Clustering oder wie zerteile ich einen Graphen ?
(... und was ist eigentlich maschinelles Lernen ?)

Jun.-Prof. Matthias Hein

Perspektiven der Informatik, 1.1.2010

Inferenz I

Wie lernen wir ? Wie wird neues Wissen gewonnen ?

In den Naturwissenschaften unterscheidet man zwei Typen von Inferenz:

- **Induktive Inferenz:** Lernen von Zusammenhängen durch Beobachtungen.
- **Deduktive Inferenz:** Ableitung spezifischer Aussagen von allgemeinen Prinzipien (Axiomen).

Deduktive Inferenz:

- **Mathematik:** System aus Axiomen ⇒ Ableitung von Theoremen
- **Physik:** Postulate über Natur ⇒ Naturgesetze

Methoden in der Künstlichen Intelligenz

- Logik
- Automatisches Beweisen von Theoremen

Probleme:

- fuer nicht-mathematische Probleme existiert keine axiomatische Darstellung

Indduktive Inferenz:

Induktive Inferenz ist das zentrale Mittel in den Naturwissenschaften.

Vorgehensweise:

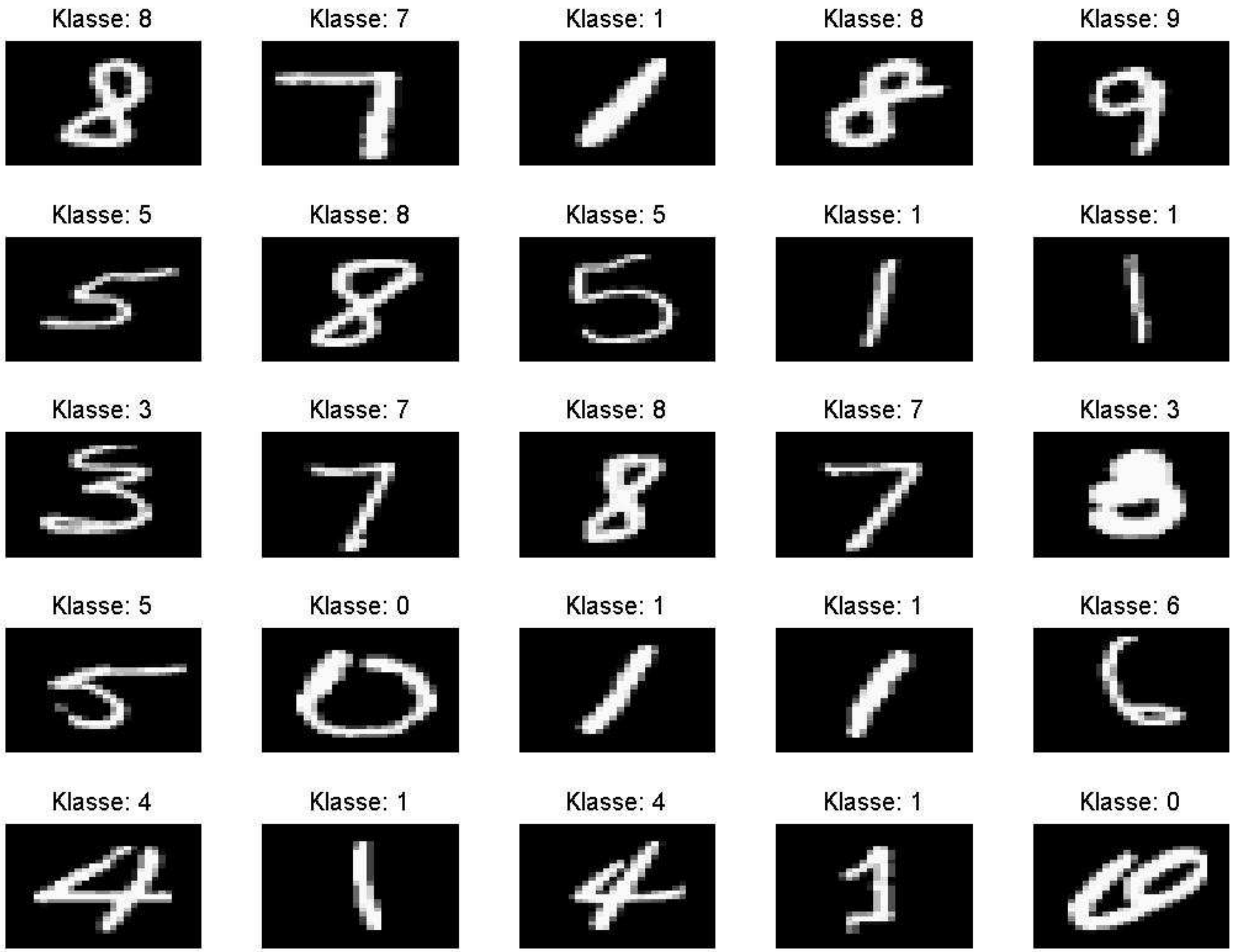
1. Sammeln von Beobachtungen.
2. Modellbildung.
3. Vorhersage

Falsifikation

Induktive Aussagen werden verworfen aber **nie** verifiziert.

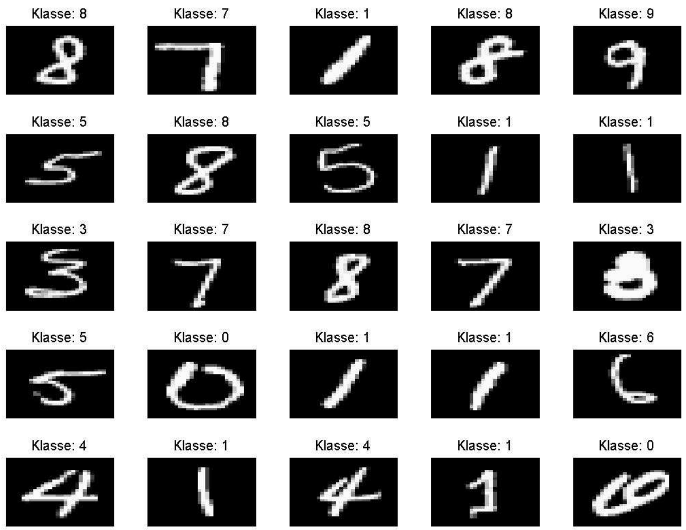
Maschinelles Lernen versucht den Prozess der Induktion zu automatisieren.

