Vorlesung Wissensentdeckung in Datenbanken

Matrix-Faktorisierung[1]

Kristian Kersting, (Katharina Morik), Claus Weihs

LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

26.06.2014

[1] Vielen Dank an Prof. Dr. Christian Bauckhage (Uni Bonn, Fraunhofer IAIS) für die Folien.

Motivation

- Die Faktorisierung von ***Datenmatrizen*** ist ein zentrales Tool im Data Mining, der Statistik, dem Maschinellen Lernen und dem Information Retrieval.
- Heutige Datenmatrizen sind typischerweise groß
- Beispiele sind:
  - Genexpressionsdaten und Microarrays
  - (Kollektionen von) digitalen Bildern
  - Wort-Dokument Matrizen
  - Adjazenzmatrizen von Graphen (z.B. soziale Netzwerke)
  - Datenbank-Indizes

Motivation

In der heutigen Vorlesung

- Matrix-Faktorisierung zur Datenanalyse
- Wir werden sehen, dass einige Aufgaben/methoden als Matrix-Faktorisierung formuliert werden können.
- Insbesondere werden wir auf die *Geometrie* des Problems eingehen und
- die Faktorisierung von sehr großen Datenmatrizen betrachten
- Zur Erinnerung

$$n \text{ Datenpunkte } \boldsymbol{x}_j \in \mathbb{R}^m$$
$$\Leftrightarrow \quad \text{Datenmatrix } \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Das heutige Menu

1. Matrix-Faktorisierung (MF): Was ist das?
2. Singulärwertzerlegung als MF
3. k-Means Cluster Analyse als MF und seine Varianten
4. Archetypenanalyse (AA)
5. Anwendungen der AA
6. Was haben wir gelernt?

Matrix Factorization: what?

Aufgabestellung:

- Sei eine Datenmatrix gegeben

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$$

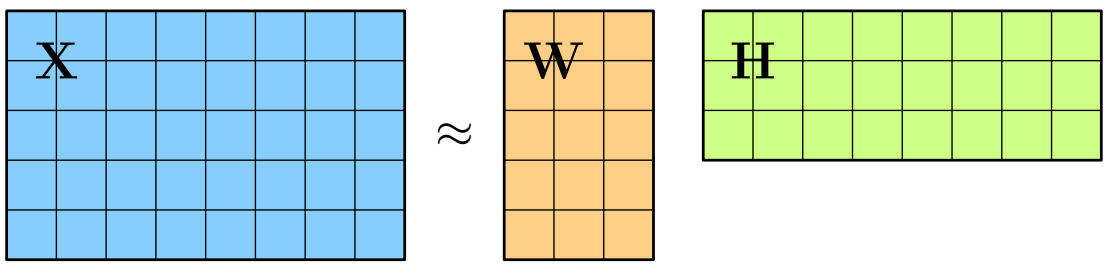
- Wähle eine natürliche Zahl $1 \leqslant k \leqslant \min\{m, n\}$ und
- bestimme zwei Matrizen — die Faktoren —

$$\boldsymbol{W} \in \mathbb{R}^{m \times k}$$

$$\boldsymbol{H} \in \mathbb{R}^{k \times n}$$

so dass

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}$$



Matrix-Faktorisierung (MF): Erinnerung

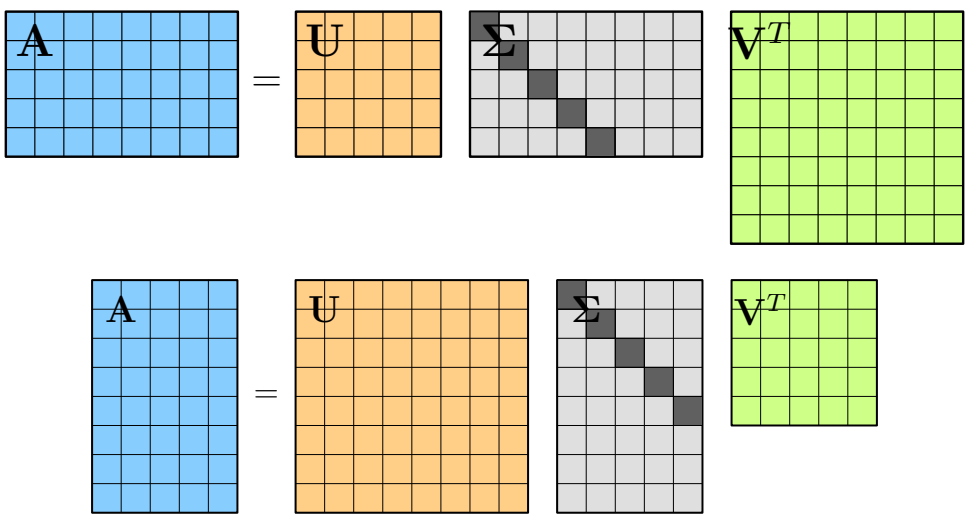
Theorem:

- Jede $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ kann als $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ zerlegt werden, wobei

  $\boldsymbol{U} \in \mathbb{R}^{m \times m} \Leftrightarrow$ orthogonale Matrix der Links-Singulärvektoren

  $\boldsymbol{V} \in \mathbb{R}^{n \times n} \Leftrightarrow$ orthogonale Matrix der Rechts-Singulärvektoren

  $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n} \Leftrightarrow$ Diagonalmatrix der Singulärwerte $\sigma_1, \ldots \sigma_{\min\{n,m\}}$



Matrix-Faktorisierung (MF): SVD

Wchtige Eigenschaft:

- Das Optimierungsproblem

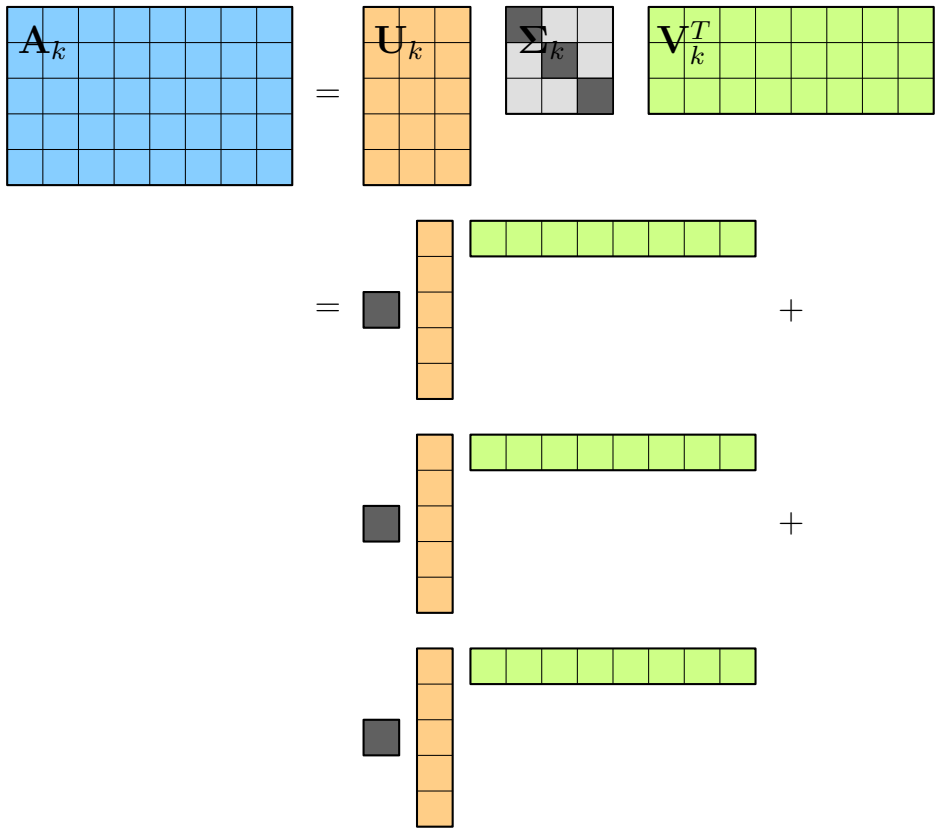
$$\min_{\mathrm{rk}(\boldsymbol{A}_k)=k} \left\| \boldsymbol{A} - \boldsymbol{A}_k \right\|^2$$

hat die Lösung

$$\begin{aligned} \boldsymbol{A}_k &= \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^T \\ &= \sum_{j=1}^{k} \sigma_j \, \boldsymbol{u}_j \, \boldsymbol{v}_j^T \end{aligned}$$

mit den Residuum

$$\left\| \boldsymbol{A} - \boldsymbol{A}_k \right\|^2 = \sum_{j=k+1}^{r} \sigma_j^2$$

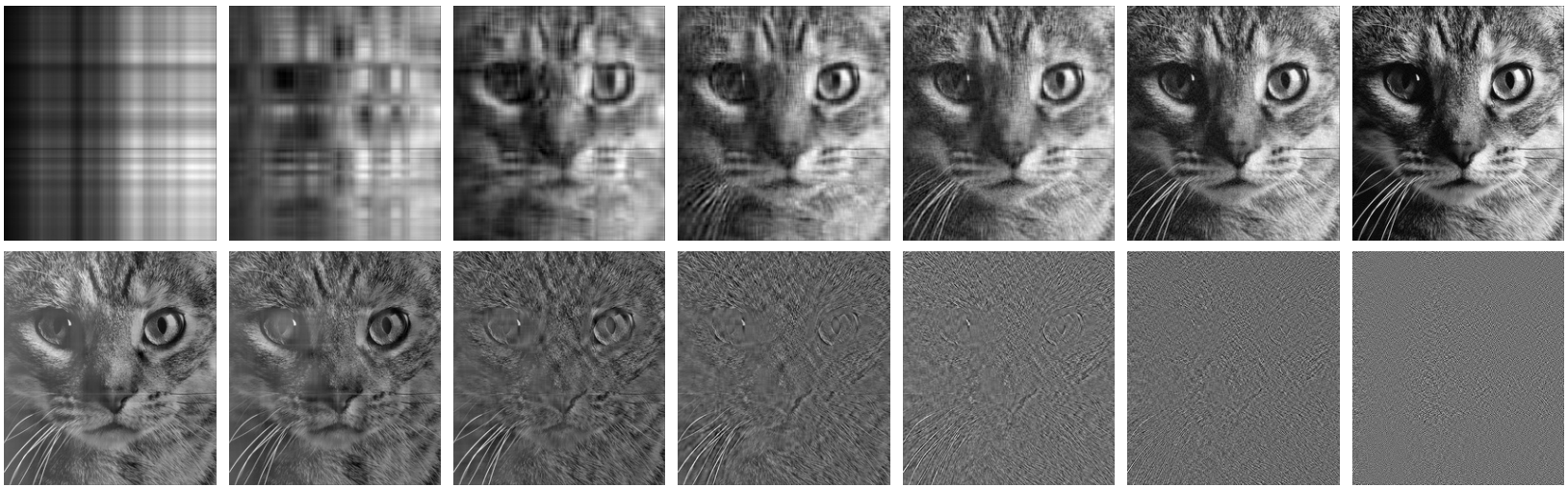


Matrix-Faktorisierung (MF): SVD

Beispiel:

