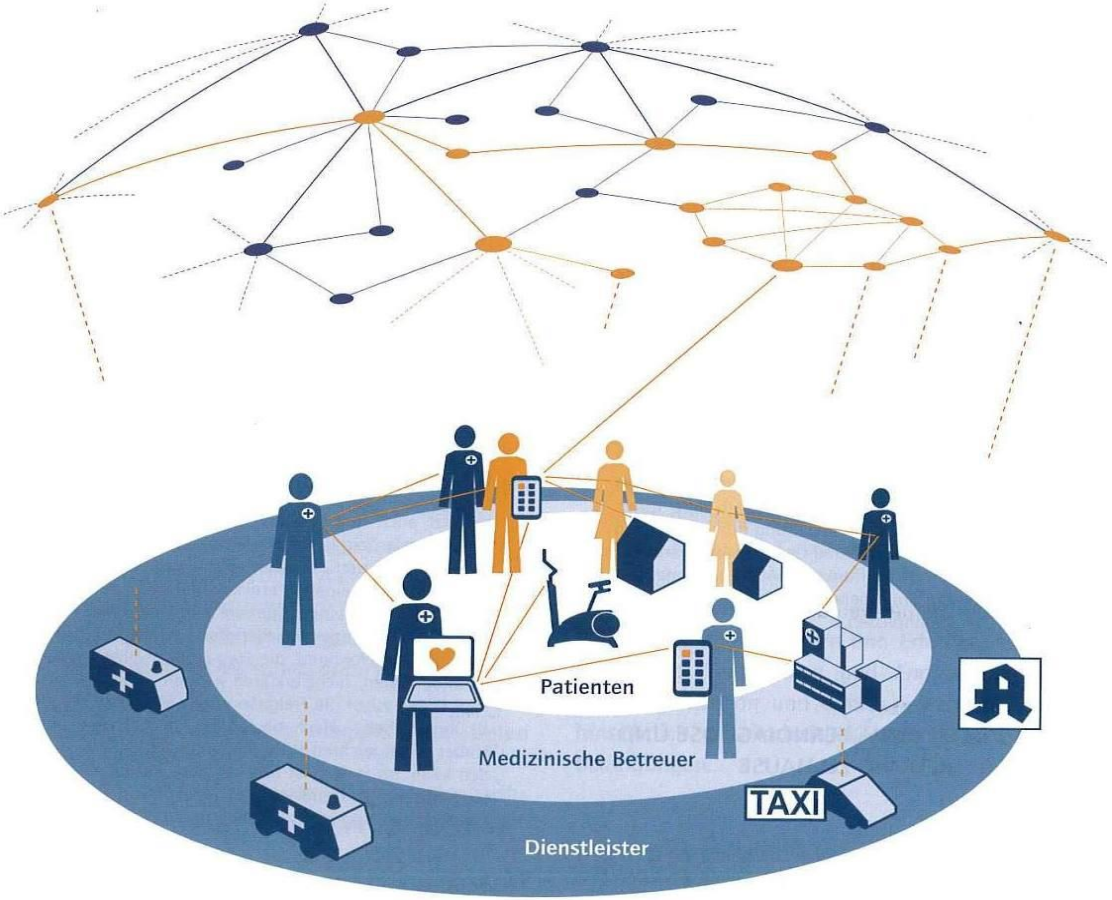


Wirkstoff	Aktivität
ZDV	1
ddC	1
ddI	0,68
d4T	1
3TC	0,0045
ABC	1
TDF	1
NVP	0,95
DLV	1
EFV	0,59
SQV	1
IDV	1
RTV	1
NFV	1
APV/FPV	1
LPV	1
ATV	1

Durch *Blutanalyse* wird die *Genomsequenz* eines *patientenspezifischen HIV-Erregers* (links) ermittelt. *Bioinformatisch* wird das *Resistenzprofil* dieses HIV-Erregerstammes gegen 17 verfügbare *AIDS-Medikamente* mit entsprechender *Resistenzwahrscheinlichkeit* (rechts) berechnet. So ergibt sich eine *Therapie* des Arztes.

Wegen der *Komplexität* des *biologischen Organismus* werden individualisierte Datenerhebungen immer notwendiger (*personalisierte Medizin*).