Lernproblem: Erkennung von handschriftlichen Zahlen



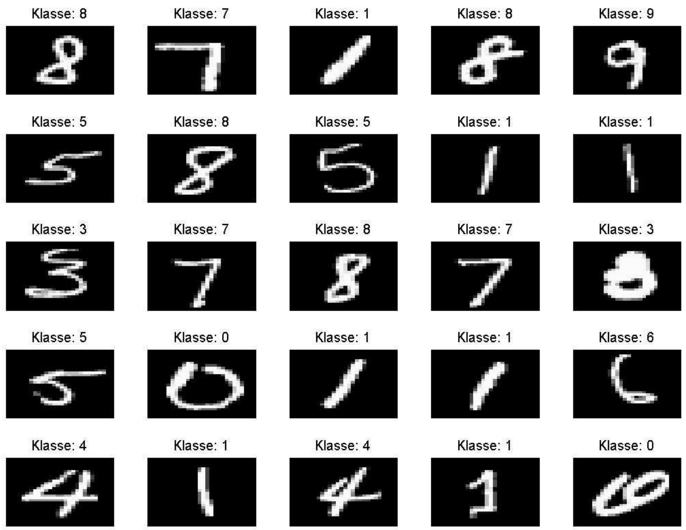
Was ist maschinelles Lernen ?

Terminologie im Maschinellen Lernen:



Eingabe	Pixeldarstellung des Bildes (hier: 28×28 - Grauwerte)
Merkmal	Eigenschaft der Eingabe (hier: Grauwert eines bestimmten Pixels
Ausgabe	eine Zahl $\{1, 2, \ldots, 10\} \Longrightarrow$ Mehrklassenproblem
Klassifikator	eine Funktion von Eingabe nach Ausgabe, hier $f : \mathbb{R}^{784} \rightarrow \{1, 2, \ldots, 10\}$.

Terminologie im Maschinellen Lernen:

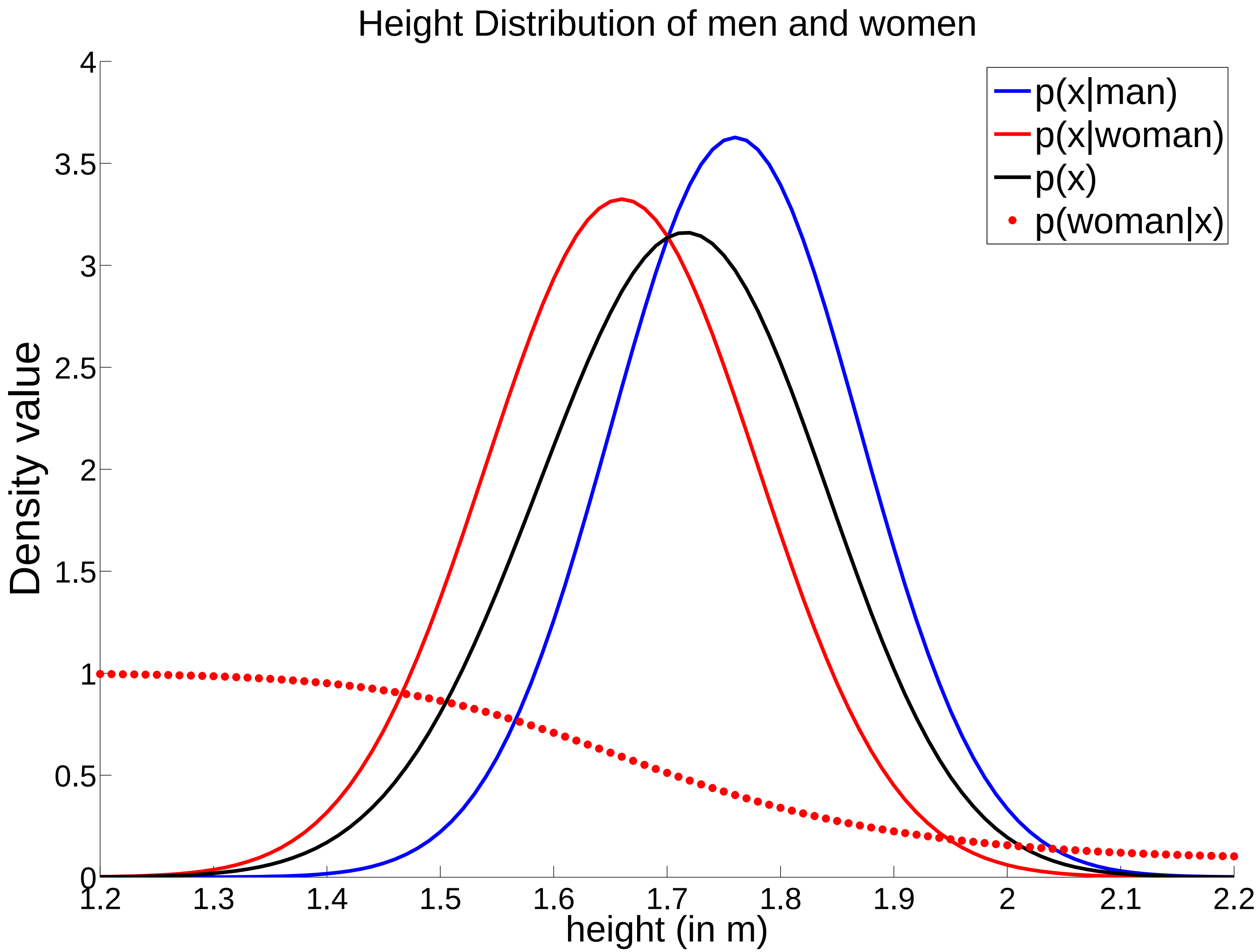


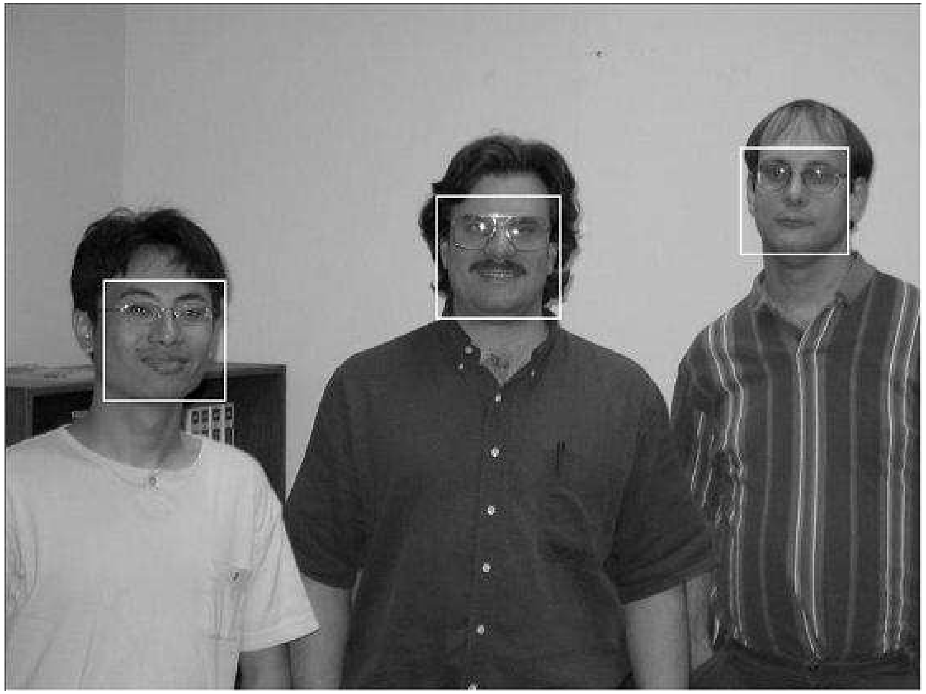
Training	Konstruktion des Klassifikators (Lernalgorithmus)
Test	Anzahl der Fehler auf neuen Bildern
Generalisierung	Klassifikator soll auf neuen Bildern funktionieren
	Kein einfaches auswendig lernen !