- Matrix $\boldsymbol{A} \in \mathbb{R}^{256 \times 232}$ mit $0 \leqslant a_{ij} \leqslant 255$

- Approximationen $\boldsymbol{A}_k$ und die Residuen für $k \in \{1, 3, 9, 18, 36, 72, 144\}$



Matrix-Faktorisierung (MF): SVD ist eine Matrix-Faktorisierung

Beobachtung:

- Um $\boldsymbol{X} \approx \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^T$ zu approximieren, können wir

$$\boldsymbol{W} = \boldsymbol{U}_k$$
$$\boldsymbol{H} = \boldsymbol{\Sigma}_k \boldsymbol{V}_k^T$$

substituieren und dann das folgende Optimierungsproblem lösen:

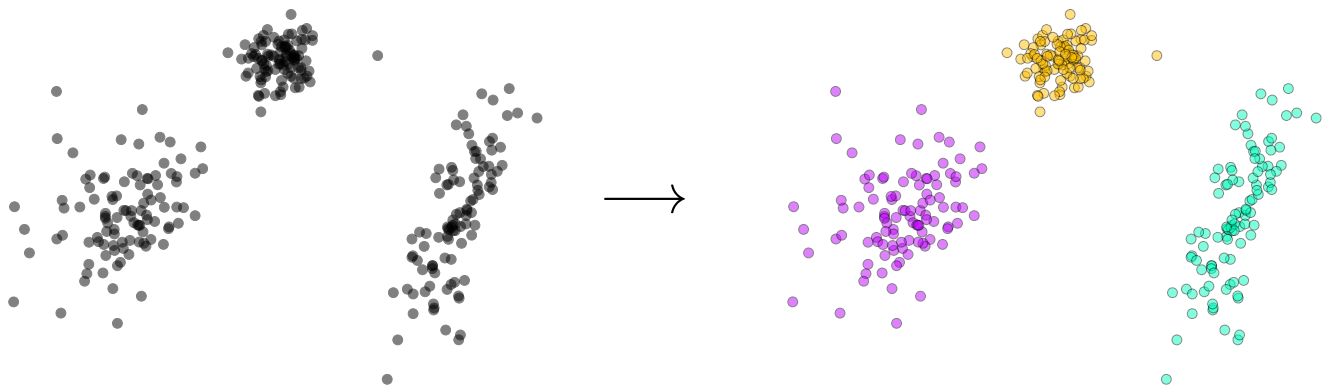
$$\min_{\boldsymbol{W},\boldsymbol{H}} \left\| \boldsymbol{X} - \boldsymbol{W}\boldsymbol{H} \right\|^2 \quad \text{s.t.} \quad \boldsymbol{W}^T \boldsymbol{W} = \boldsymbol{I}$$

$\Rightarrow$ Die Singulärwertzerlegung (SV) kann als Matrix-Faktorisierung unter Nebenbedingungen formuliert werden!

k-Means Cluster Analyse: Erinnerung

Aufgabe:

- Gegeben eine Datenmenge (/-matrix) $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^m$
- finde k Clusterzentren $C_i \subset X$ so dass
  1. Datenpunkte, die zu C_i zugewiesen sind, sind zu einander ähnlich, und
  2. $C_i \cap C_j = \emptyset$ für $i \neq j$



k-Means Cluster Analyse: Erinnerung

Prinzipielles Vorgehen:

- Wähle eine *Ähnlichkeitfunktion* und definiere *Clusters* mittels Zentroiden $\boldsymbol{\mu}_i \in \mathbb{R}^m$

$$C_i = \left\{ \boldsymbol{x}_j \in X \,\middle|\, \left\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\right\|^2 \leqslant \left\|\boldsymbol{x}_j - \boldsymbol{\mu}_l\right\|^2 \, \forall \, l \neq i \right\}$$

- Wir bestimmen eine "gute" Menge von Zentroiden
- und betrachten die folgende Zielfunktion

$$E(k) = \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in C_i} \left\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\right\|^2$$

k-Means Cluster Analyse etwas anders betrachtet

- k-means lernt implizit eine Mischverteilung von Gauß'schen Zufallsvariable
- k-means ist NP-schwer (Aloise et al., ML, 75(2), 2009)
- k-means ist "nur" eine Heuristik

Um die Mischverteilung zu sehen, führen wir Indikatorvariablen ein:

$$r_{ij} = \begin{cases} 1, & \text{falls } \boldsymbol{x}_j \in C_i \\ 0, & \text{sonst} \end{cases}$$

Damit können wir die Zielfunktion schreiben als

$$E(k) = \sum_{i=1}^{k} \sum_{j=1}^{n} r_{ij} \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_i \right\|^2$$

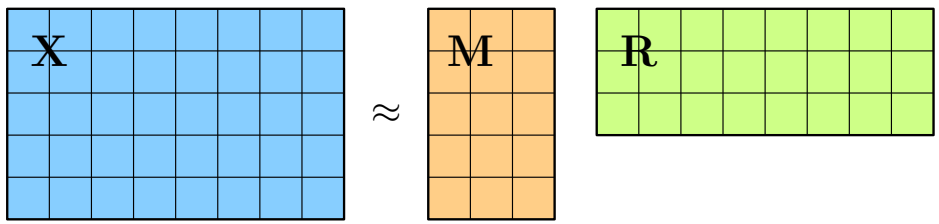
Warum ist das spannend?

k-Means Cluster Analyse als Matrix-Faktorisierung

k-means führt eine Matrix-Faktorisierung durch!

Wenn es konvergiert, dann haben wir

$$\boldsymbol{X} \approx \boldsymbol{MR}$$
$$\Leftrightarrow \quad \boldsymbol{x}_j \approx \boldsymbol{M}\boldsymbol{r}_j$$



mit

$\boldsymbol{X} \in \mathbb{R}^{m \times n} \Leftrightarrow$ data matrix

$\boldsymbol{M} \in \mathbb{R}^{m \times k} \Leftrightarrow$ centroid matrix

$\boldsymbol{R} \in \mathbb{R}^{k \times n} \Leftrightarrow$ indicator matrix

k-Means Cluster Analyse als Matrix-Faktorisierung

Die k-means Cluster-Analyse löst das folgende Optimierungproblem

$$\min_{\boldsymbol{M},\boldsymbol{R}} \left\| \boldsymbol{X} - \boldsymbol{M}\boldsymbol{R} \right\|^2$$

mit verschiedenen Nebenbedingungen

- Jeder Zentroid ist ein konvexe Kombination von Datapunkten

$$\mu_i = \boldsymbol{X}\boldsymbol{s}_i \quad \Leftrightarrow \quad \boldsymbol{M} = \boldsymbol{X}\boldsymbol{S}$$

- $\boldsymbol{S}$ ist zeilen-stochastisch ($\boldsymbol{s}_i$ besteht aus 0s und n_i Werten in $\frac{1}{n_i}$)
- $\boldsymbol{R}$ is zeilen-stochastisch und binär ($\boldsymbol{r}_j$ enthält genau eine 1)

k-Means Cluster Analyse: Umschreiben der Nebenbedingungen

Nebenbedingungen an S:

$$s_{ji} \geqslant 0 \Leftrightarrow \quad \boldsymbol{S} \succeq \mathbf{0}$$

$$\sum_{j=1}^{n} s_{ji} = 1 \Leftrightarrow \quad \mathbf{1}^T \boldsymbol{s}_i = 1$$

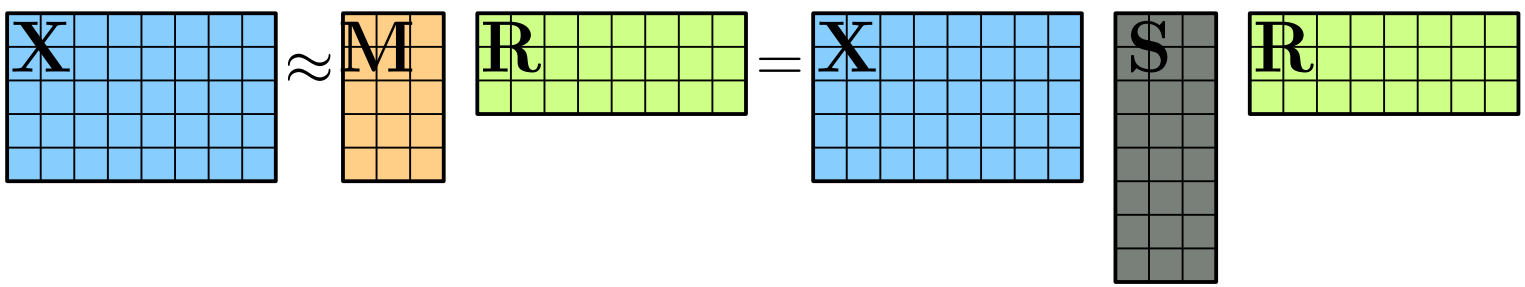
$$s_{ji} \in \left\{0, \frac{1}{n_i}\right\} \Leftrightarrow H(\boldsymbol{s}_i) = -\sum_{j=1}^{n} s_{ji} \log s_{ji} \gg 0$$

Nebenbedingungen an R:

$$r_{ij} \geqslant 0 \Leftrightarrow \quad \boldsymbol{R} \succeq \mathbf{0}$$

$$\sum_{i=1}^{k} r_{ij} = 1 \Leftrightarrow \quad \mathbf{1}^T \boldsymbol{r}_j = 1$$