Internet der Dinge im Gesundheitssystem



acatech Cyberphysical Systems Studie: 2012, 42

Technische Grundlagen von Big Data Mining

	In-Memory Datenbanken	MPP Datenbank	Big Data Appliance	Hadoop	No SQL Datenbank
Konsistent	●	●	●	▲	▲
verfügbar	●	●	●	▲	▲
fehler = tolerant	●	●	▲	●	●
geeignet für Echtzeit Transaktionen	●	●	●	◆	◆
geeignet für Analyse	▲	▲	●	●	◆
geeignet für extreme Big Data	◆	▲	▲	●	●
geeignet für unstrukturierte Daten	◆	◆	▲	●	●

***Stärken* und *Schwächen* von *Big Data Technologien*:**

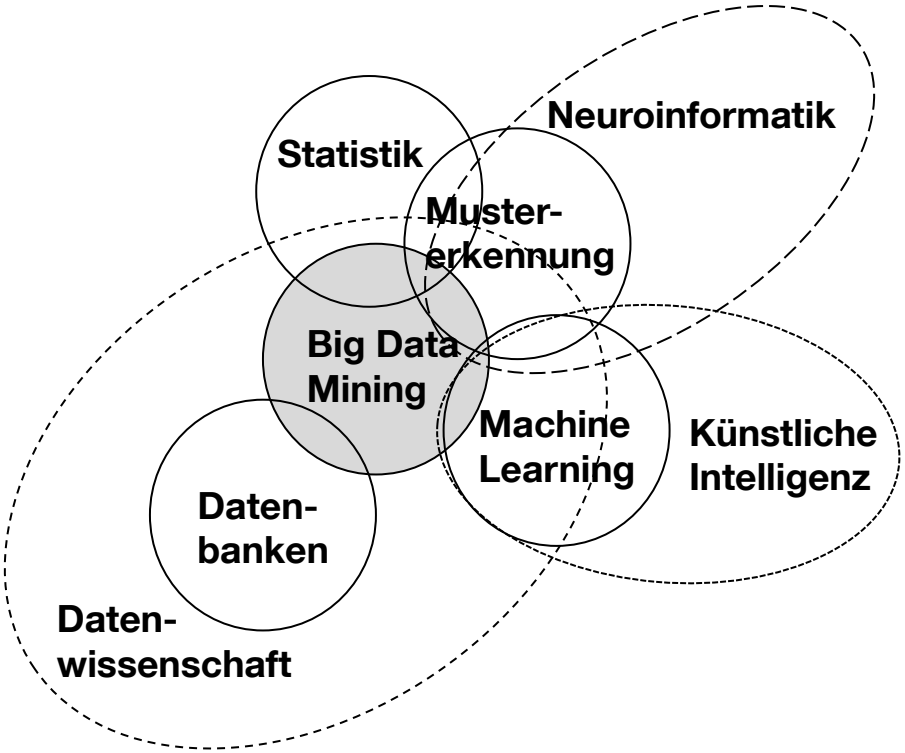
● = entspricht den Erwartungen

▲ = entspricht eingeschränkt den Erwartungen

◆ = entspricht nicht den Erwartungen

Vor *Einrichtung* einer *High-Performance Big Data Mining Plattform* muss eine *Stärken- und Schwächenanalyse* für die jeweiligen *Anwendungen* durchgeführt werden.

Interdisziplinäre Grundlagen von Big Data Mining



Voraussagemodelle **(*predicative modeling*) sind das zentrale Ziel von *Big Data Mining* als Teil der Datenwissenschaft. Dazu werden *Algorithmen* des *Machine Learning* nach dem Vorbild des menschlichen *Gehirns* aus den *Neurowissenschaften* und der *KI-Forschung* mit z.B. *Musterbildung* und *Clustering* ebenso angewendet wie *Methoden* der *Statistik* und *Datenbanken*.**

Methodologie für Voraussagemodelle im Data Mining

Stichprobe:

- *groß genug* für *signifikante* Information
- *klein genug* für *schnelle* Manipulation
- *statistisch repräsentativ*

Datenexploration:

- *Suche* nach *Trends*, *Anomalien* und *Muster*
- *Statistische Methoden* (z.B. Faktoranalyse, Clustering)

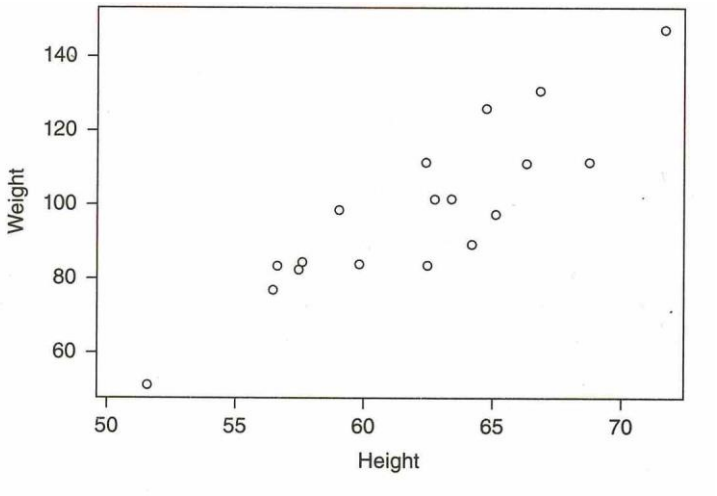
Modell:

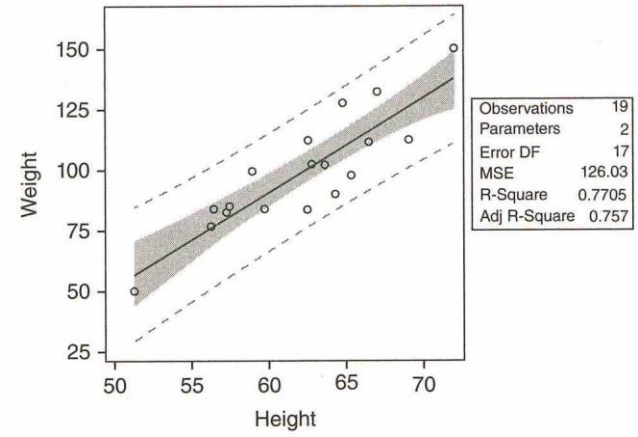
- Entwicklung einer *Software*, die *automatisch* in Daten *Korrelationen* und *Muster* entdeckt, um *Trends* und *Profile* vorauszusagen
- Unterschiedliche *Algorithmen* haben unterschiedliche *Vor- und Nachteile* (z.B. Neuronale Netze, Entscheidungsbäume, stochastische und Bayesianische Modelle, Zeitreihenanalyse)

Modellselektion:

- *Bewertung* (assess, evaluation) und *Auswahl* (selection) aus einer möglichen Modellklasse
- *Verbesserung* des Modells

Regressionsanalyse für Voraussagemodelle





Regressionsanalyse **erlaubt eine *lineare Trendvoraussage*, wenn eine *lineare Beziehung* zwischen der *abhängigen Variable* und allen *unabhängigen Variablen* angenommen wird:**

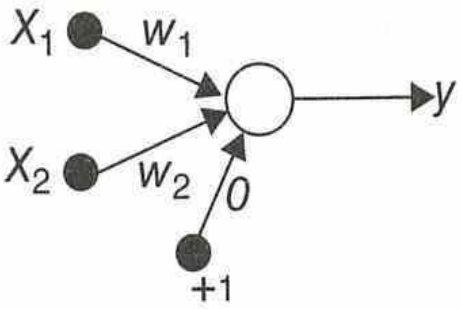
$H_i = \beta_0 + W_i\beta_1$ **mit z.B.** H_i ***Größe*** **und** W_i ***Gewicht*** **eines Patienten *i*.**

Die *beste Trendlinie* wird nach der (bereits auf C.F. Gauß (1777-1855) und A.M. Legendre (1752-1833) zurückgehenden) *Methode der kleinsten Quadrate* bestimmt:

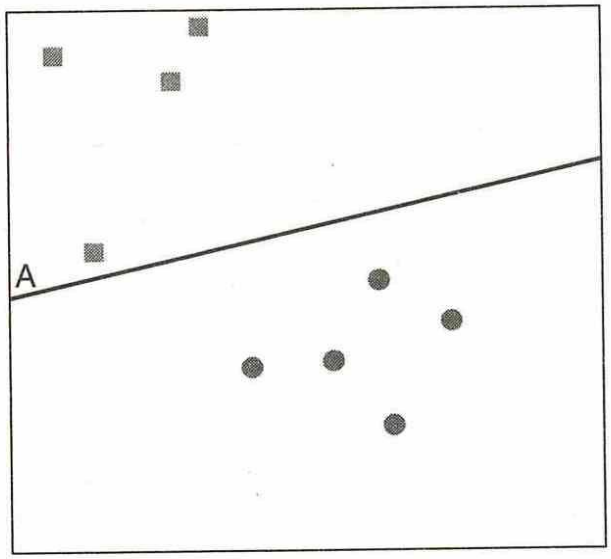
Die *beste Linie minimiert* die *Summe der Abweichungen* aller *Datenpunkte* von der vorausgesagten Trendlinie. Die *Abweichung einer Beobachtung* wird durch den *quadratischen Abstand* zwischen der *vorausgesagten Linie* und dem *Beobachtungswert* bestimmt.

Weiterentwicklungen **(z.B. Generalized Linear Models =GLM) werden auch in *Big Data* angewendet (z.B. Versicherungsprämien).**

Lernalgorithmus für (lineare) Mustererkennung



Nach dem Netzmodell (1943) von W. McCulloch und W. Pitts feuert ein Neuron (d.h. output y=1), wenn die *Summe* seiner *erregten Inputs* (z.B. x_1, x_2) gewichtet mit den *Synapsenstärken* W_1, W_2 größer als ein *Schwellenwert* ist und keine *hemmenden Inputs gleich* 1 sind.