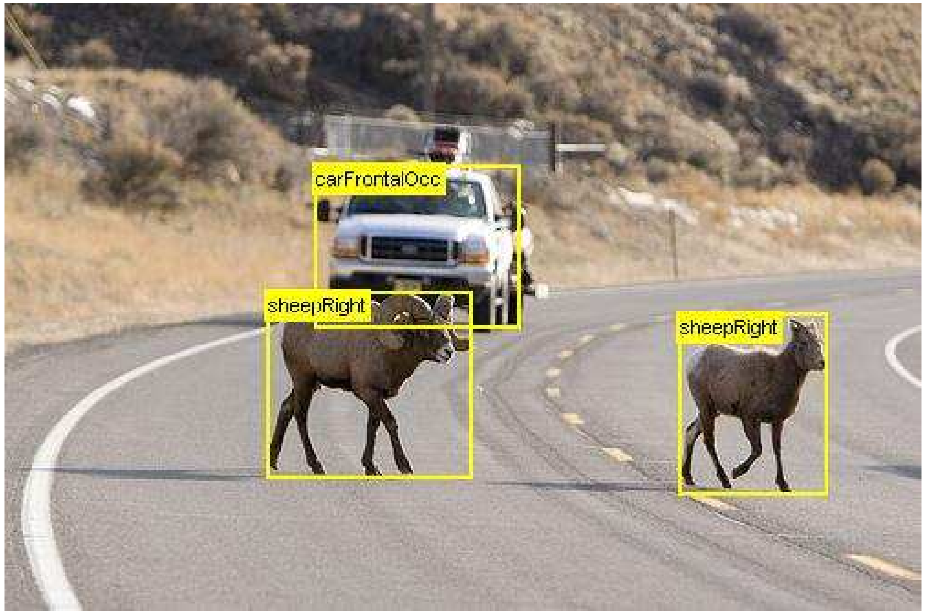
- Annahme: Es existiert ein datengenerierendes Wahrscheinlichkeitsmaß P on $\mathcal{X} \times \mathcal{Y}$.
- Was bedeutet das ?

  1. Trainingsdaten sind eine **zufällige Stichprobe** von P,
  2. Die Ausgaben $y \in \mathcal{Y}$ sind **nicht-deterministisch**, d.h. es existiert nicht notwendigerweise $y = g(x)$. Stattdessen für ein x gibt es eine Verteilung über $\mathcal{Y}$.
  3. Letzteres heißt, daß eine "perfekte" Lösung nicht existiert.

Lernproblem: Vorhersage des Geschlechts, $Y = \{\text{male}, \text{female}\}$, basierend auf der Körpergröße (Eingaberaum: $\mathcal{X} = \mathbb{R}$).







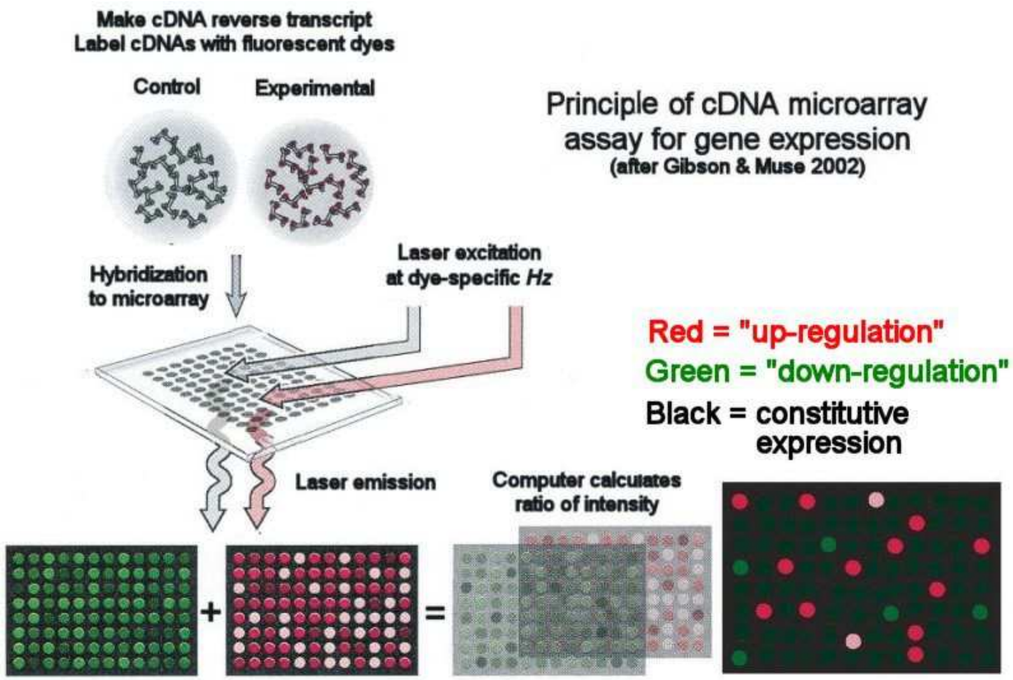
Objekterkennung im maschinellen Sehen:

- Gesichtserkennung (funktioniert gut - aber nur bei gerader Draufsicht)
- allgemeine Objekterkennung (Wettbewerbe heute mit 20 Objektklassen).

Robotik/Computer Vision

Machine learning in robotics and autonomous driving:

- Stanford won the DARPA Grand Challenge 2005 (Left: the navigation),
- The Grand Urban challenge 2007 has been won by CMU (Right: crashes of other cars).