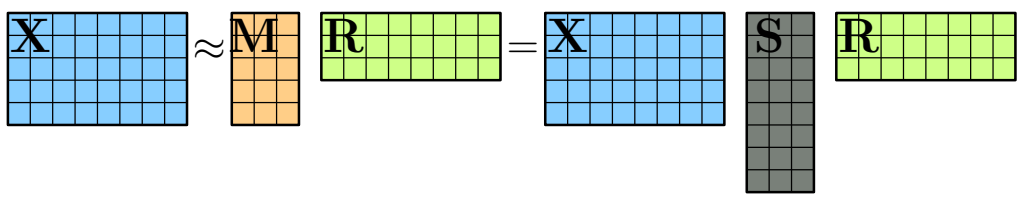
$$r_{ij} \in \{0, 1\} \Leftrightarrow H(\boldsymbol{r}_j) = -\sum_{i=1}^{k} r_{ij} \log r_{ij} = 0$$



k-Means Cluster Analyse: Optimierungsproblem der "strikten" Variante

hard *k*-means clustering

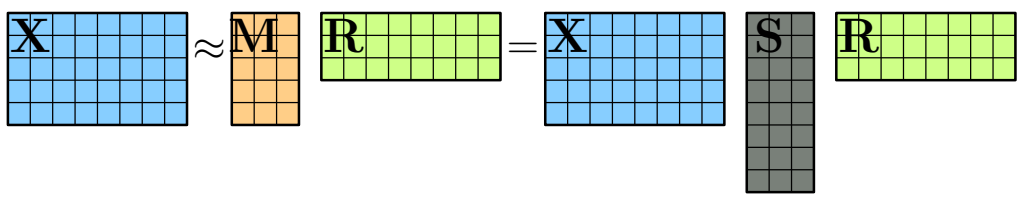
$$\begin{aligned} \min_{\boldsymbol{S},\boldsymbol{R}} \left\| \boldsymbol{X} - \boldsymbol{XSR} \right\|^2 - \sum_{i=1}^{k} H(\boldsymbol{s}_i) \\ \boldsymbol{S} \succeq \mathbf{0} \\ \mathbf{1}^T \boldsymbol{s}_i = 1 \\ \text{s.t.} \quad \boldsymbol{R} \succeq \mathbf{0} \\ \mathbf{1}^T \boldsymbol{r}_j = 1 \\ H(\boldsymbol{r}_j) = 0 \end{aligned}$$



k-Means Cluster Analyse: Optimierungsproblem der "soften" Variante

soft *k*-means clustering

$$\begin{aligned} \min_{\boldsymbol{S},\boldsymbol{R}} \left\| \boldsymbol{X} - \boldsymbol{XSR} \right\|^2 - \sum_{i=1}^{k} H(\boldsymbol{s}_i) \\ \boldsymbol{S} &\succeq \mathbf{0} \\ \mathbf{1}^T \boldsymbol{s}_i &= 1 \\ \text{s.t.} \quad \boldsymbol{R} &\succeq \mathbf{0} \\ \mathbf{1}^T \boldsymbol{r}_j &= 1 \\ H(\boldsymbol{r}_j) &= 0 \end{aligned}$$



Wir haben das Problem also relaxiert!!!

k-Means Cluster Analyse: Andere Varianten

Können wir also weitere Varianten durch das Relaxieren des Optimierungsproblems erhalten?

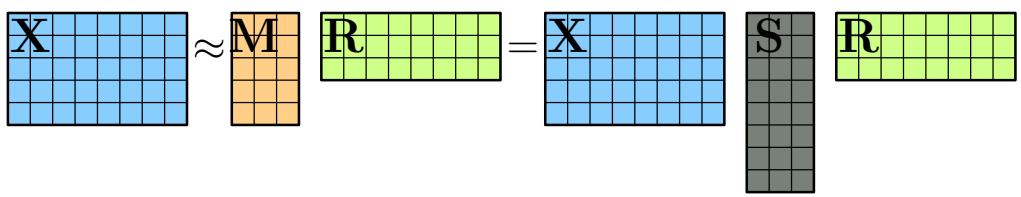
Und wenn ja, was bestimmen diese Varianten?

k-Means Cluster Analyse: Andere Varianten

Archetypenanalyse

$$\min_{\boldsymbol{S},\boldsymbol{R}} \left\| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{S}\boldsymbol{R} \right\|^2 - \sum_{i=1}^{k} H(\boldsymbol{s}_i)$$

$$\begin{aligned} \boldsymbol{S} &\succeq \mathbf{0} \\ \mathbf{1}^T \boldsymbol{s}_i &= 1 \\ \text{s.t.} \quad \boldsymbol{R} &\succeq \mathbf{0} \\ \mathbf{1}^T \boldsymbol{r}_j &= 1 \\ H(\boldsymbol{r}_j) &= 0 \end{aligned}$$



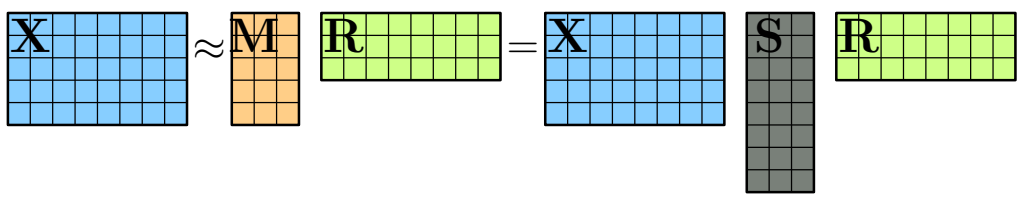
Stellt alle Datenpunkte durch extreme Datenpunkte dar, den Archetypen.

k-Means Cluster Analyse: Andere Varianten

Nicht-negative Matrix-Faktorisierung

$$\min_{\boldsymbol{S},\boldsymbol{R}} \left\| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{S}\boldsymbol{R} \right\|^2 - \sum_{i=1}^{k} H(\boldsymbol{s}_i)$$

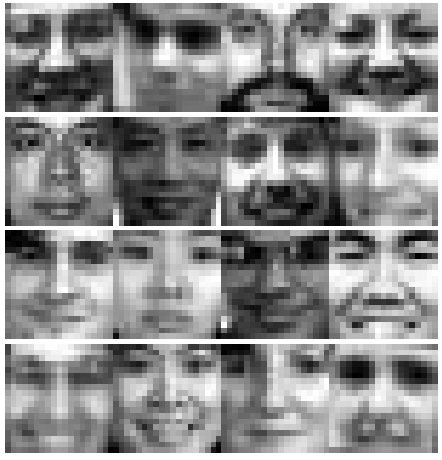
$$\begin{aligned} \boldsymbol{S} &\succeq \mathbf{0} \\ \mathbf{1}^T \boldsymbol{s}_i &= 1 \\ \text{s.t.} \quad \boldsymbol{R} &\succeq \mathbf{0} \\ \mathbf{1}^T \boldsymbol{r}_j &= 1 \\ H(\boldsymbol{r}_j) &= 0 \end{aligned}$$



Findet oft einfacher zu interpretierende Cluster in den Daten.

k-Means Cluster Analyse: Vergleich der Varianten

Daten (einige Samples)



SVD

NMF

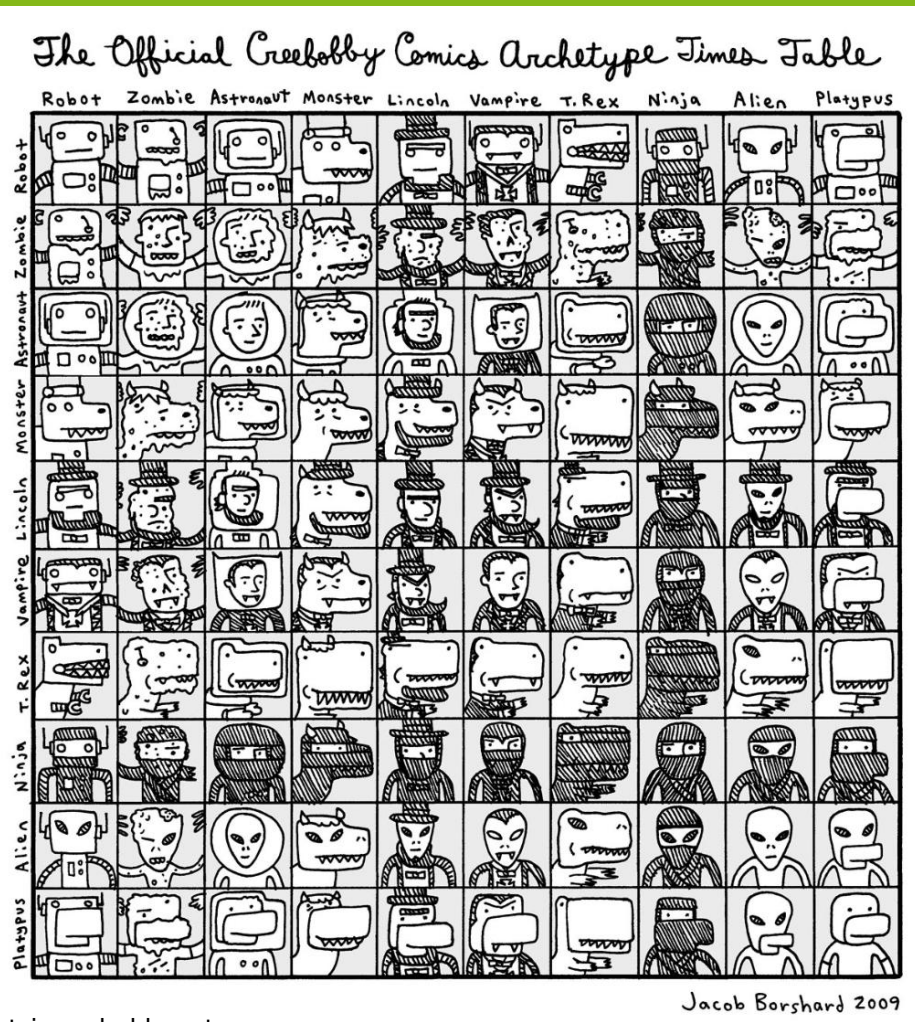
k-means

AA

Archetypenanalyse (AA): Andere Varianten

Im Folgenden wollen wir die Archetypenanalyse etwas genauer betrachten

Archetypen



- **Plato**:

  ideals and/or pure forms that embody fundamental characteristics of a thing rather than its specific peculiarities

- **C.G. Jung**:

  innate, universal forms (the *hero*, the *great mother*, the *wise old man*, . . .) that channel experiences and emotions, resulting in recognizable and typical behaviors with certain probable outcomes

- **A. Cutler** and **L. Breiman** (in Technometrics 36(4), 1994):

  Archetypenanalyse $\Leftrightarrow$ ein "neuer" Ansatz zur Analyse von multivariaten Daten, der zur effizienten Faktorisierung von großen Datenmatrizen führt.