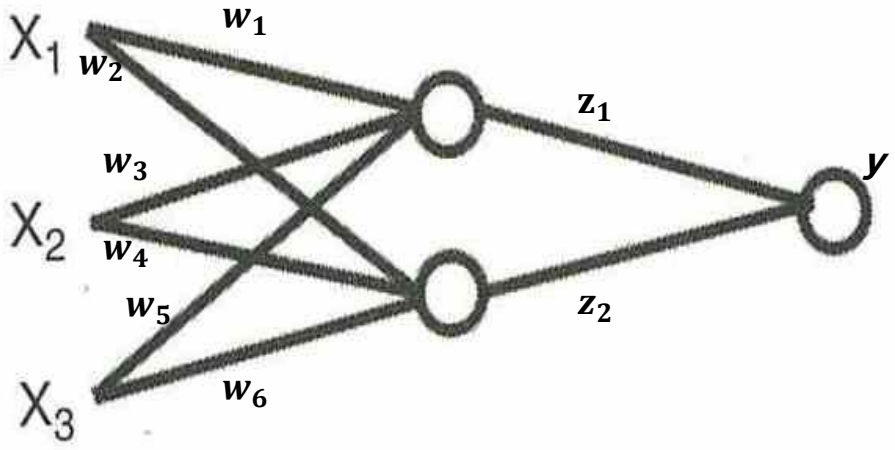


Der *Perzeptron Lernalgorithmus* (1950) beginnt mit einer Zufallsmenge von *Gewichten* und *modifiziert* diese *Gewichte* nach einer *Irrtumsfunktion*, um die *Differenz* zwischen *aktuellen Output* eines Neuron und *gewünschten Output* eines *trainierten Datenmusters* (z.B. Buchstabenfolgen, Pixelbild) zu *minimieren*.

Dieser *Lernalgorithmus* kann nur *trainiert* werden, um solche *Muster wiederzuerkennen* (supervised learning), die *linear trennbar* sind – *keine nichtlinearen Beziehungen* (M. Minsky/ S. Papert 1969).

Lernalgorithmus für (nichtlineare) Regressionsanalyse

In einem *nichtlinearen Regressionsproblem* wird eine *nichtlineare Funktion* von *Inputvariablen* bestimmt, indem die *Gewichte* der Funktion $y(W,X)$ mit dem zu berechnenden *Gewichtsvektor W*, dem Vektor X der bekannten *Inputs* und dem bekannten *Output y optimiert* werden.

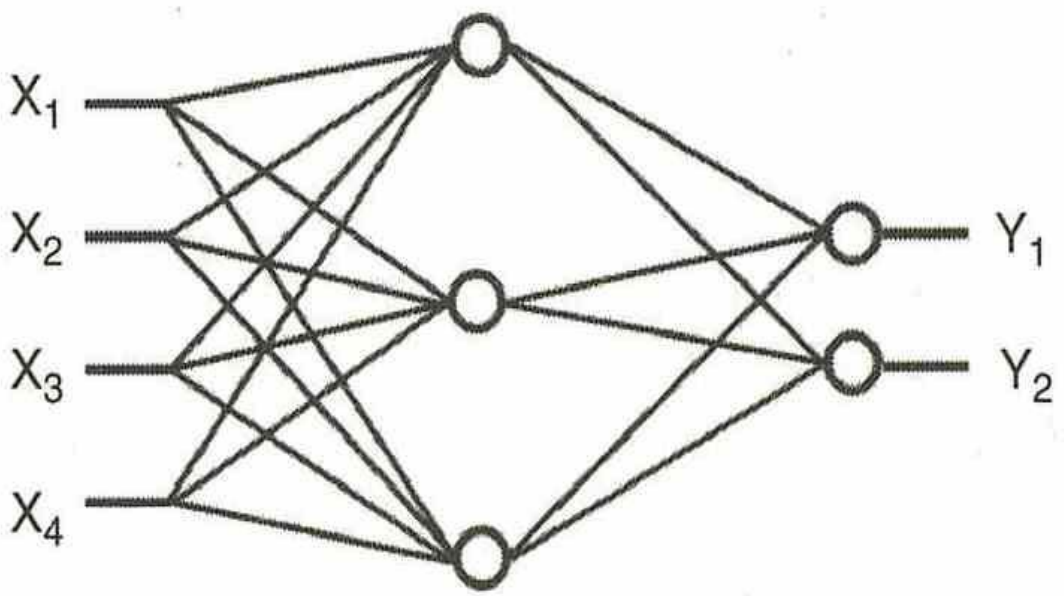


Ein (feedforward) *neuronales Netz* mit 3 Schichten von *Inputneuronen*, mittleren („*versteckten*“) *Neuronen* und *Outputneuronen* ist bestimmt durch die *Outputfunktion*

$$y(Z,W,X) = o(Z \cdot h(W \cdot X))$$

mit *Inputvektoren X, Gewichtsvektoren W* zwischen Inputschicht und versteckten Neuronen, *Aktivierungsfunktion h* der versteckten Neuronen, *Gewichtungsvektor Z* zwischen versteckten Neuronen und Outputneuronen und *Aktivierungsfunktion* o des Outputneurons. Mit einem *Outputneuron* können auch *einzelne numerische Werte vorausgesagt* werden.