Bioinformatik:

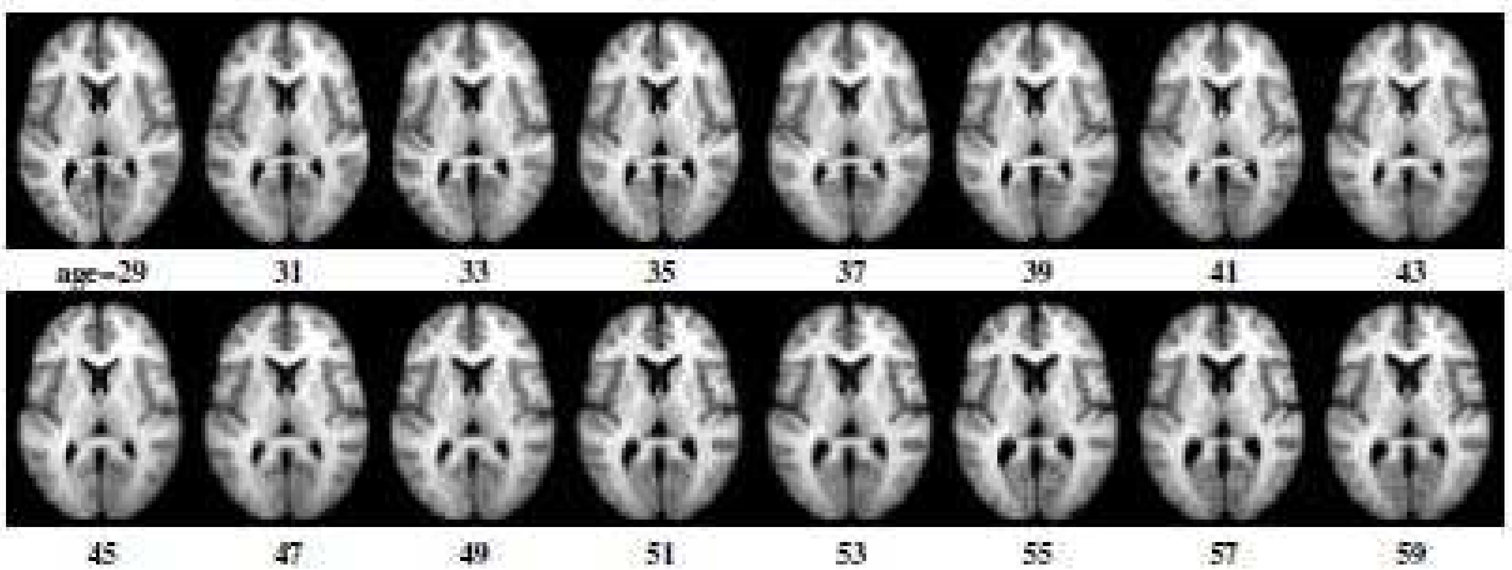
- bestimmte Gene reagieren im kranken Gewebe weniger oder mehr als im normalen Gewebe ⇒ Verwendung von maschinellem Lernen zur Diagnose von Krankheiten.

Regression:

Lernen einer allgemeinen Funktion $f : M \rightarrow N$.

Beispiele

- Vorhersage der Temperatur, Windrichtung,
- Vorhersage von ganzen Voxel-Bildern,



Wie verändert sich das Gehirn einer Person mit dem Alter ?

Anwendungsgebiete:

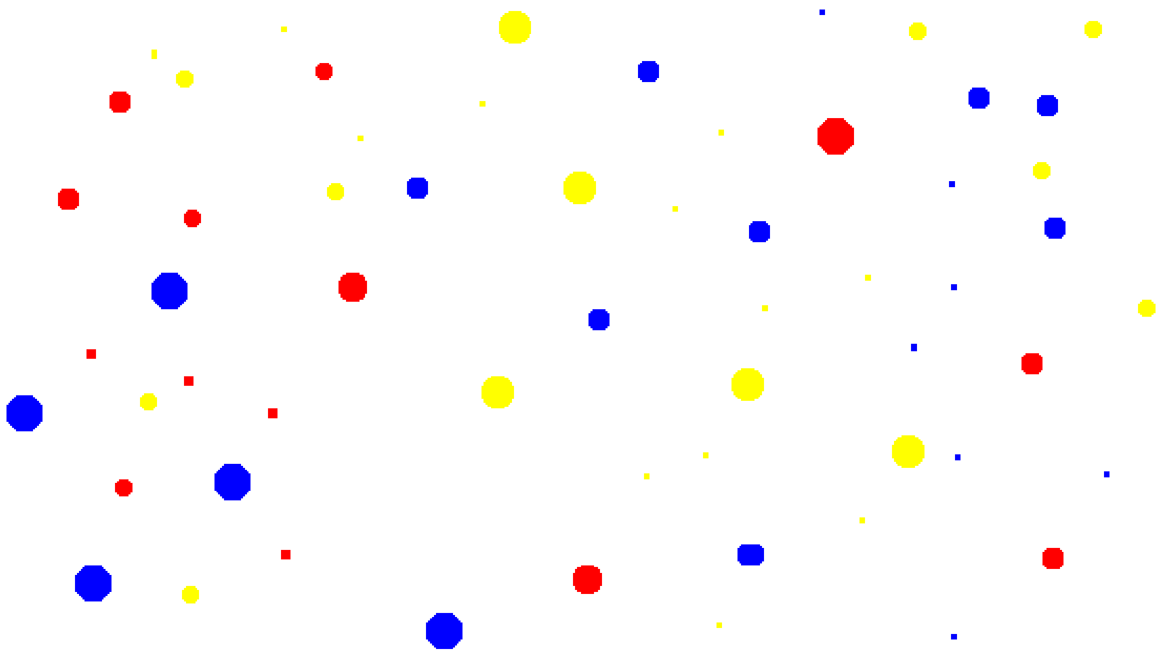
- Bioinformatik,
- Computer Vision/Image Processing/Computer Graphics,
- Information Retrieval/Collaborative Filtering,
- Robotik,
- Spam Filter/Intrusion Detection,
- neu: Machine Learning in Games und Software Engineering
- jedes Problem wo Daten analysiert werden müssen.

Mehr und mehr Daten werden gesammelt. Menschen allein können sie nicht analysieren.

$\Longrightarrow$ **Nachfrage nach maschinellem Lernverfahren steigt !**

$\Longrightarrow$ **Informatik auch als experimentelle Naturwissenschaft !**

Clustering von Daten



Clustering von Bällen (Farbe oder Größe ?)

Was ist Clustering ?