Archetypenanalyse (AA)

Aufgabe:

- Gegen eine Datenmatrix

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$$

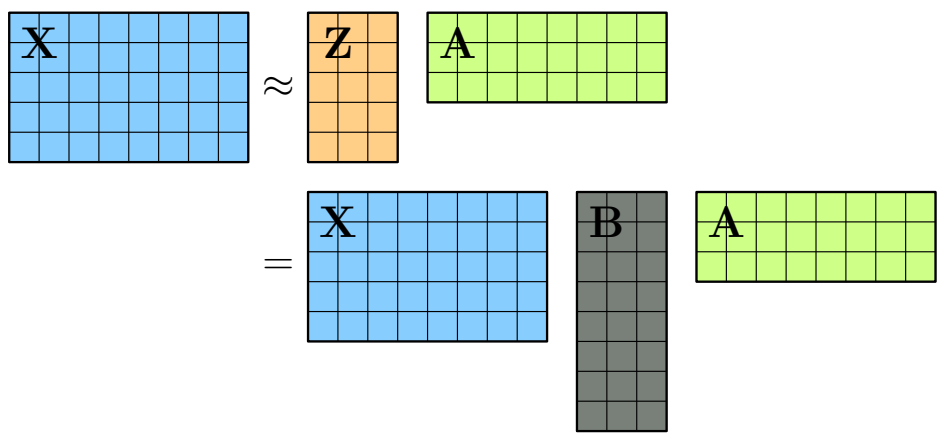
- wähle eine natürliche Zahl $1 \leqslant k \leqslant \min\{m, n\}$ und
- bestimme zwei zeilen-stochastische Matrizen — die Faktoren —

$$\boldsymbol{B} \in \mathbb{R}^{n \times k}$$

$$\boldsymbol{A} \in \mathbb{R}^{k \times n}$$

so dass

$$\begin{aligned} \boldsymbol{X} &\approx \boldsymbol{ZA} \\ &= \boldsymbol{XBA} \end{aligned}$$



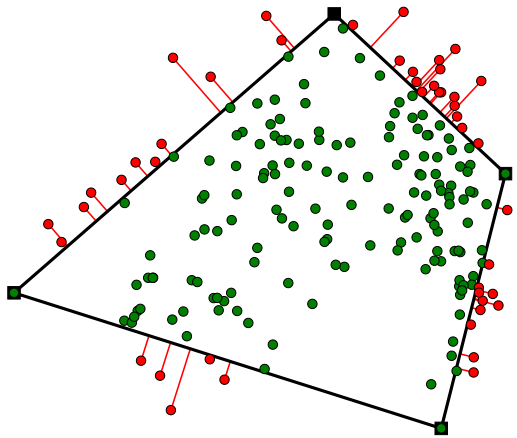
Archetypenanalyse (AA)

Ansatz:

- Löse das folgende Optimierungsproblem

$$\min_{\boldsymbol{B},\boldsymbol{A}} \left\| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A} \right\|^2$$

$$\text{s.t.} \quad \begin{aligned} \boldsymbol{B} &\succeq \mathbf{0},\ \mathbf{1}^T \boldsymbol{b}_i = 1 \\ \boldsymbol{A} &\succeq \mathbf{0},\ \mathbf{1}^T \boldsymbol{a}_j = 1 \end{aligned}$$



Beobachtung:

- Einsetzen von $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} \in \mathbb{R}^{m \times k}$ ergibt

$$\boldsymbol{z}_i = \sum_{j=1}^{n} \boldsymbol{x}_j \, b_{ji} \qquad \text{and} \qquad \boldsymbol{x}_j \approx \sum_{i=1}^{k} \boldsymbol{z}_i \, a_{ij}$$

Archetypenanalyse (AA): Semantik

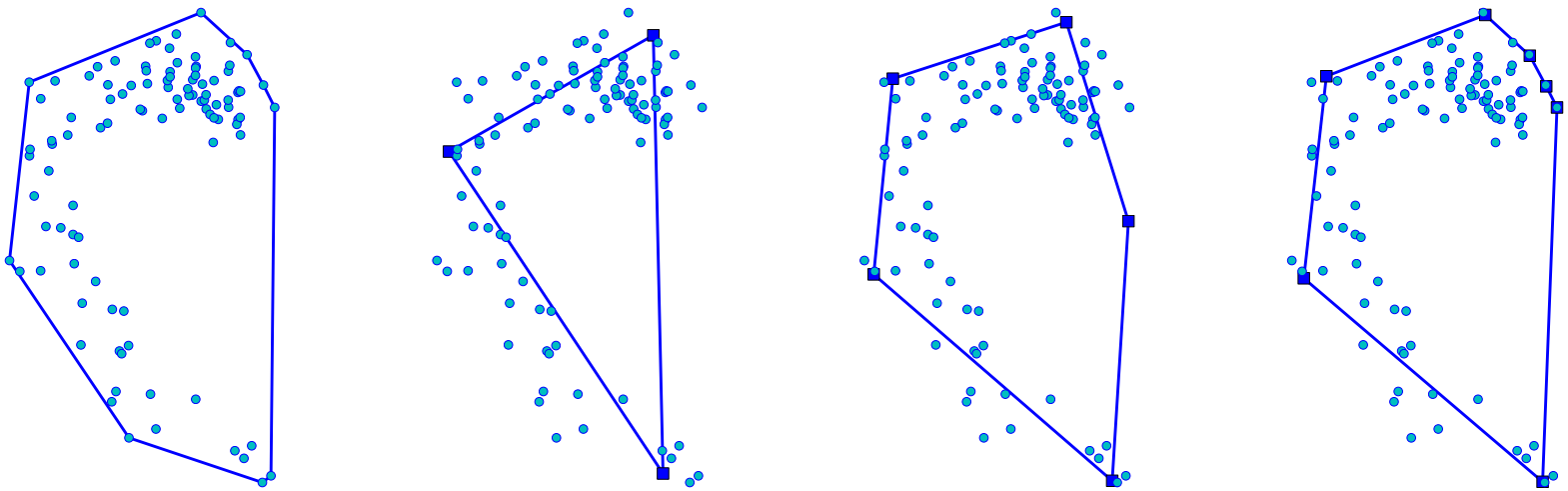
Oder anders gesagt:

- Die Archetypen $\boldsymbol{z}_i$ sind spärliche konvexe Kombinationen der Datenpunkte $\boldsymbol{x}_j$
- Die Datenpunkte $\boldsymbol{x}_j$ sind spärliche konvexe Kombinationen der Archetypen $\boldsymbol{z}_i$

The four basic personality types

Archetypenanalyse (AA): Eigenschaften

- Für $k > 1$ liegen die Archetypen auf der konvexen Hülle der Datenmenge
- Je größer k desto besser approximiert die "archetypische Hülle" die konvexe Hülle der Daten



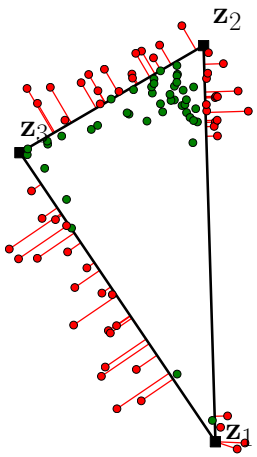
convex hull $k = 3$ $k = 5$ $k = 7$

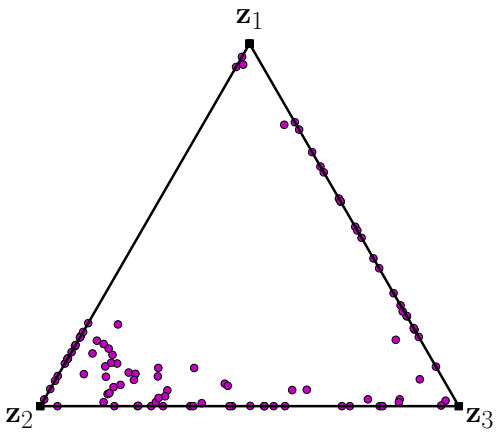
Archetypenanalyse (AA): Eigenschaften

- Die "Koeffizienten"-vektoren $\boldsymbol{a}_j$ in $\boldsymbol{x}_j = \boldsymbol{Z}\,\boldsymbol{a}_j$ sind stochastisch

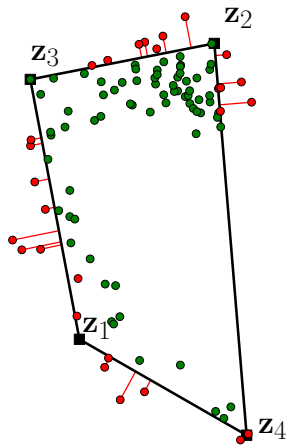
$\Rightarrow$ Wir können a_{ij} als $p(\boldsymbol{x}_j \mid \boldsymbol{z}_i)$ interpretieren

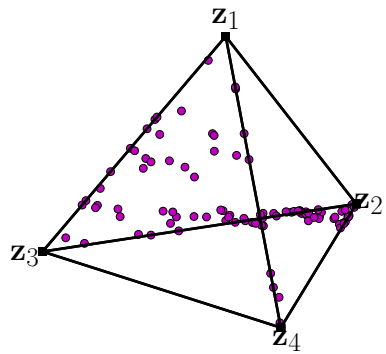
$\Rightarrow$ (soft)clustering, classification, ranking, ...





$k = 3$

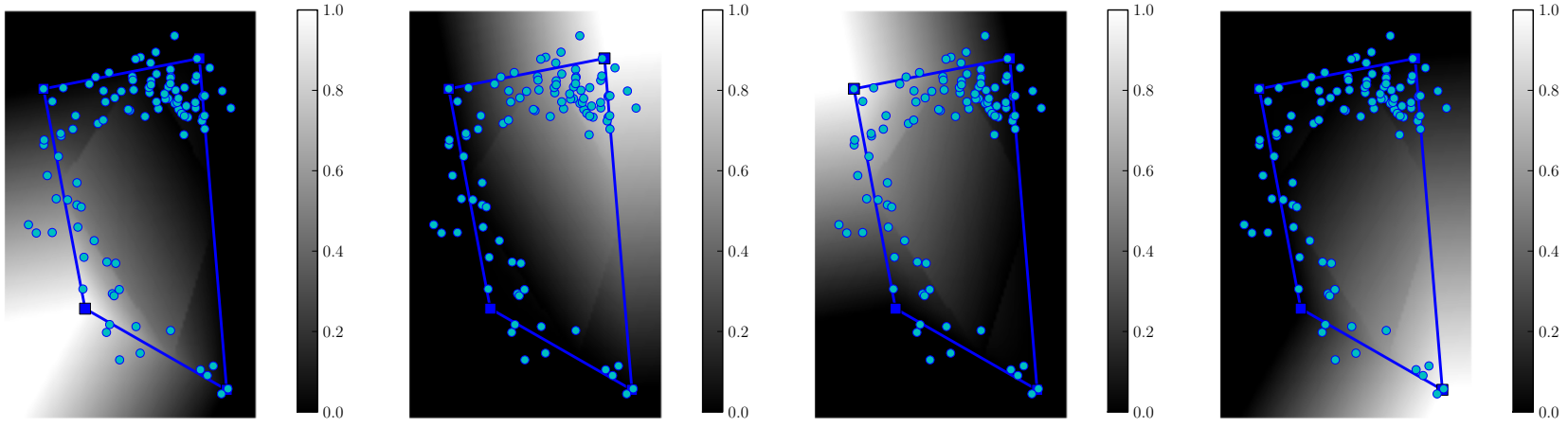




$k = 4$

Archetypenanalyse (AA): Eigenschaften

Die Archetypenanalyse verbindet Geometrie mit Wahrscheinlichkeiten



$p(\boldsymbol{x} \mid \boldsymbol{z}_1)$ $p(\boldsymbol{x} \mid \boldsymbol{z}_2)$ $p(\boldsymbol{x} \mid \boldsymbol{z}_3)$ $p(\boldsymbol{x} \mid \boldsymbol{z}_4)$

Aber wie berechnen wir das denn nun?