Lernalgorithmus für (nichtlineare) Klassifikationen

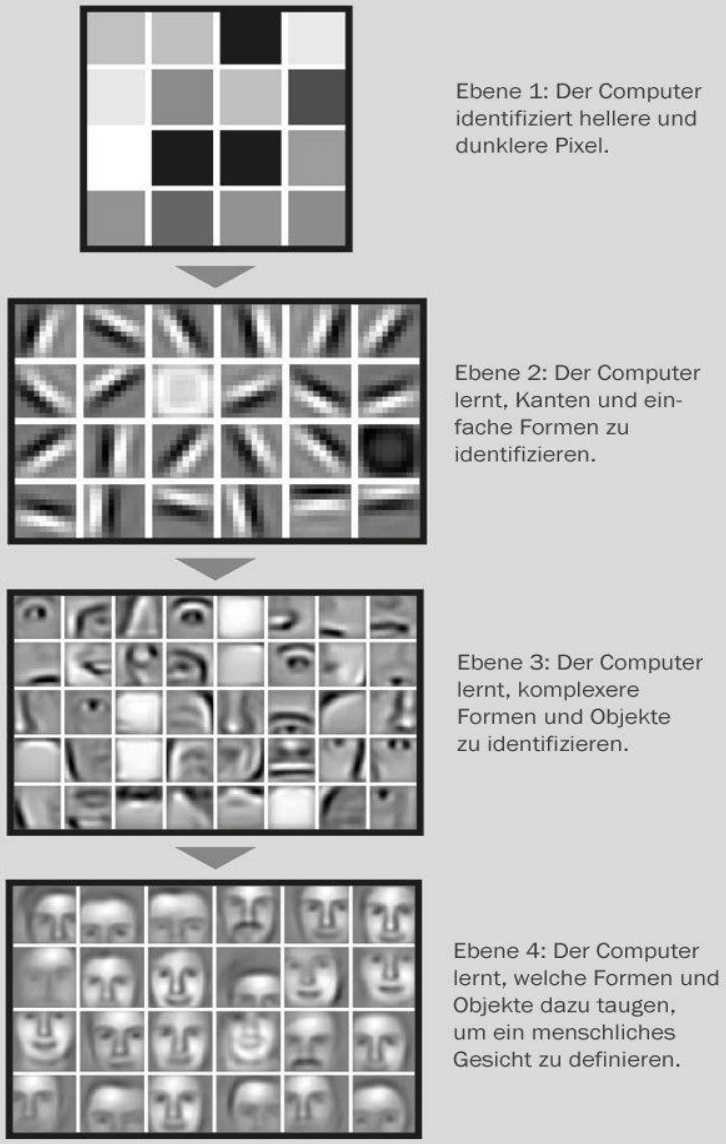


Ein *feedforward neuronales Netz* mit 3 Schichten und *zwei Outputneuronen* ist bestimmt durch die Outputfunktion

$$y_1(Z_1, W, X) = o(Z_1 \cdot h(W \cdot X))$$
$$y_2(Z_2, WX) = o(Z_2 \cdot h(W \cdot X))$$

mit *Gewichtsvektoren* Z_1 und Z_2 zwischen den *verstecken Neuronen* und den *beiden Outputneuronen.*

Bei *Klassifikationsaufgaben* lernen neuronale Netze *vorauszusagen*, zu *welcher Klasse* (entsprechend der Anzahl der Outputneuronen) ein Input gehört (z.B. Gesichts-, Profil-, Krankheitserkennung).