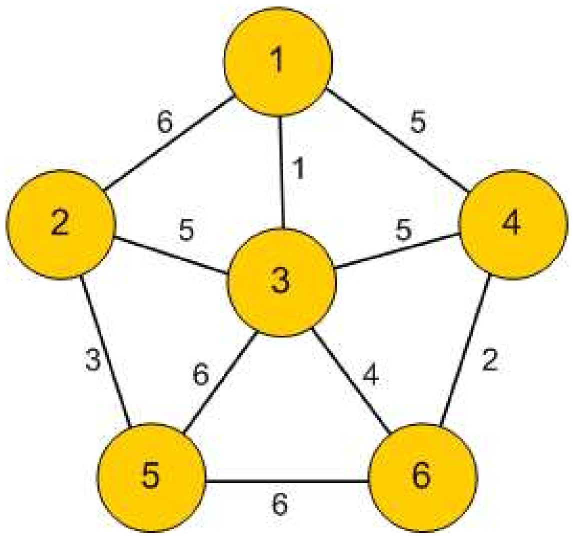
Definition 1. *Ein **Clustering** ist eine Gruppierung der Daten, so daß alle Daten in einem Cluster ähnlich sind und Daten in verschiedenen Clustern unähnlich sind.*

$\Rightarrow$ keine exakte mathematische Definition !

$\Rightarrow$ Ziel oft anwendungsabhängig.

Was bedeutet ähnlich ? Ideen ?

Definition 2. *Ein* **Graph** (V, W) *ist eine Menge* V *von* **Knoten** *und eine Gewichtsmatrix* W, *wobei* $w_{ij} > 0$ *genau dann wenn eine (ungerichtete)* **Kante** *vom Knoten* i *zum Knoten* j *besteht.*

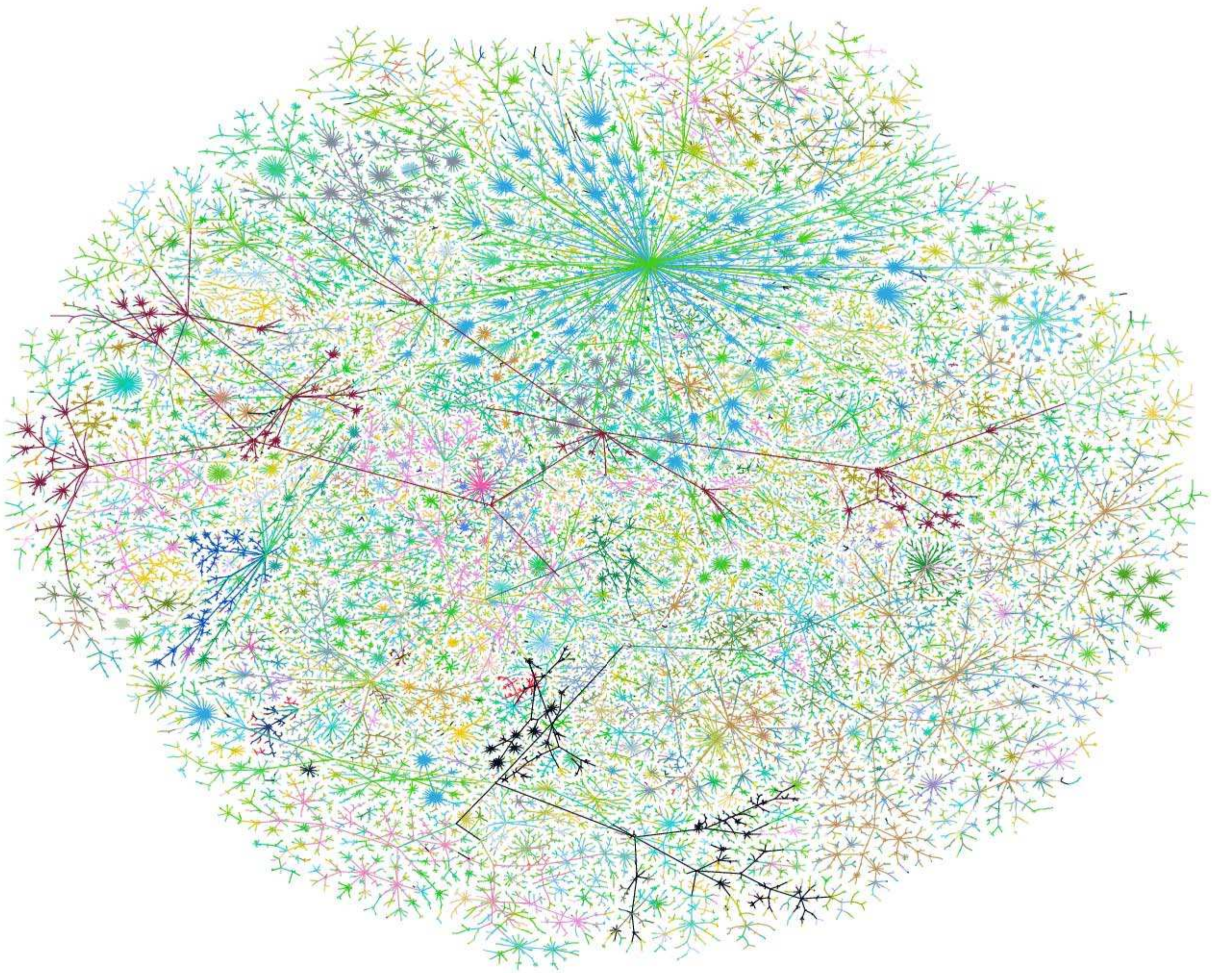


$$W = \begin{pmatrix} 0 & 6 & 1 & 5 & 0 & 0 \\ 6 & 0 & 5 & 0 & 3 & 0 \\ 1 & 5 & 0 & 5 & 6 & 4 \\ 5 & 0 & 5 & 0 & 0 & 2 \\ 0 & 3 & 6 & 0 & 0 & 6 \\ 0 & 0 & 4 & 2 & 6 & 0 \end{pmatrix}$$

Ein ungerichteter, gewichteter Graph.