Archetypenanalyse (AA): Algorithmen

Alternating Least Squares:

initialisiere $\boldsymbol{B}$ zufällig

wiederhole bis zur Konvergenz

Berechnet $\boldsymbol{Z} = \boldsymbol{XB}$

Löse unter Nebenbedingungen

$$\boldsymbol{A} = \text{argmin}_{\boldsymbol{A}} \left\| \boldsymbol{X} - \boldsymbol{ZA} \right\|^2$$

Berechne $\hat{\boldsymbol{Z}} = \boldsymbol{XA}^{\dagger}$

Löse unter Nebenbedingungen

$$\boldsymbol{B} = \text{argmin}_{\boldsymbol{B}} \left\| \hat{\boldsymbol{Z}} - \boldsymbol{XB} \right\|^2$$

Projektierte Gradienten:

$$\left\| \boldsymbol{X} - \boldsymbol{XBA} \right\|^2 = \text{tr}\left[\boldsymbol{X}^T\boldsymbol{X} - 2\boldsymbol{X}^T\boldsymbol{XBA} + \boldsymbol{A}^T\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{XBA} \right]$$

$$\frac{\partial E}{\partial \boldsymbol{A}} = 2\left[\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{XBA} - \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X} \right]$$

$$\frac{\partial E}{\partial \boldsymbol{B}} = 2\left[\boldsymbol{X}^T\boldsymbol{XBAA}^T - \boldsymbol{X}^T\boldsymbol{XA}^T \right]$$

Initialisiere $\boldsymbol{A}$ und $\boldsymbol{B}$ zufällig

wiederhole bis zur Konvergenz

$\boldsymbol{A} \leftarrow \boldsymbol{A} - \eta_A \frac{\partial E}{\partial \boldsymbol{A}}$, projektiere $\boldsymbol{a}_j$ auf Δ^{k-1}

$\boldsymbol{B} \leftarrow \boldsymbol{B} - \eta_B \frac{\partial E}{\partial \boldsymbol{B}}$, projektiere $\boldsymbol{b}_i$ auf Δ^{n-1}

Archetypenanalyse (AA): Zusammenfassung

- Achetypenanalyse (AA) machte keinerlei (direkten) Annahmen über die Datenverteilung
- AA ist NP-schwierig
- Beide bisherigen Ansätze sind Heuristiken
- beide benutzen (implizit) $\boldsymbol{X}^T\boldsymbol{X} \in \mathbb{R}^{n \times n}$

  $\Rightarrow$ Damit können wir den *Kerntrick* wider anwenden

  $\Rightarrow$ Beide Ansätze skalieren mit mindestens $O(n^2)$

$\Rightarrow$ Beide skalieren nicht zu BIG DATA

Archetypenanalyse (AA): Auf zu Big Data !

Drei wichtige Einsichten:

1. $\boldsymbol{Z} = \boldsymbol{XB}$ stellen im Wesentlichen *extreme* Datenpunkte dar
2. Etreme Datenpunkte können relative einfach berechnet werden.
3. Würden wir $\boldsymbol{Z}$ kennen, dann

$$\left\|\boldsymbol{X} - \boldsymbol{ZA}\right\|^2 = \left\|\boldsymbol{x}_1 - \boldsymbol{Z}\boldsymbol{a}_1\right\|^2 + \left\|\boldsymbol{x}_2 - \boldsymbol{Z}\boldsymbol{a}_2\right\|^2 + \ldots$$

Zwei Hauptideen:

1. Berechne AA nur auf einer angemessenen Teilmenge der Daten
2. Entkopple die Berechnung von $\boldsymbol{Z}$ und $\boldsymbol{A}$

Archetypenanalyse (AA): Literaturverweise

- C. Bauckhage and C. Thurau: Making Archetypal Analysis Practical, DAGM, 2009.
- C. Thurau, K. Kersting, and C. Bauckhage: Yes we can: simplex volume maximization for descriptive web-scale matrix factorization. ACM CIKM, 2010.
- C. Thurau, K. Kersting, M. Wahabzada, and C. Bauckhage: Convex non-negative matrix factorization for massive datasets. Knowledge and Information Systems, 29(2):457-478, 2011.
- K. Kersting, C. Bauckhage, C. Thurau, and M. Wahabzada: Matrix Factorization as Search. ECML/PKDD, 2012.
- C. Thurau, K. Kersting, M. Wahabzada, C. Bauckhage: Descriptive matrix factorization for sustainability Adopting the principle of opposites. Data Min. Knowl. Discov. 24(2): 325-354, 2012.

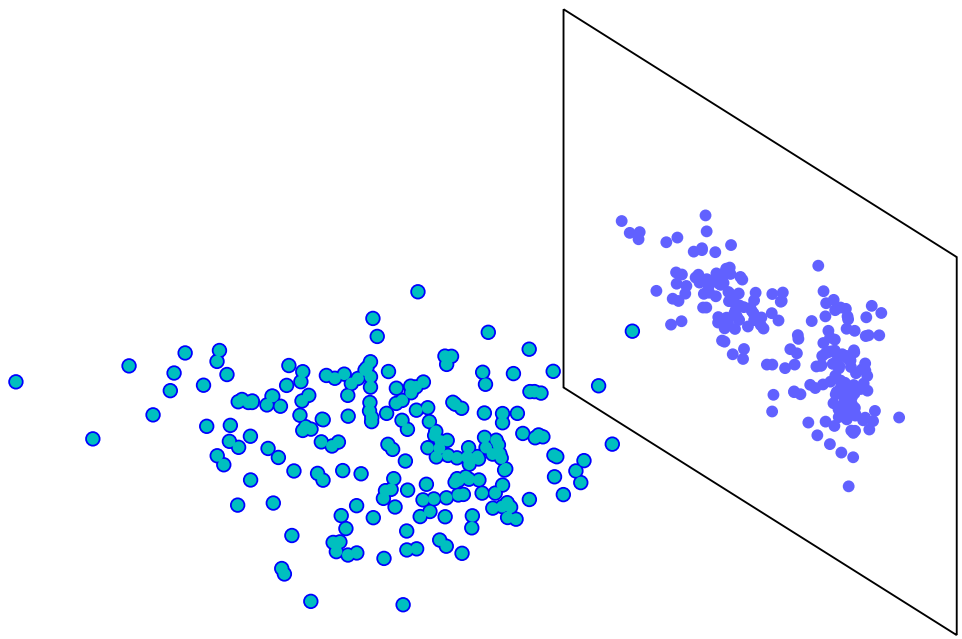
Archetypenanalyse (AA): Subsampling der konvexen Hülle

- Archetypen sind Mischungen von Punkten auf der konvexen Hülle der Daten

$\Rightarrow$ Lass AA nur auf der konven Hülle $X^H \subseteq X$ laufen

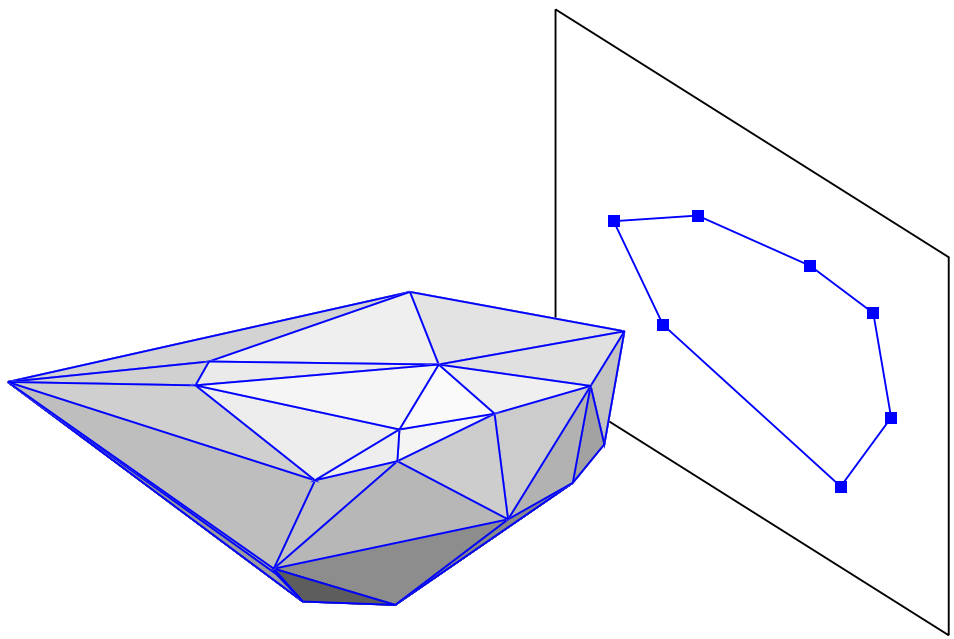
- Aber Achtung, in $\mathbb{R}^m$ ist die Berechnung der konvexen Hülle "teuer" ($O(n^{\lfloor m/2 \rfloor + 1})$)

$\Rightarrow$ Daher betrachten wir (viele) 2D Projektionen der Daten und berechnen ihre konvexen Hüllen (z.B. mittels PCA/SVD der Kovarianzmatrix und der paarweisen Kombination der ersten Eigenvektoren)



Archetypenanalyse (AA): Warum funktioniert das?

- Fals P ein Polytop ist (z.B. die konvex Hülle) and $\pi : \boldsymbol{x} \rightarrow \boldsymbol{M}\boldsymbol{x} + \boldsymbol{t}$ eine affine Abbildung,
  dann ist auch $\pi(P)$ ein Polytop
- Jeder Vertex von $\pi(P)$ korrespondiert zu einem Vertex von P
- Der Aufwand zum Berechnen der Vertices ist $O(\eta^2)$ mit $\eta \in \Omega(\sqrt{\log n})$. Sehr viel kleiner als im Ausgansraum!