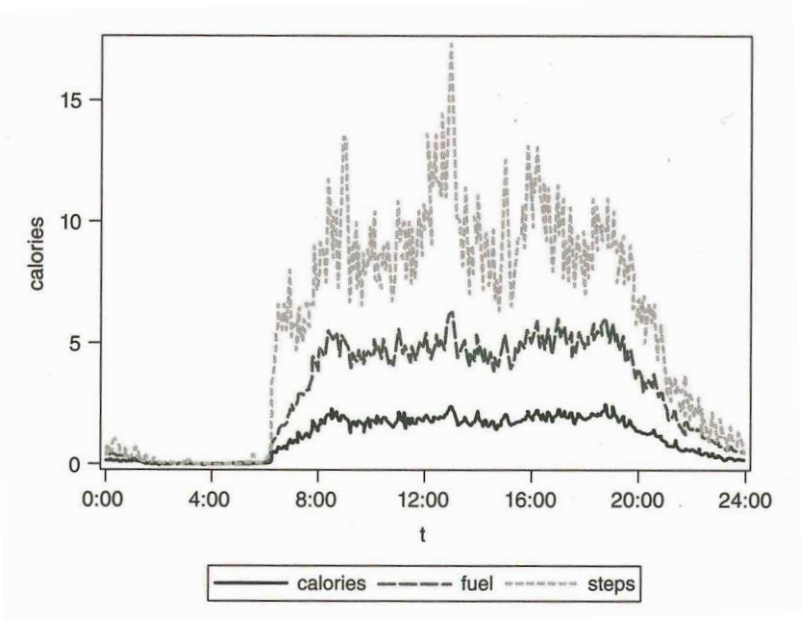
Deep Learning: Wie Maschinen lernen lernen

Beim *Deep Learning* werden *neuronale Netze* zu *Ebenen* angeordnet, die *immer komplexere Merkmale* verwenden, um z.B. den *Inhalt eines Bildes* zu erkennen. So lassen sich große *Datenmassen* in Kategorien einteilen.

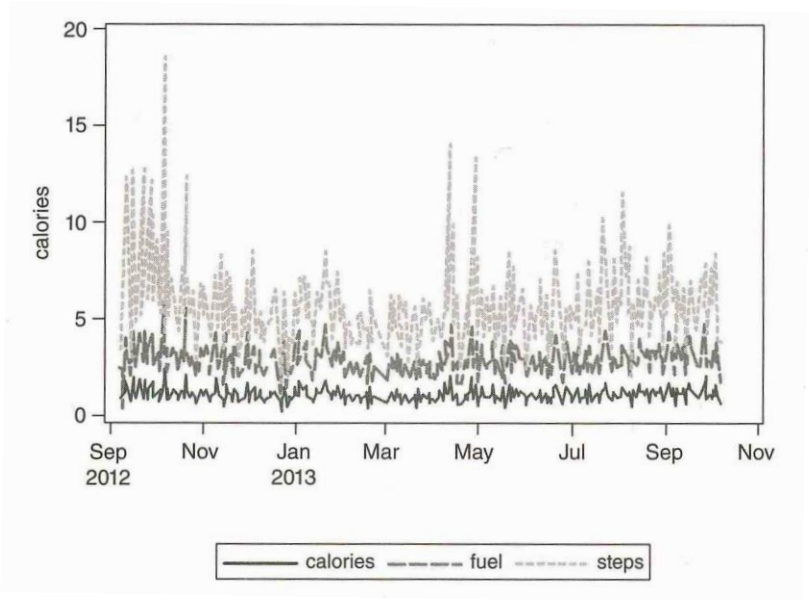
Im „*Google Brain*“ (Mount View CA 2014) werden ca. 1 *Million Neuronen* und 1 *Milliarde Verbindungen* (Synapsen) simuliert. *Big Data Technologie* macht *neuronale Netze* mit *mehrfachen Zwischenschritten* möglich, die in den 1980er Jahr nur theoretisch denkbar waren.

Zeitreihenanalyse im Gesundheitsbereich

Im *Internet der Dinge* werden *Wearables* (z.B. Nike-FuelBand) eingesetzt, um *Verhaltens-* und *Gesundheitsdaten* durch *Biosensoren* zu messen. *Zeitreihenanalyse* entdeckt *Korrelationen* und *Muster*, um aus *Big Data-Massen Informationen* zu gewinnen.