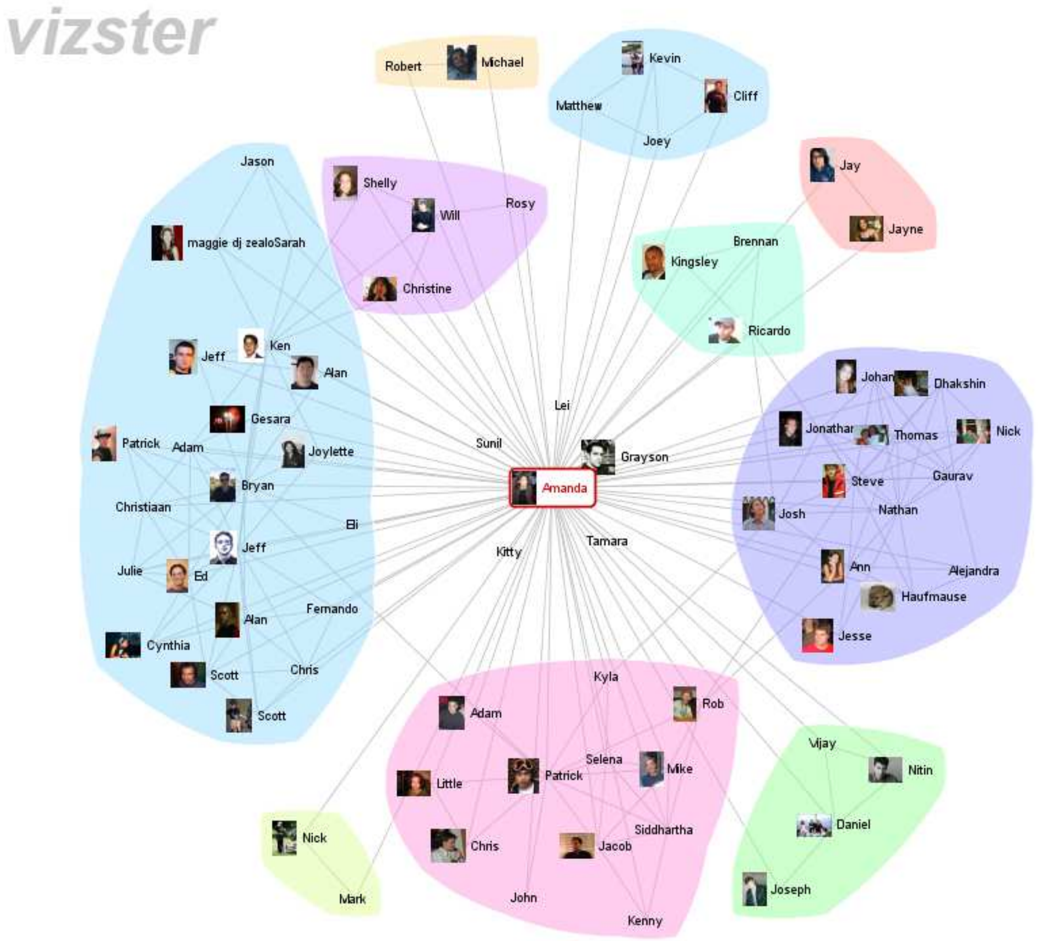
- **Knoten:** ein Datenpunkt (z.B. ein Protein, ein Link im Suchergebnis,...)
- **Kanten:** Ähnlichkeit zwischen Datenpunkten.

Beispiele von Graphen



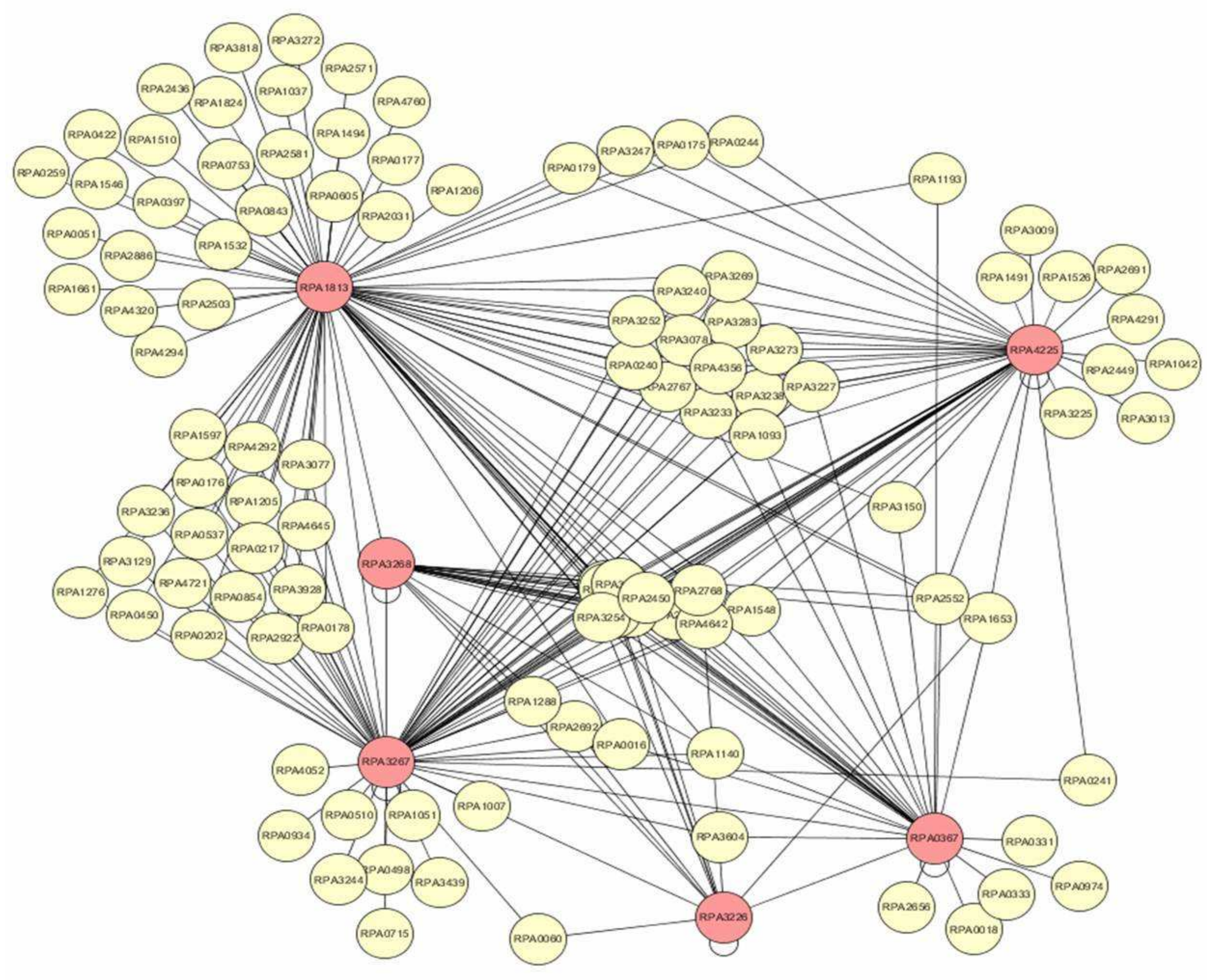
Eine Visualisierung des Internets auf dem Level von Serververbindungen.

Beispiele von Graphen I



Soziale Netzwerke: Extraktion von Gruppen.