Archetypenanalyse (AA): Mittels Distanzgeometrie

- Falls die Distanzen d_{ij} zwischen den k Vertices eines $k-1$ Simplex S bekannt sind, dann ist das Volumen des Simplex gerade

$$V(S)^2 = \frac{-1^k}{2^{k-1}\left((k-1)!\right)^2} \det(\boldsymbol{A})$$

wobei $\det(\boldsymbol{A})$ die sogenannte Cayley-Menger Determinante ist.

Cayley-Menger Determinante

$$\begin{vmatrix} 0 & 1 & 1 & 1 & \dots & 1 \\ 1 & 0 & d_{11}^2 & d_{12}^2 & \dots & d_{1k}^2 \\ 1 & d_{11}^2 & 0 & d_{22}^2 & \dots & d_{2k}^2 \\ 1 & d_{12}^2 & d_{22}^2 & 0 & \dots & d_{3k}^2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & d_{1k}^2 & d_{2k}^2 & d_{3k}^2 & \dots & 0 \end{vmatrix}$$

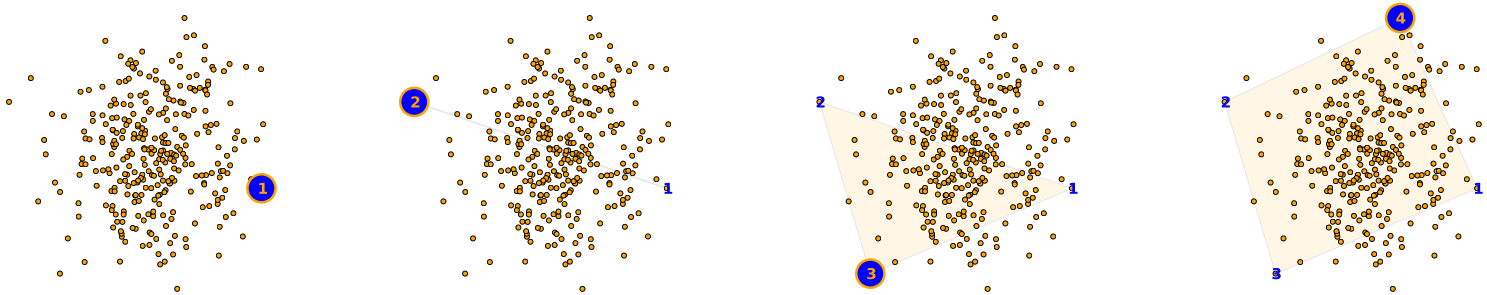
Archetypenanalyse (AA): Simplex Volume Maximization (SiVM)

- Bestimme $\boldsymbol{Z}$ durch das Einpassen eines Simplex mit maximalem Volume in die Daten

$\Rightarrow$ schneller Greedy-Heuristik

- Gegeben S mit $k-1$ Vertices, bestimme den nächsten Vertex $\boldsymbol{z}_k \in \boldsymbol{X}$ so, dass

$$\boldsymbol{z}_k = \underset{j}{\operatorname{argmax}}\ V(S \cup \boldsymbol{x}_j)^2$$



Archetypenanalyse (AA)
how?

greedy stochastic hill climbing:

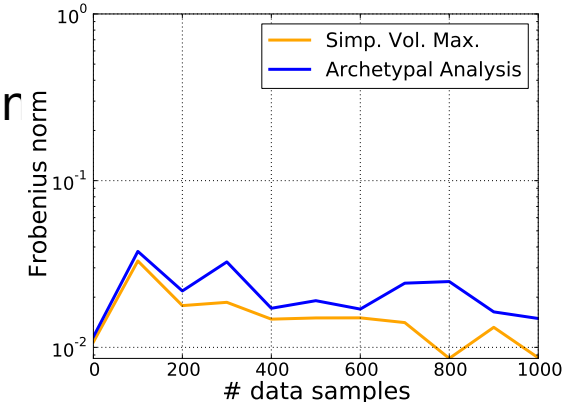
Wähle $Z \subset X$ zufällig

Iteriere über $\boldsymbol{x}_j \in X$

Iteriere über $\boldsymbol{z}_i \in Z$

Falls $V\left(Z \setminus \{\boldsymbol{z}_i\} \cup \{\boldsymbol{x}_j\}\right) > V(Z)$ dann

$$Z \leftarrow Z \setminus \{\boldsymbol{z}_i\} \cup \{\boldsymbol{x}_j\}$$

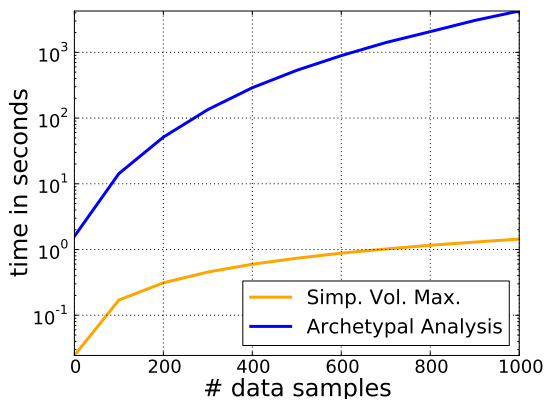


divide and conquer:

Teile die Daten aud in $X = X_1 \cup X_2 \cup \ldots \cup X_p$

Für jedes X_i, berechne $Z_i(X_i)$

Berechne $Z(Z_1 \cup Z_2 \cup \ldots \cup Z_p)$



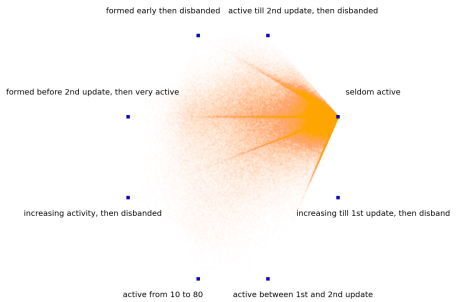
Archetypenanalyse (AA): Zusammenfassung

note:

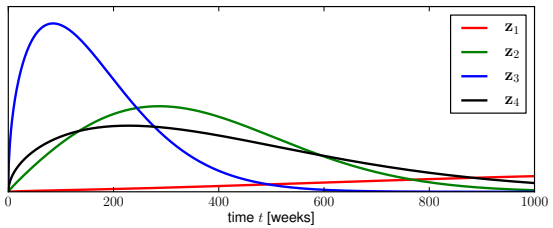
- AA analysiert Daten mittels extremen Datenpunkten
- Ergebnisse sind intuitive und einfach zu interpretieren
- Beruht implizit auf einer Distanzfunktion ...
- und kann daher in vielen Fällen eingesetzt werden
- Es skaliert zu Big Data

Anwendungen

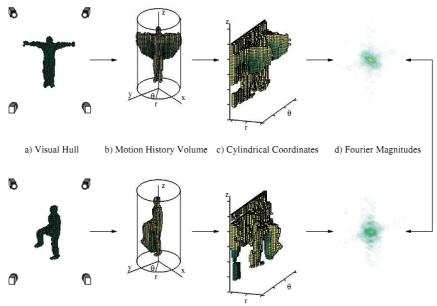
Game Mining



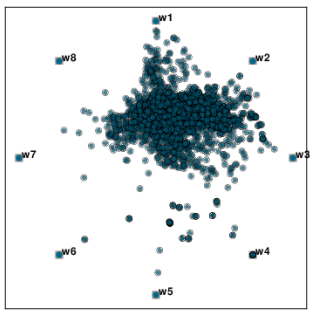
kollektive Aufmerk-samkeitsprozesse



Aktionserkennung



Twitter Forensik



Stressdetektion bei Pflanzen

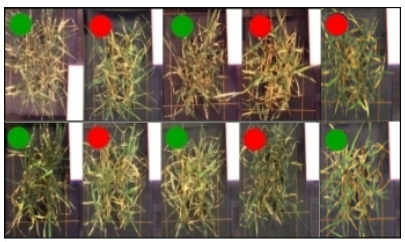
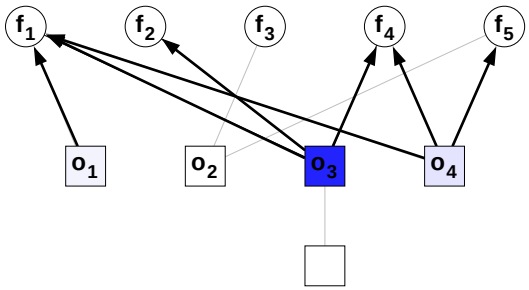


Image Retrieval



Anwendungen: Mining World of Warcraft

Warum?

- Automatische Charakterisierung von MMORPG Teams / Gilden
- Verstehen von Gruppenverhalten in MMORPGs
- Vereinfachung des Game Design Prozesses

Anwendungen: Mining World of Warcraft

data:

- 150 Million in-game Beobachtungen von
  - 18 Million Spiel-Charakteren
  - aus 1.4 Million Gilden
- Wir haben uns die Erfahrungslevel (10-80) der Spieler angeschaut
- Für jede Gilde haben wir also das 70 dim. Erfahrungshistogram angeschaut

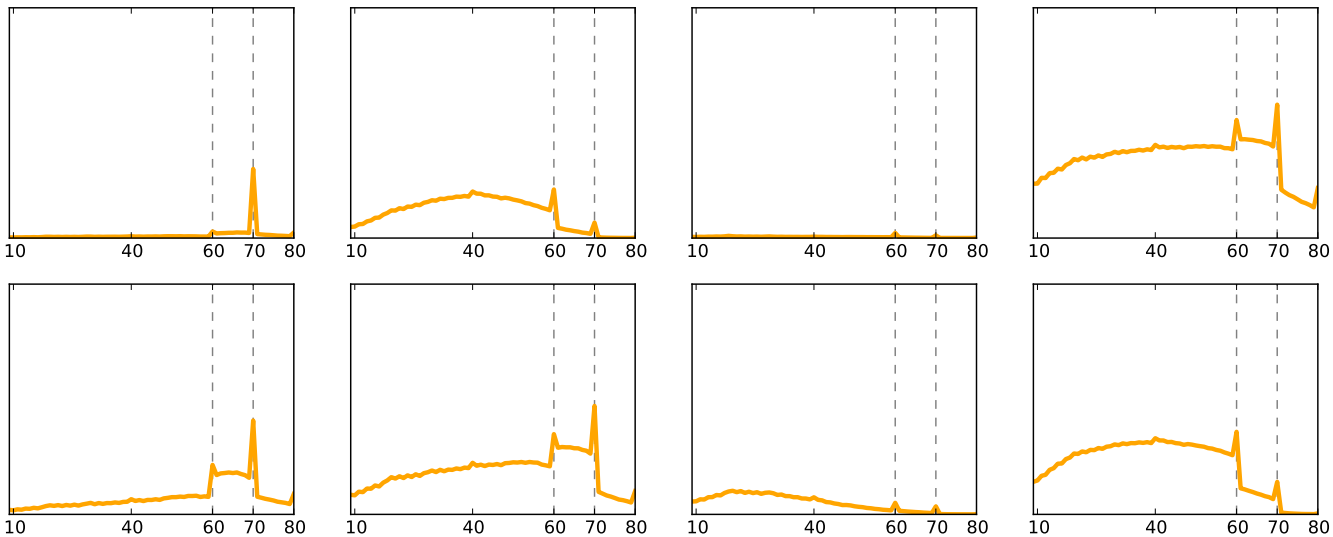
$\Rightarrow$ Datenmatrix mit 98 Million Einträgen

details in:

- C. Thurau and C. Bauckhage: Analyzing the evolution of social groups in World of Warcraft. IEEE CIG, 2010.