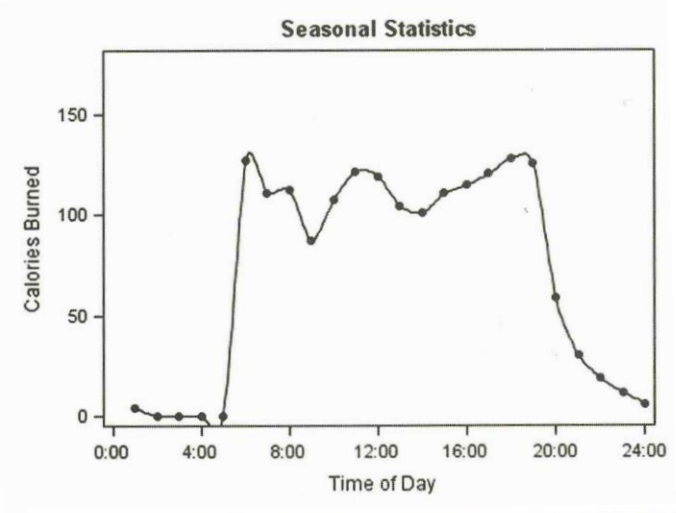


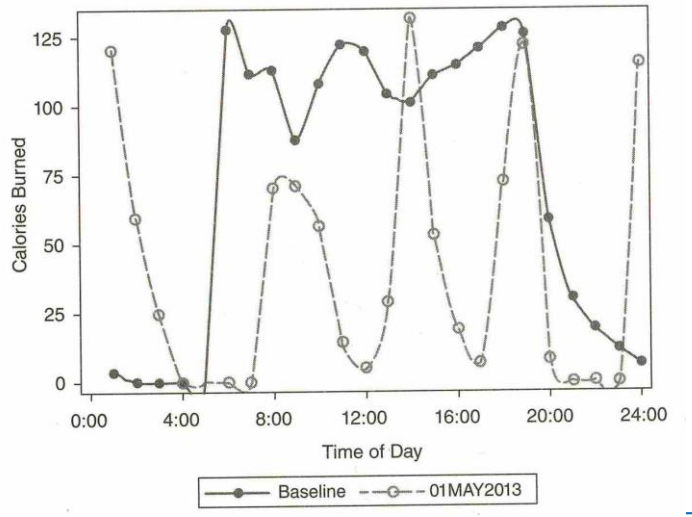
Saisonaler Aktivitätsindex



Minuten-zu-Minuten Aktivitätsindex

Ähnlichkeitsanalyse führt von Big Data zu Gesundheitsinformationen





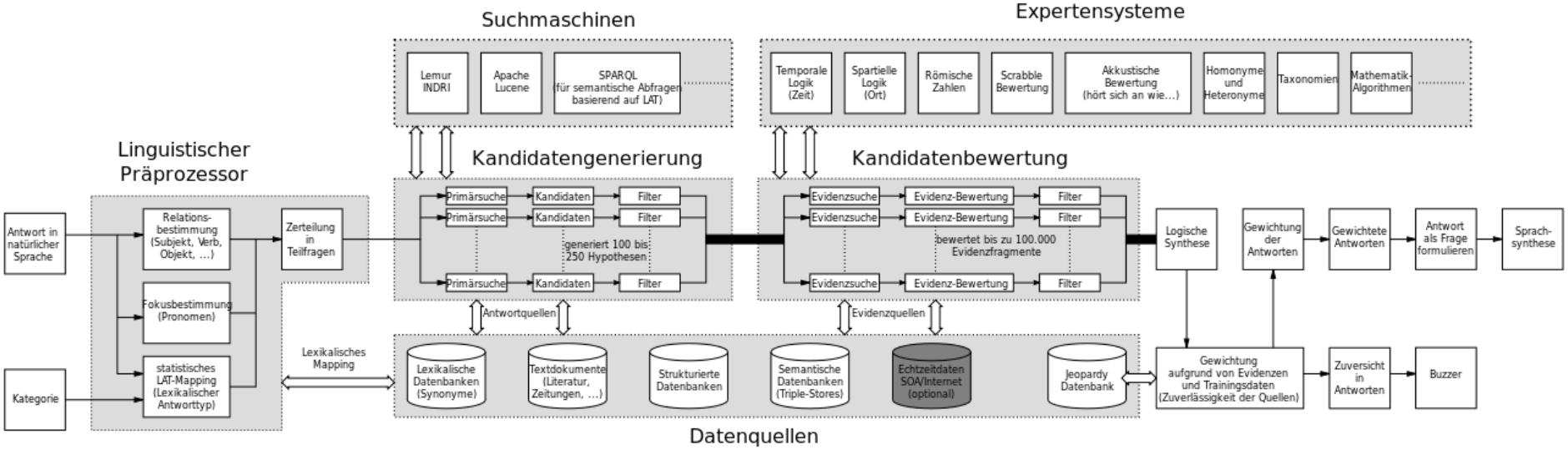
Um aus den *Datenmassen* zu lernen, wird zunächst ein *normales* (durchschnittliches) *Tagesmuster* z.B. des Kalorienverbrauchs (target) ermittelt. In einer *Ähnlichkeitsanalyse* werden *abnormale Abweichungen* bestimmt. (hier: Büro- und häuslicher Alltag werden durch Konferenz und Messebetrieb unterbrochen).

In der *Gesundheitsfürsorge* für ältere Patienten/Senioren, die *zu Hause* leben (wollen), spielt *automatische Ähnlichkeitsanalyse* zur Feststellung von *abnormalen Datenmustern* eine grundlegende Rolle.