Beispiele von Graphen II

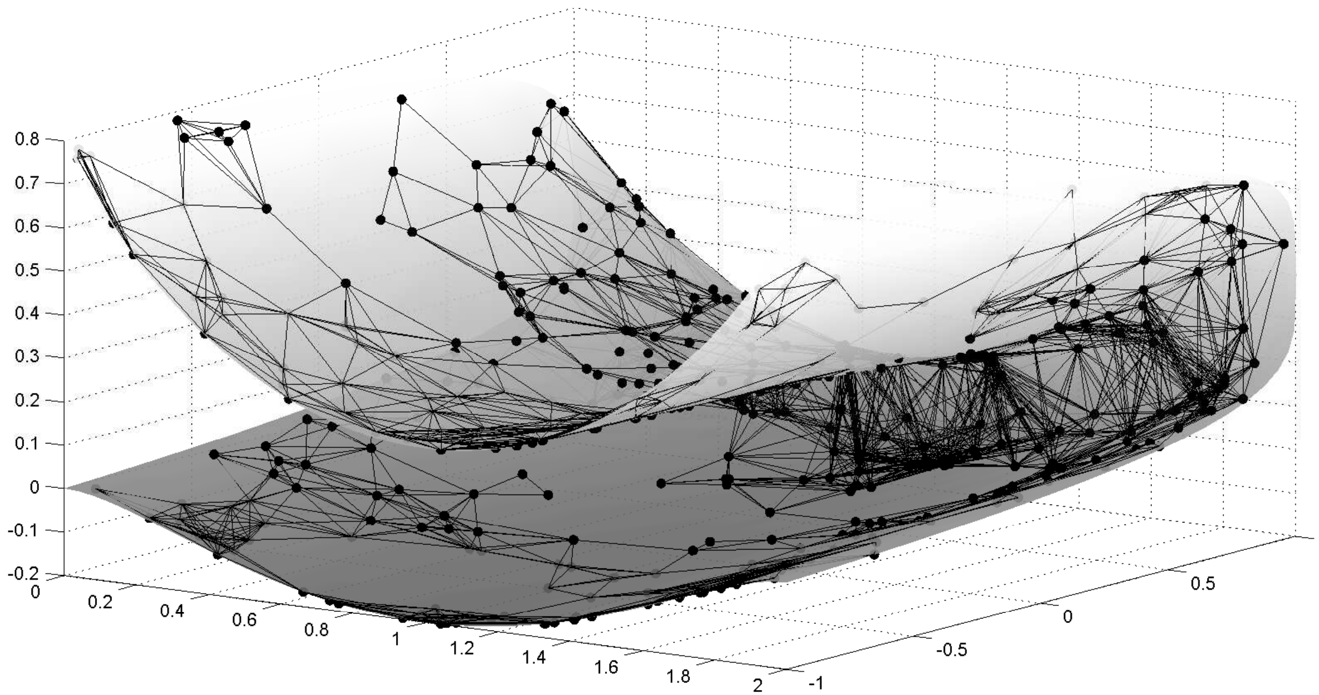


Biologisches Netzwerk: Ein Protein-Wechselwirkungs-Netzwerk.

Beispiele von Graphen III

Ähnlichkeitsgraph: zwei Pixel(Knoten) sind ähnlich, wenn ihr Grauwert ähnlich ist ⇒ Clustering entspricht Bildsegmentierung.

Beispiele von Graphen III



Nachbarschaftsgraph: zwei Punkte werden verbunden, wenn sie räumlich nah sind - der Graph approximiert die Oberfläche auf der die Punkte liegen.

Was haben alle Beispiele gemeinsam ?

Was haben alle Beispiele gemeinsam ?

Alle Graphen modellieren Beziehungen zwischen den Knoten !

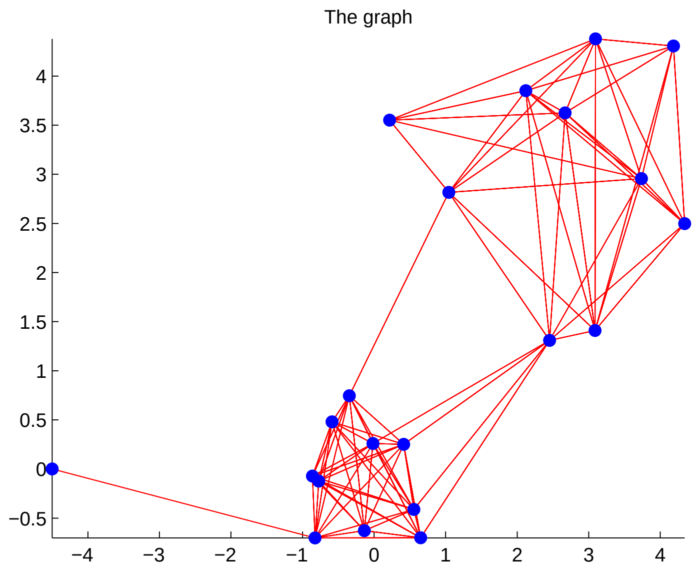
- sehr flexible Datenstruktur,
- keine “absoluten” Daten, nur “relationale” Daten,
- mit einem (selbstdefinierten) Maß für die Ähnlichkeit zwischen zwei Datenpunkten können wir einen Graphen auf fast allen Daten bilden.

Definition 3. *Ein* ***Clustering*** *ist eine Gruppierung der Daten, so daß alle Daten in einem Cluster ähnlich sind und Daten in verschiedenen Clustern unähnlich sind.*

Frage:

Was wäre ein gutes Kriterium um einen Graph in zwei Cluster zu zerteilen, wenn wir die Kantengewichte mit der Ähnlichkeit der Knoten assoziieren (hohes Gewicht=sehr ähnlich) ?

Outlier graph



Führt das Kriterium hier zu einem guten Clustering ?

Definitionen:

- Sei $C \subset V$ dann bezeichnet $\overline{C} := V \backslash C$ das Komplement von C.
- Gegeben eine Partition $C, \overline{C}$ des Graphen definieren wir den **Schnitt** als

$$\mathrm{cut}(C, \overline{C}) = \sum_{i \in C,\, j \in \overline{C}} w_{ij}.$$

- Der Ratio-Cut RCUT ist definiert als

$$\mathrm{RCUT}(C, \overline{C}) = \frac{\mathrm{cut}(C, \overline{C})}{|C|} + \frac{\mathrm{cut}(C, \overline{C})}{|\overline{C}|}.$$

Ratio-Cut II

- Sei $C \subset V$ dann bezeichnet $\overline{C} := V \backslash C$ das Komplement von C.
- Der Grad d_i eines Knotens i ist definiert als $d_i = \sum_{j=1}^{n} w_{ij}$.
- Der Ratio-Cut RCUT ist definiert als

$$\mathrm{RCUT}(C, \overline{C}) = \frac{\mathrm{cut}(C, \overline{C})}{|C|} + \frac{\mathrm{cut}(C, \overline{C})}{|\overline{C}|}.$$

Optimaler ratio cut: $\mathrm{RCUT}^* = \min_{C \subset V} \mathrm{RCUT}(C, \overline{C})$.

Aufgaben:

- Wann gilt $\mathrm{RCUT}^* = 0$?
- Zeige: $0 \leq \mathrm{RCUT}^* \leq \max_{i \in V} d_i$,
- Was ist der optimale Ratio-Cut eines vollständig verbundenen Graphen mit Kantengewichten 1 ?
- Welcher verbundene Graph hat den niedrigsten Ratio-Cut ?

Brute Force Lösung:

- Wieviele verschiedene Partitionen eines Graphen mit n Knoten gibt es ?
- Ein Bild hat heute ca. 1 Million Pixel - kann man alle möglichen Segmentierungen zu testen ?

Lösung des Ratio-Cut-Problems

Brute Force Lösung:

- Wieviele verschiedene Partitionen eines Graphen mit n Knoten gibt es ?
- Ein Bild hat heute ca. 1 Million Pixel - kann man alle möglichen Segmentierungen zu testen ?

Relaxierung:

- Das Optimierungsproblem für den Ratio-Cut wird umformuliert in ein äquivalentes Problem bei dem über alle Funktionen, die nur zwei Werte annehmen, minimiert werden.
- Relaxierung: anstatt nur über die zwei-wertigen Funktionen minimiert man über alle Funktionen $\Rightarrow$ Lösung des relaxierten Problems führt auf ein Eigenwert-Problem. Dieses kann noch für Graphen mit Millionen von Knoten berechnet werden.

Umformulierung des Problems

Gegeben eine Partition $C, \overline{C}$ definiere $f : V \to \mathbb{R}$,

$$f_i = \begin{cases} \sqrt{|\overline{C}|/|C|} & i \in C, \\ -\sqrt{|C|/|\overline{C}|} & i \in \overline{C}. \end{cases}$$

$$\begin{aligned} \langle f, (D-W)f \rangle =& \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 = \sum_{i \in C, j \in \overline{C}} w_{ij} \left(\sqrt{\frac{|\overline{C}|}{|C|}} + \sqrt{\frac{|C|}{|\overline{C}|}} \right)^2 \\ =& \mathrm{cut}(C,\overline{C}) \left(\frac{|\overline{C}|}{|C|} + \frac{|C|}{|\overline{C}|} + 2 \right) = \mathrm{cut}(C,\overline{C}) \left(\frac{|C| + |\overline{C}|}{|C|} + \frac{|C| + \overline{C}}{|\overline{C}|} \right) \\ =& |V| \ \mathrm{RCUT}(C,\overline{C}) \end{aligned}$$

$$\sum_{i=1}^{n} f_i = \sum_{i \in C} \sqrt{\frac{|\overline{C}|}{|C|}} - \sum_{i \in \overline{C}} \sqrt{\frac{|C|}{|\overline{C}|}} = 0, \qquad \|f\|^2 = \sum_{i=1}^{n} f_i^2 = |C| \frac{|\overline{C}|}{|C|} + |\overline{C}| \frac{|C|}{|\overline{C}|} = n.$$

Damit ist der optimale ratio cut

$$\mathrm{RCUT}^* = \min_{C \subset V} \Big\{ \langle f, (D-W)f \rangle \;\big|\; \langle f, \mathbb{1} \rangle = 0,\ \|f\| = \sqrt{n} \Big\}.$$

Dies ist immer noch ein kombinatorische Optimierungsproblem.

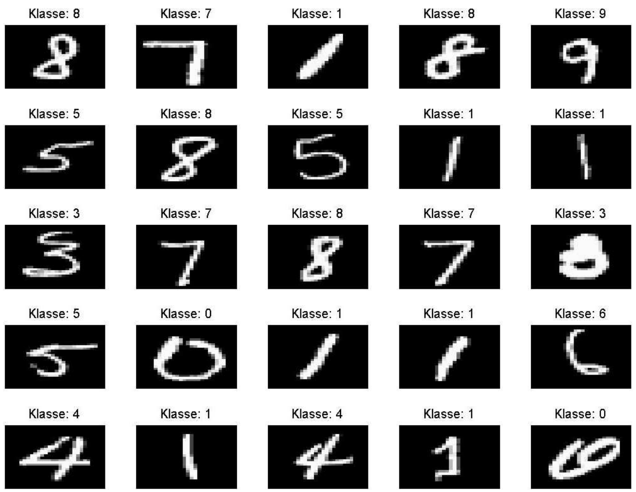
$\Rightarrow$ erlaube beliebige Funktionen f,

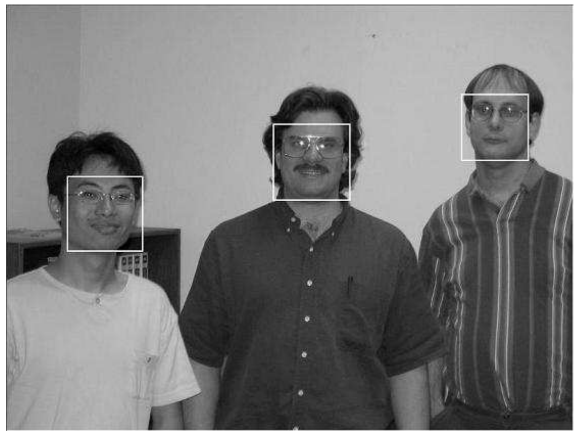
$$\min_{f \in \mathbb{R}^V} \Big\{ \langle f, (D-W)f \rangle \;\big|\; \langle f, \mathbb{1} \rangle = 0,\ \|f\| = \sqrt{n} \Big\}.$$

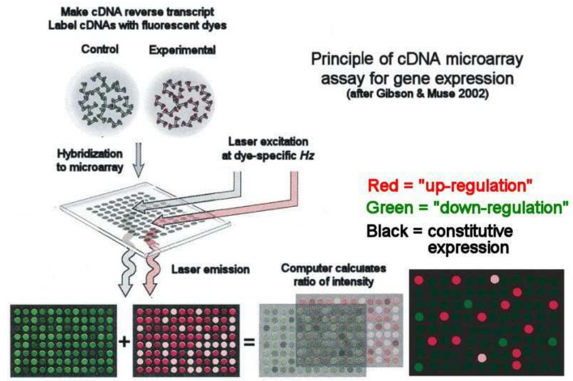
- Rayleigh-Ritz principle $\Rightarrow$ Minimum ist der zweite Eigenwert von $D-W$
- Partition durch optimales Thresholding $C_t = \{j \in V \,|\, u_2(j) > t\} \rightarrow \min_{t \in \mathbb{R}} \mathrm{RCUT}(C_t, \overline{C_t})$.

Segmentierung von Bildern mittels Clustering.

Machine Learning: Core Lecture 4+2, We 14-16, Fr 10-12 in HS III.







Content:

- Optimal decisions under uncertainty ? (Bayesian decision theory)
- Classification and Regression (SVM, Boosting, ...)
- Feature Selection, Dimensionality Reduction, Clustering.
- Statistical Learning Theory and many more.