Anwendungen: Mining World of Warcraft

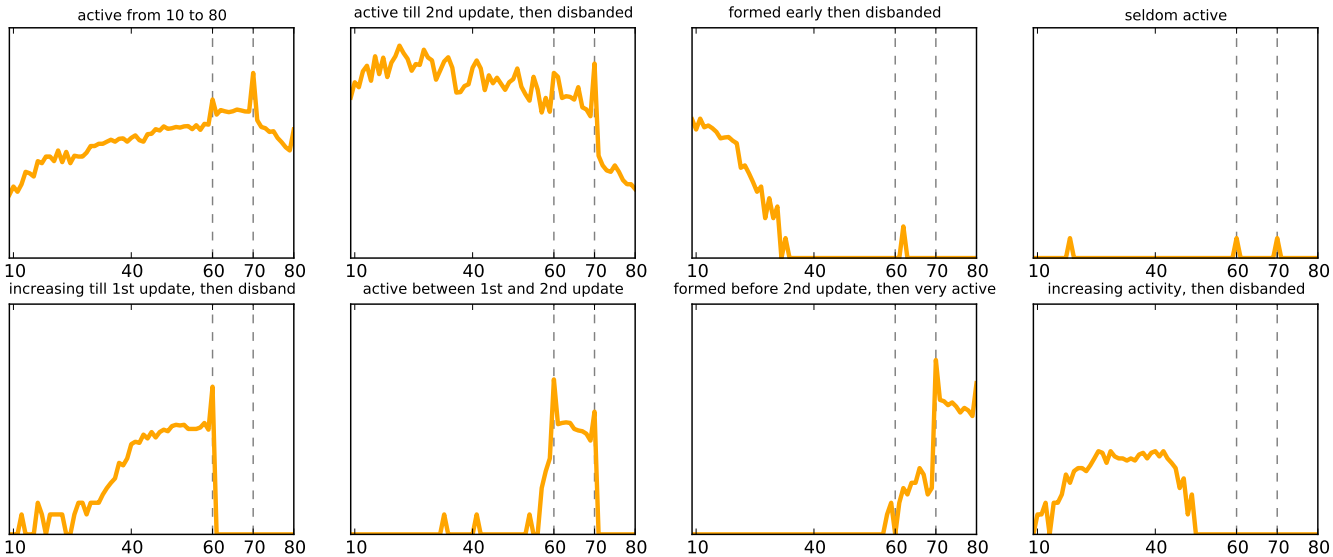
Zentroide mittels k-means:



- sehen sehr ähnlich aus
- schwierig zu verstehen

Anwendungen: Mining World of Warcraft

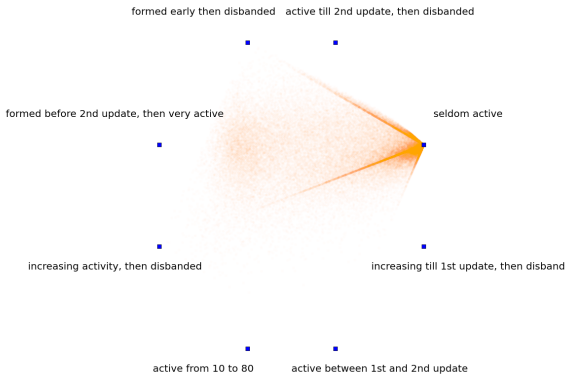
"Zentroide" mittels SiVM:



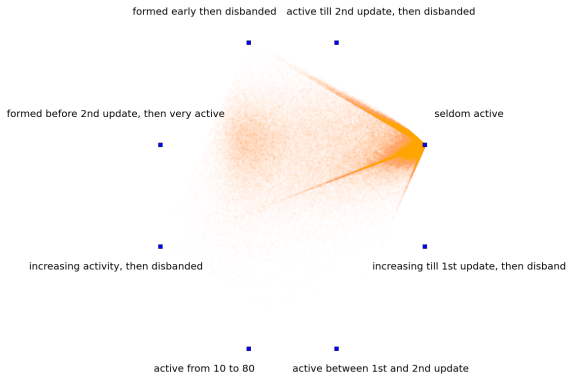
- sehr viel einfacher zu verstehen, weil es *extreme* Gilden sind.

Anwendungen: Mining World of Warcraft

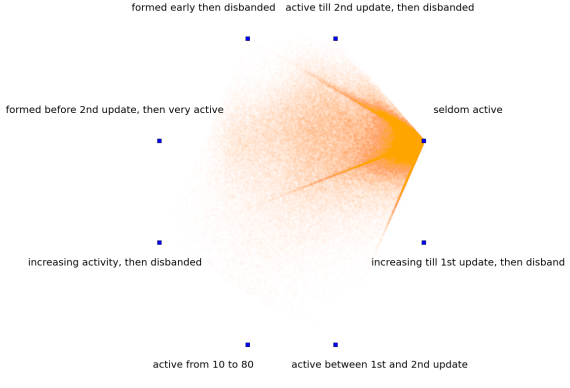
über die Zeit:



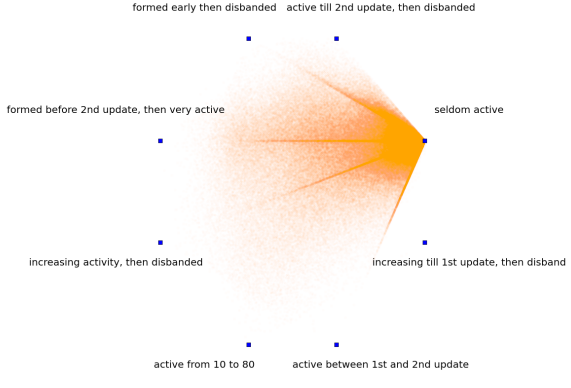
1 year



2 years

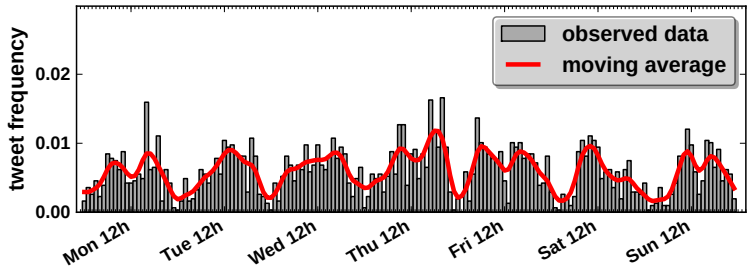


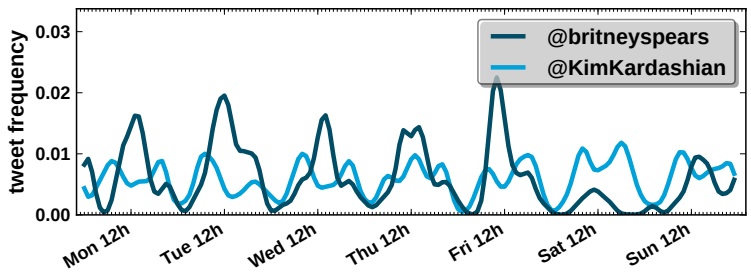
3 years



4 years

Anwendungen: Wer schreibt die Tweets?





- do celebrities really tweet themselves?

Daten / Merkmale:

- 300 celebrity twitteres
- 1.5 Millionen Tweets
- 30 dim. Merkmalvektoren, die stilistische Eigenarten bescreiben

$\Rightarrow$ Datenmatrix mit 45 Millionen Einträgen

Details in:

- C. Thurau, K. Kersting, and C. Bauckhage: Yes we can: simplex volume maximization for descriptive web-scale matrix factorization. ACM CIKM, 2010.

Anwendungen: Wer schreibt die Tweets?

@aplusk

z_1	I often ponder as to how similar the name Adam and the word "atom" are to one another	0.871
z_2	@thesettlers Its R	0.051
z_3	new blog... http://tinyurl.com/cggp43	0.024
z_4	http://www.ustream.tv/ashton	0.020
z_5	@toojiggy what?	0.011
z_6	@DillonPrat amazing!	0.009
z_7	HOlY GOOD GOATS MILK WE'VE GOT A QB!!!!!!!!!! THANK YOU BABY JESUS BUDAH ALLAH KABBALAH DARWIN WHOEVER!!!!!	0.007
z_8	Victory is ours!!!!!!!!	0.007

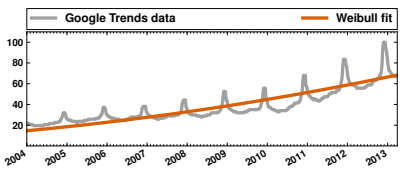
@britneyspears

z_1	Brit's up for a People's Choice Award for Fave Scene Stealing Guest Star from How I Met Your Mother. Please vote: http://tinyurl.com/5ns9n3	0.317
z_2	Over a million hits in 24 hours on the new http://www.britneyspears.com. You guys really are the best fans in the world!	0.301
z_3	New leg, new song, new outfits, new do.... -Brit	0.301
z_4	Have you joined the Britney social network yet? Connect with other Britney fans at http://www.circusvip.com	0.032
z_5	NYC....Here I come!!! -Brit	0.022
z_6	HAPPY BIRTHDAY BRETT!!!!! XOXO -Britney	0.011
z_7	Happy Holidays from Britney!! http://tinyurl.com/7uyrr4	0.011
z_8	Thanks @britneylive @filmmaker311 @singlemomindebt @youbetheanchor @perezhilton! Were you there? Share @ http://britneyspears.com/tour	0.005

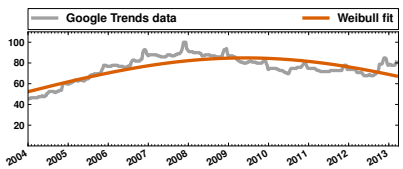
Anwendungen: Kollektives Verhalten im Web

idea / motivation: Wu and Huberman, PNAS, 104(45), 2007:

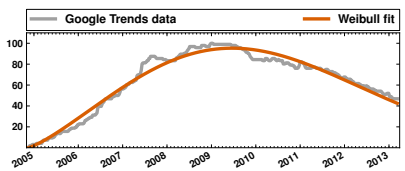
- "understanding the dynamics of collective attention is a key scientific challenge of the information age"



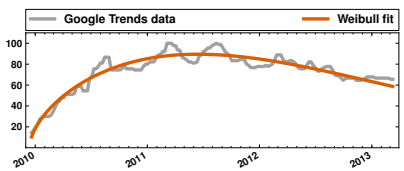
amazon



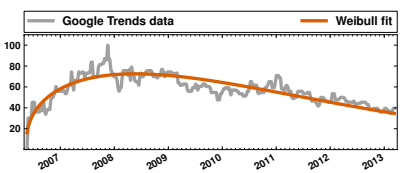
ebay



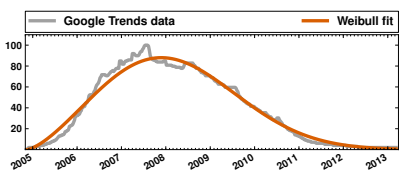
flickr



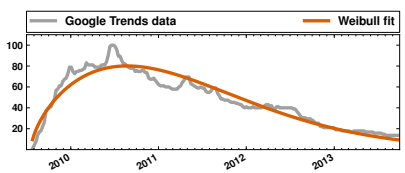
foursquare



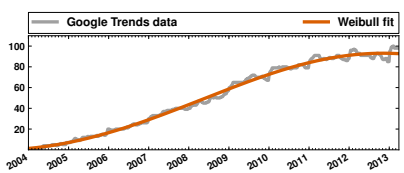
librarything



myspace



renren



wordpress

Anwendungen: Kollektives Verhalten im Web

Daten:

- Google Trends Daten für
  - 175 "social media"-Services
  - 45 Länder + weltweit

Ansatz:

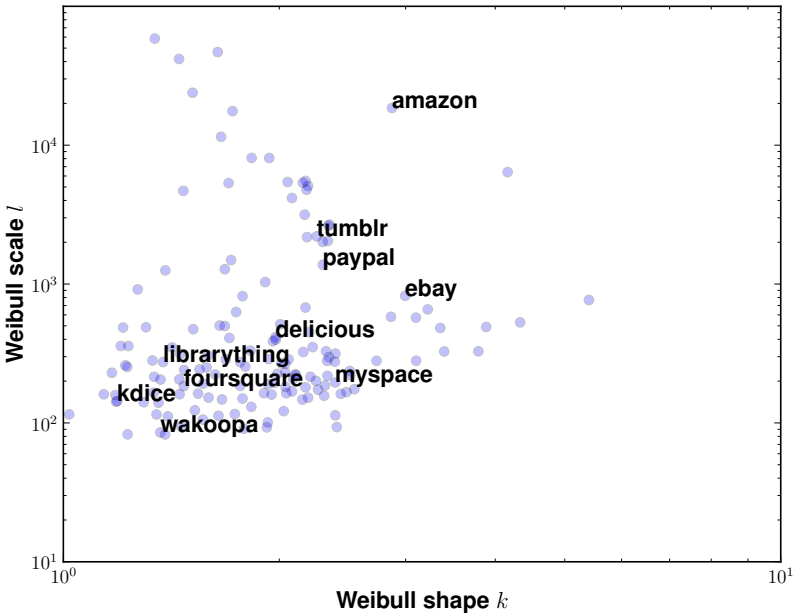
- Bestimme Weibull-Verteilungen der Suchhäufigkeiten über die Zeit hinweg
- Einbetten der "social media"-Services " in den 2-D Weibull Parameterraum
- Anwenden einer *Kern-AA*

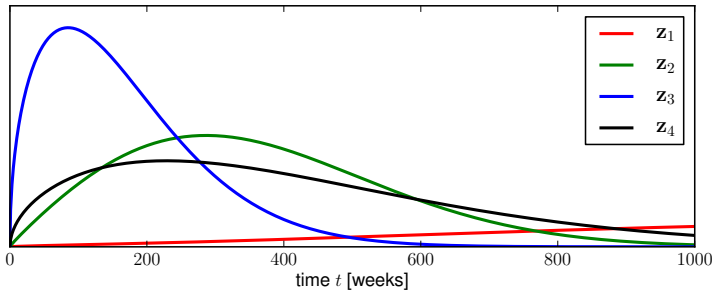
Details in:

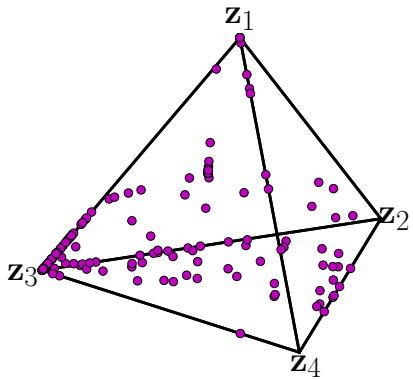
- C. Bauckhage, K. Kersting, and B. Rastagarpana: Collective attention to social media evolves according to diffusion models. ACM WWW, 2014.

Anwendungen: Kollektives Verhalten im Web

Ergebnisse:





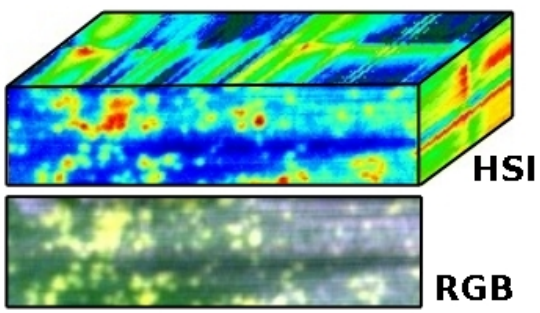


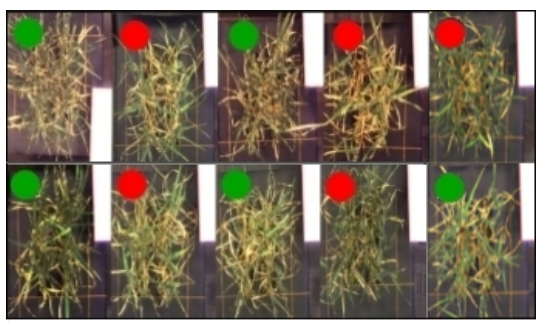
at	corresponding service(s)	represents
$\boldsymbol{z}_1$	*airbnb*, *foursquare*	22% of the data
$\boldsymbol{z}_2$	*delicious*, *foursquare*	11% of the data
$\boldsymbol{z}_3$	*foursquare*	58% of the data
$\boldsymbol{z}_4$	*spotify*, *foursquare*	09% of the data

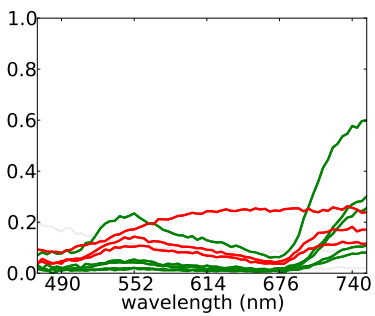
Anwendungen: Trockenstress bei Pflanzen

Motivation:

- Trocken- und anderer abiotischer Stress ist Schuld an 70% des Ertragverlusts
- Um dem gegenzuwirken, müssen wir die Stressreaktionsmechanismen von Pflanzen verstehen
- Hyperspektrale Bildaufnahmen erlauben das am lebenden Objekt







Anwendungen: Trockenstress bei Pflanzen

Daten / Ansatz:

- 205 tägliche Messungen von gestressten und gesunden Pflanzen
- Hypersektrale Kammera Surface Optics Corp. SOC-700
- Jedes Bild ist ein $640 \times 640 \times 69$ Datenwürfel

$\Rightarrow$ Datenmatrix mit 5.8 Milliarden Einträgen

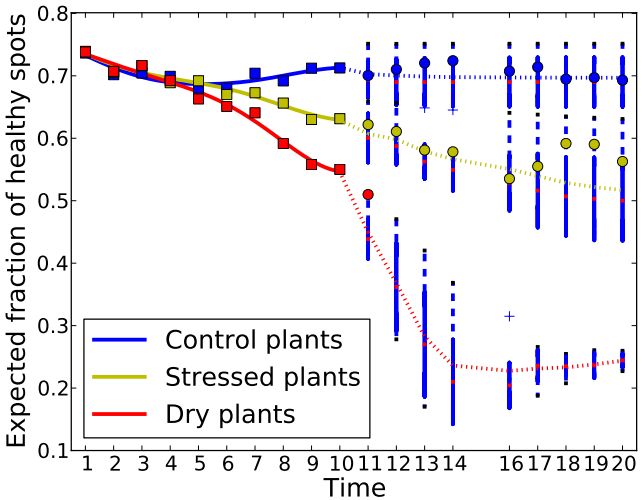
- Wir haben erst die extremen Verteilungen bestimmt
- Dann Dirichlet-Verteilungen auf den Simplices per Tag bestimmt
- Und dann eine Bayes'sche Regression über die Zeit durchgeführt

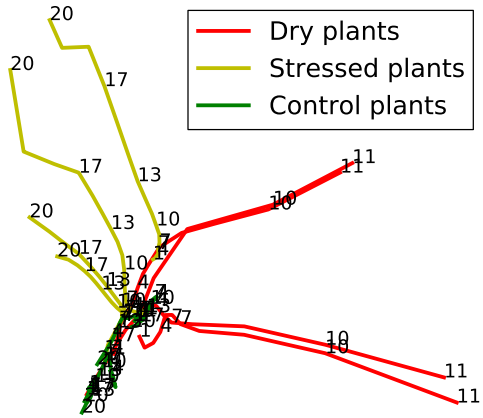
Details in:

- C. Römer, M. Wahabzada, A. Ballvora, F. Pinto, M. Rossini, C. Panigada, J. Behmann, J. Leon, C. Thurau, C. Bauckhage, K. Kersting, U. Rascher, and L. Plümer: Early drought stress detection in cereals: simplex volume maximisation for hyperspectral image analysis. Functional Plant Biology 39(11), 2012.
- C. Bauckhage, K. Kersting: Data Mining and Pattern Recognition in Agriculture. KI 27(4), 2013.

Anwendungen: Trockenstress bei Pflanzen

Ergebnisse:





Was haben wir gelernt?

- Matrixfaktorisierung (MF) ist eine zentrale Aufgabe im Data Mining
- MF reduziert/kompremiert die Daten
- kMeans und viele andere DM-Algorithmen können als MF formuliert werden
- Dazu ändert man die Zielfunktion und die Nebenbedingungen
- Je nach den benutzten Nebenbedingungen sind die Ergebnisse einfacher zu verstehen.