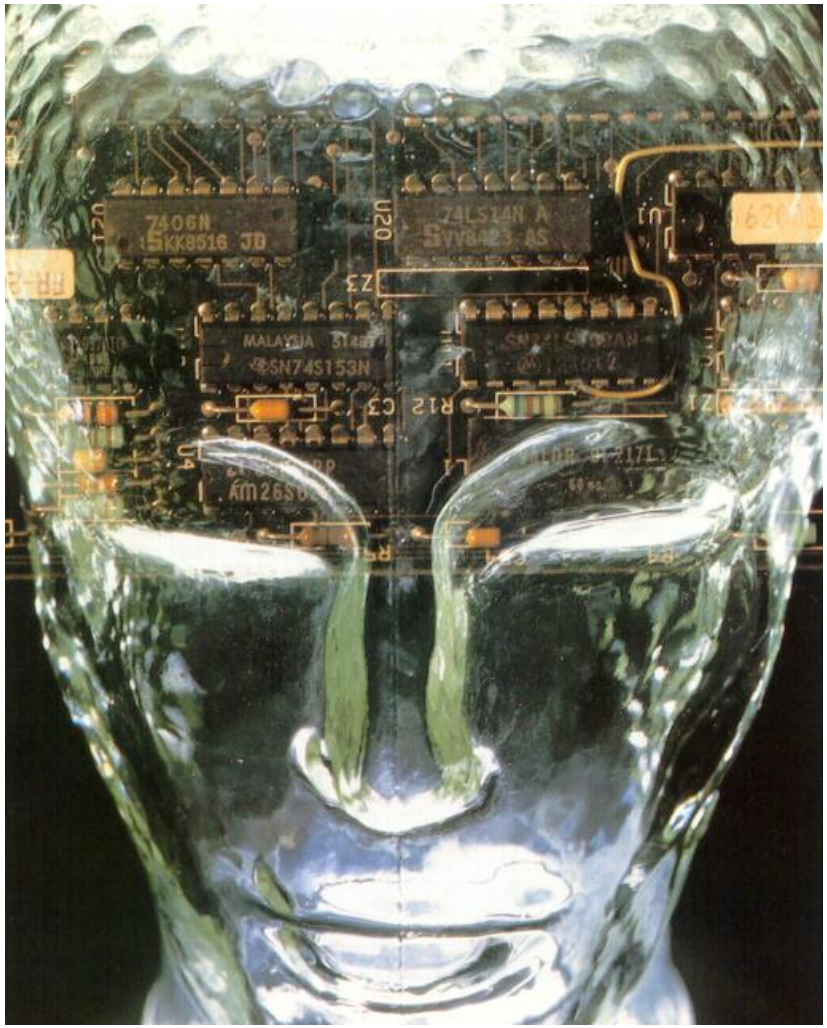
Big Data Technologie übertrifft den Menschen? WATSON

WATSON ist eine *semantische Suchmaschine* (IBM), die in *natürlicher Sprache* gestellte *Fragen* erfasst und in einer *Big Data Datenbank passende Fakten* und *Antworten* in *kurzer Zeit* findet.

Sie orientiert sich *nicht* am *menschlichen Gehirn* (Deep Learning), sondern integriert *Sprachalgorithmen*, *Expertensysteme*, *Suchmaschinen* und *linguistische Prozessoren* auf der Grundlage der *Rechen-* und *Speicherkapazitäten* von *Big Data Technologie*.



Mensch als Big Data Träger



Im Gesundheitswesen werden *Verhalten*, *Kognition* und *Emotionen von Menschen* mit *komplexen Datenmustern* verbunden.

Mit *neuronalen Netzwerken*, *Machine Learning* und *Big Data Algorithmen* werden *Profile* berechnet , *zukünftiges Verhalten* antizipiert und *passende Therapien* abgeleitet.

Wie weit können/dürfen wir gehen?

Big Data in der Medizin – Chancen und Risiken

***Intelligente Suchalgorithmen* (*machine learning*) müssen für den einzelnen Patienten die *passenden Schlüsselinformationen* finden (*personalisierte Medizin*).**

***Personalisierte medizinische Daten* sind durch *anonymisierte* und *pseudo-anonymisierte Verschlüsselung* vor *Missbrauch* (z.B. Versicherung, Marketing, Berufsmarkt) zu schützen.**

Big Data – „Eine neue Art der Wissenschaft“ ?

Führt *Big Data* in einer *komplexen Welt* zu einer „*neuen Art daten-getriebener Wissenschaft*“ mit *effizienten Algorithmen* - „*ohne Theorie*“ (C. Anderson)? Algorithmen *ohne Theorie und Gesetze* sind *blind* ! *Korrelationen* und *Datenmuster* ersetzen *keine Erklärungen und Begründungen von Ursachen*. Daher brauchen wir URTEILSKRAFT !

Literaturhinweise: