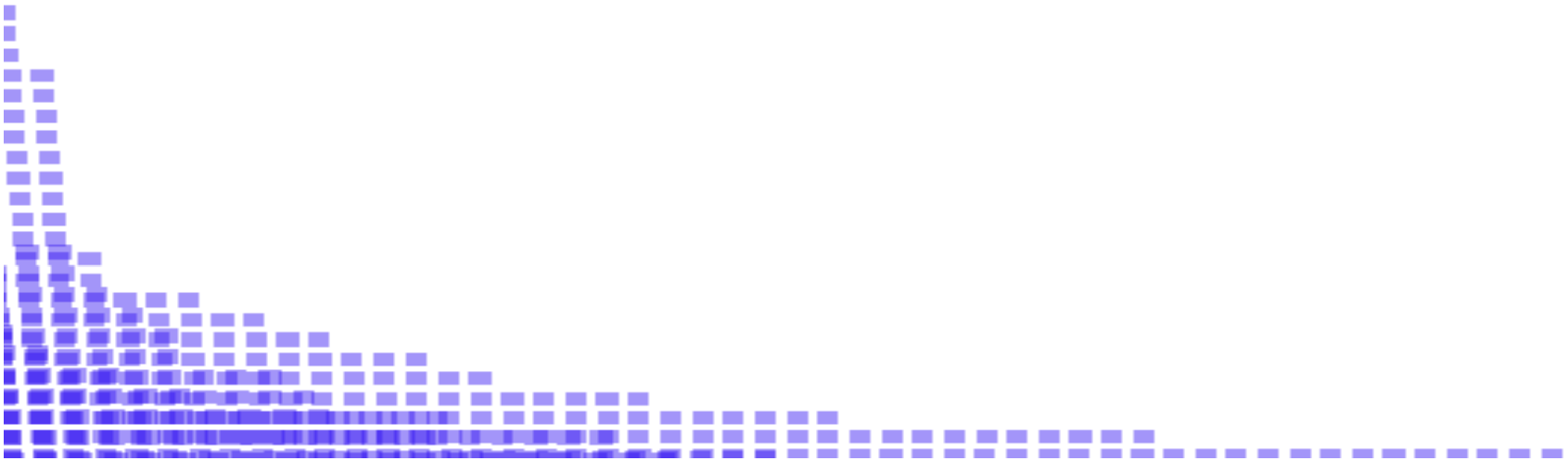
Informationsextraktion mit LastFM im Vergleich zu Google

Spezielle Kapitel der Informatik: Music Information Retrieval | KV SS 2009

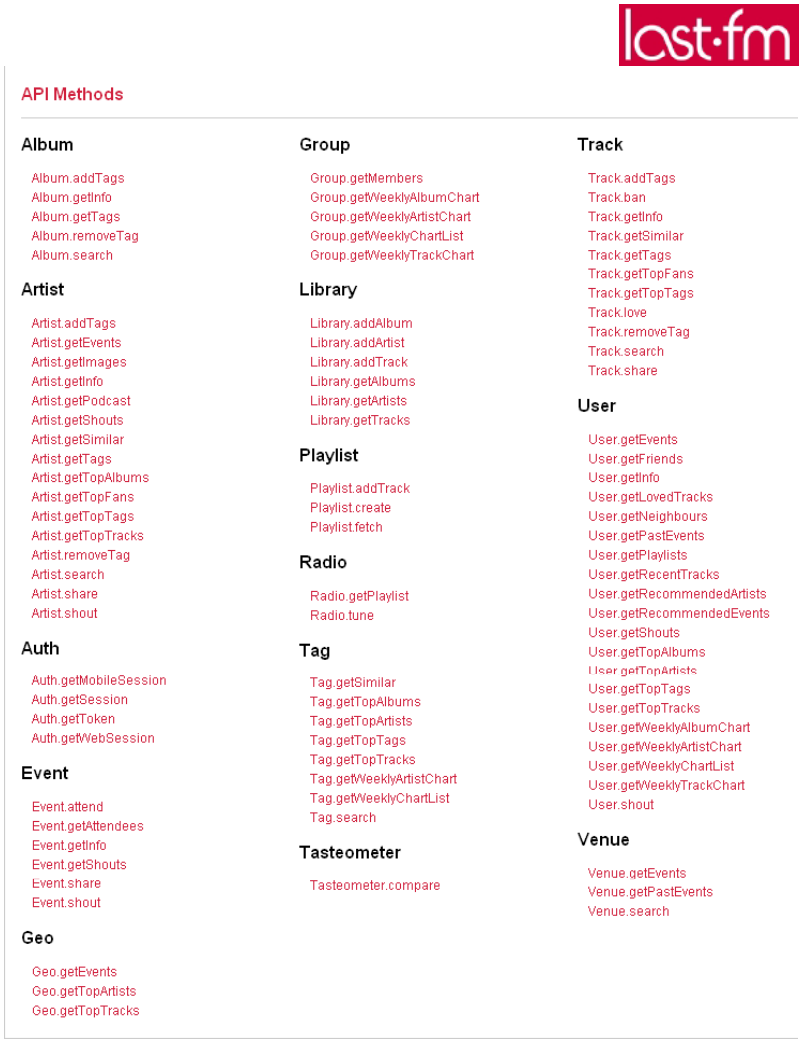
Jakob Doppler, Matthias Husinsky, Doris Zachhuber



Einleitung

Aufgabenstellung

- Context-basierte Feature Extraktion (Google, LastFM)
- Music Community Portal LastFM,
  - Informationskategorien
  - http://www.lastfm.de/api
- Ähnlichkeitsmaße berechnen
- Optional
  - Visualisierung
  - Klassifikation



API Methods

Album
Album.addTags
Album.getInfo
Album.getTags
Album.removeTag
Album.search

Artist
Artist.addTags
Artist.getEvents
Artist.getImages
Artist.getInfo
Artist.getPodcast
Artist.getShouts
Artist.getSimilar
Artist.getTags
Artist.getTopAlbums
Artist.getTopFans
Artist.getTopTags
Artist.getTopTracks
Artist.removeTag
Artist.search
Artist.share
Artist.shout

Auth
Auth.getMobileSession
Auth.getSession
Auth.getToken
Auth.getWebSession

Event
Event.attend
Event.getAttendees
Event.getInfo
Event.getShouts
Event.share
Event.shout

Geo
Geo.getEvents
Geo.getTopArtists
Geo.getTopTracks

Group
Group.getMembers
Group.getWeeklyAlbumChart
Group.getWeeklyArtistChart
Group.getWeeklyChartList
Group.getWeeklyTrackChart

Library
Library.addAlbum
Library.addArtist
Library.addTrack
Library.getAlbums
Library.getArtists
Library.getTracks

Playlist
Playlist.addTrack
Playlist.create
Playlist.fetch

Radio
Radio.getPlaylist
Radio.tune

Tag
Tag.getSimilar
Tag.getTopAlbums
Tag.getTopArtists
Tag.getTopTags
Tag.getTopTracks
Tag.getWeeklyArtistChart
Tag.getWeeklyChartList
Tag.search

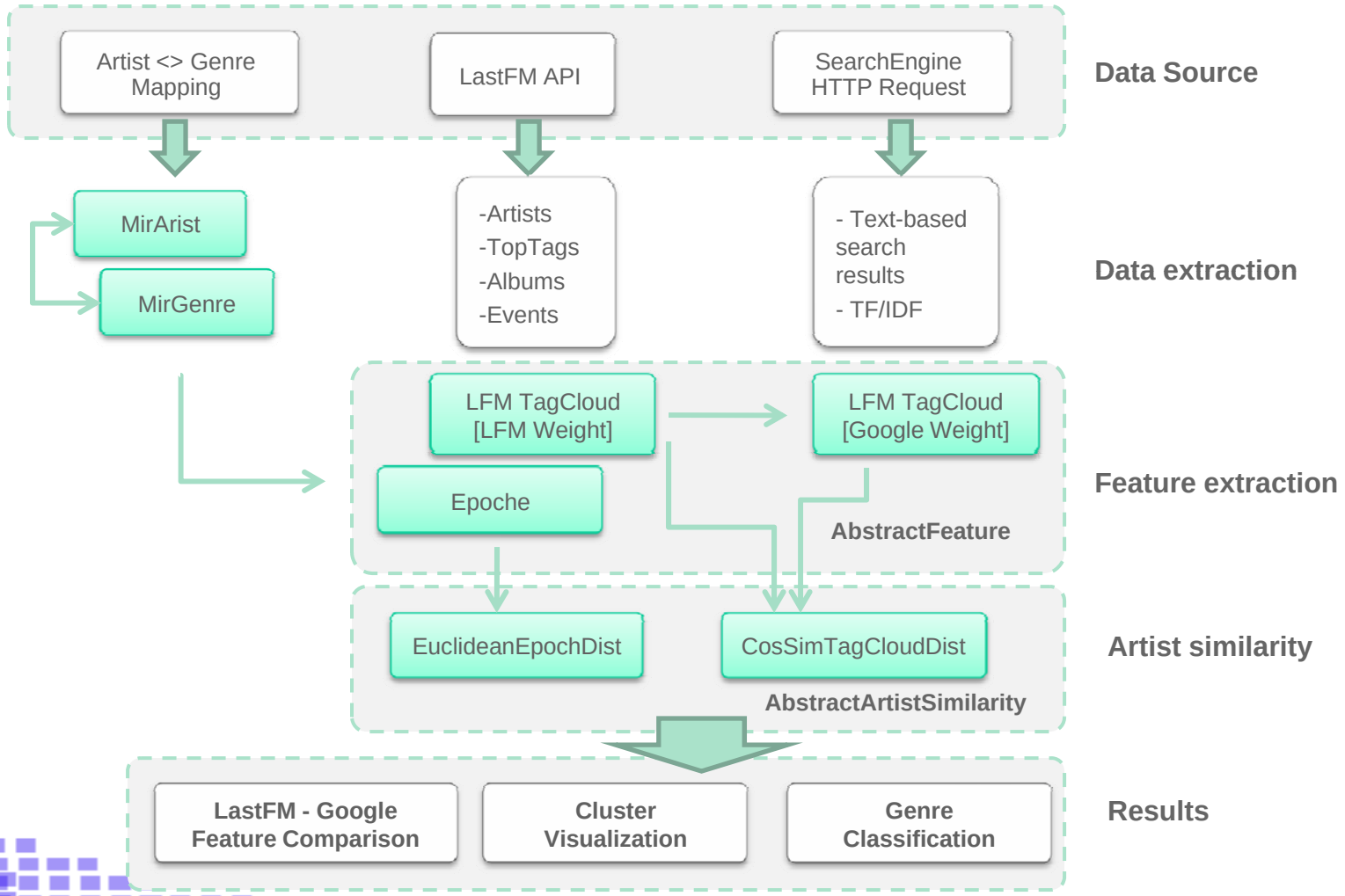
Tasteometer
Tasteometer.compare

Track
Track.addTags
Track.ban
Track.getInfo
Track.getSimilar
Track.getTags
Track.getTopFans
Track.getTopTags
Track.love
Track.removeTag
Track.search
Track.share

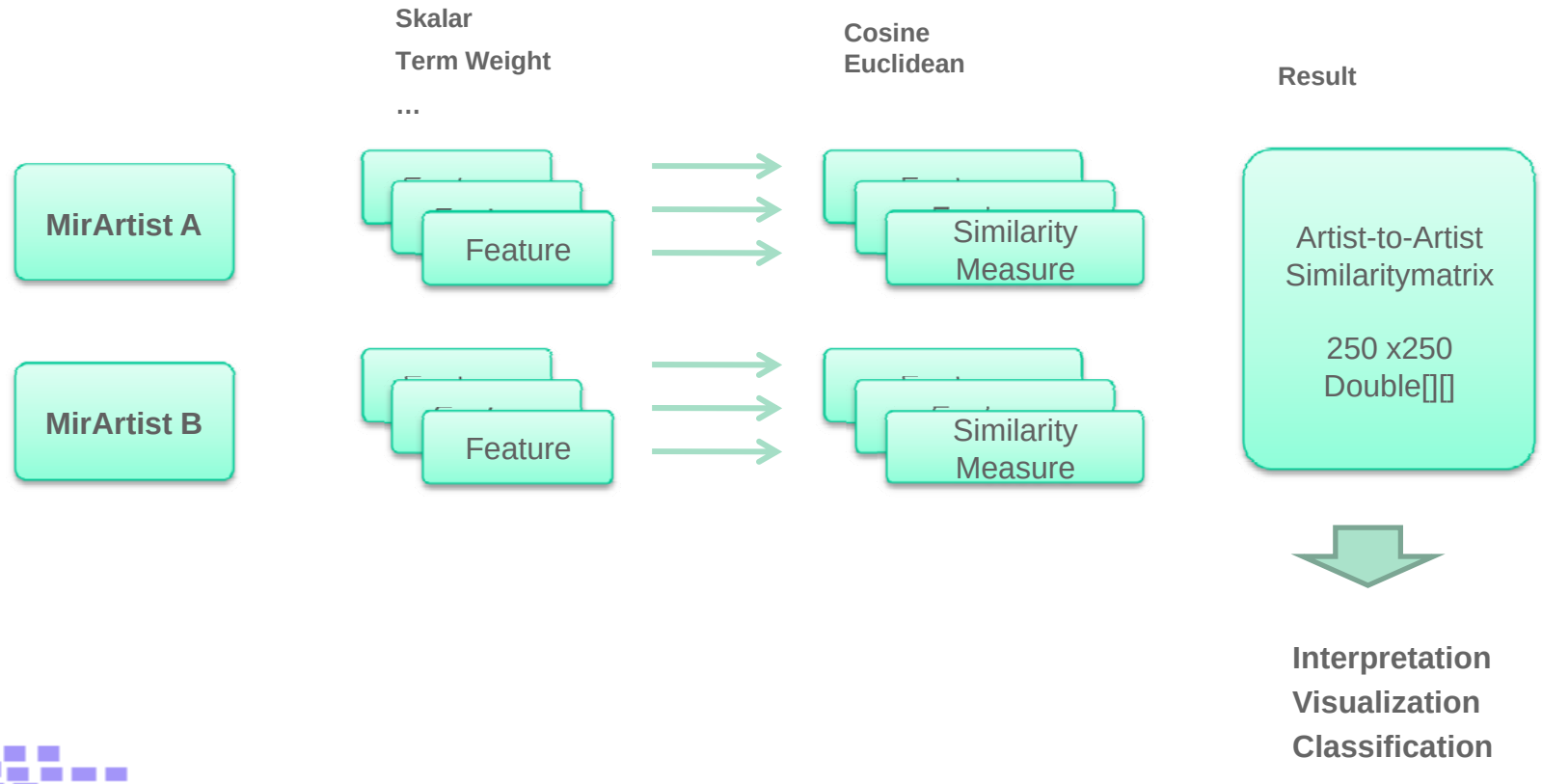
User
User.getEvents
User.getFriends
User.getInfo
User.getLovedTracks
User.getNeighbours
User.getPastEvents
User.getPlaylists
User.getRecentTracks
User.getRecommendedArtists
User.getRecommendedEvents
User.getShouts
User.getTopAlbums
User.getTopArtists
User.getTopTags
User.getTopTracks
User.getWeeklyAlbumChart
User.getWeeklyArtistChart
User.getWeeklyChartList
User.getWeeklyTrackChart
User.shout

Venue
Venue.getEvents
Venue.getPastEvents
Venue.search

Systemarchitektur



Systemarchitektur



Systemarchitektur

Implementierung

- Umfangreiche Implementierung in Java
- **MIR Entitäten** als Objekte mit abstrakten Features und Similarities
- 40 Klassen, ~4000 LoC, unendlich viel Geduld ;-)
  SVN - Google Code Repository

Zahlreiche Libraries

- LastFM Java API
- Matrix Utils
- Text Utils Apache Commons Lang
- CoMirva (Anysearch, UrlRetriever)
- **Visualisierung** JUNG (Java Universal Network Graph)
- **Klassifikation** Machine Learning Toolkit Weka

Informationsgewinnung

Datenquellen

- **Artist zu Genre Mapping**
  - Ausgangsmaterial für Feature Extraction
  - ~ 250 Artists, 14 Genres
  - Genre-Labeling -Groundtruth für Klassifizierung
- **LastFM**
  - Benutzeraccount zum Generieren eines API Keys
  - Informationskategorien Artist ,Top Tags, Top Artists in Tags, Top Albums in Tags,
  - Features: Artist Tag Cloud, Wirkzeit
  - Ähnlichkeitsmaß: Similar Artist Ranking

- **Search Engine Google**
  - Informationskategorien: Term-based filtering and weighting
  - Feature: LastFM Artist Tag Cloud Neu gewichtet
  - Ähnlichkeitsmaß TF basierend auf

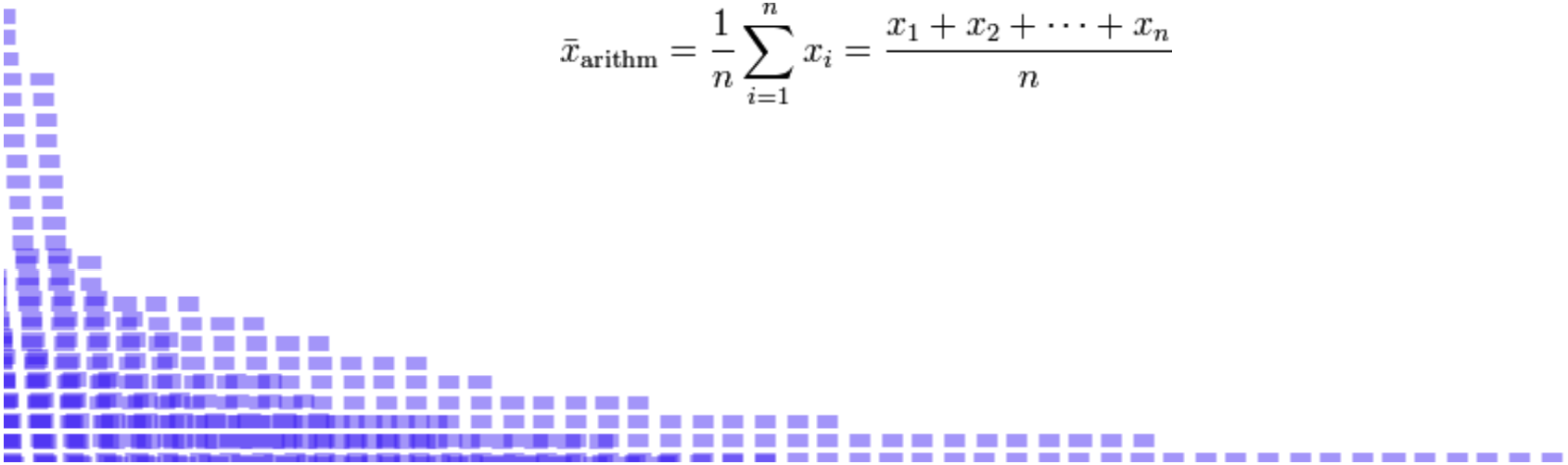
Informationsgewinnung : Albumbasierte Wirkungszeit

Feature Extraktion und Ähnlichkeitsmaß (I) – Albumbasierte Wirkungszeit

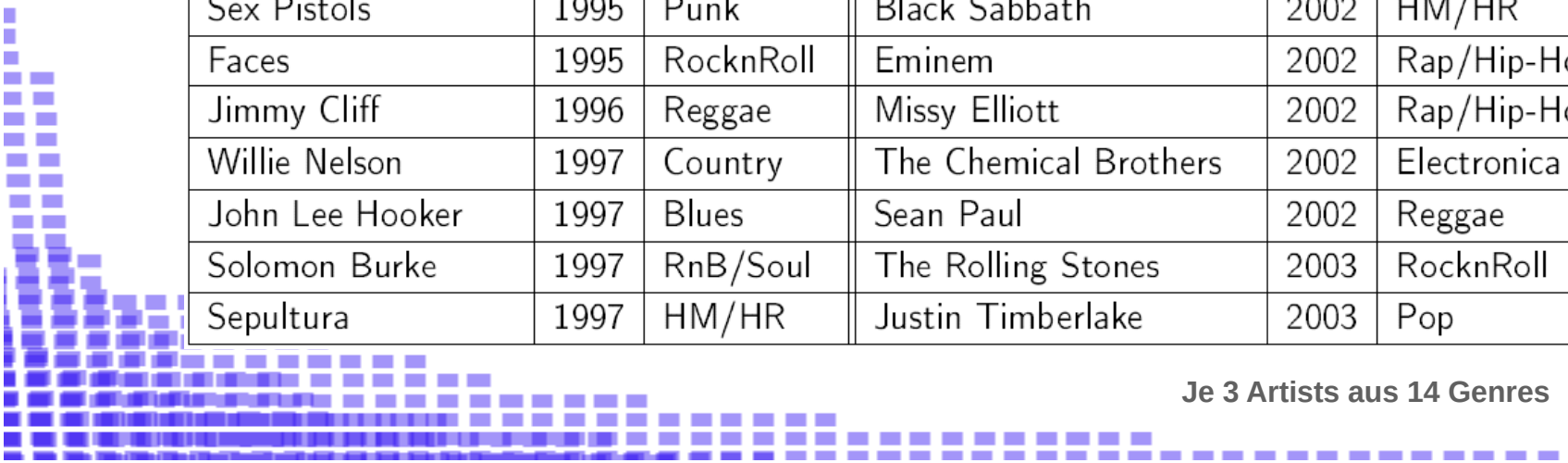
LastFM Alben-Releasedates

- Extraktion der Alben eines Artists
- Extraktion des Releasedates eines Albums → Jahr
- Arithmetisches Mittel aller Releasedates
- Absolute Distanz zweier skalarer Werte → Similarity-Matrix
- Normierung und Invertierung der Ähnlichkeitswerte

$$\bar{x}_{\mathrm{arithm}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$



Künstler	WZ	Genre	Künstler	WZ	Genre
Miles Davis	1982	Jazz	Ramones	1997	Punk
Dave Brubeck	1987	Jazz	Bad Religion	1997	Punk
Leonard Cohen	1989	Folk	PublicEnemy	1997	Rap/Hip-Hop
Kraftwerk	1990	Electronica	Bob Marley	1997	Reggae
Taj Mahal	1993	Blues	Madonna	1997	Pop
Aretha Franklin	1993	RnB/Soul	Sam Cooke	1998	RnB/Soul
Wolfgang A. Mozart	1993	Classical	Nirvana	1998	Alt.Rock/Indie
Johnny Cash	1994	Country	Beck	1998	Alt.Rock/Indie
John Mayall	1994	Blues	The Smashing Pumpkins	1998	Alt.Rock/Indie
Iron Maiden	1994	HM/HR	ABBA	1998	Pop
The Animals	1994	RocknRoll	Bob Dylan	1999	Folk
Johannes Brahms	1994	Classical	Johann Sebastian Bach	2000	Classical
Hank Williams	1995	Country	Fatboy Slim	2001	Electronica
Joan Baez	1995	Folk	Billie Holiday	2002	Jazz
Sex Pistols	1995	Punk	Black Sabbath	2002	HM/HR
Faces	1995	RocknRoll	Eminem	2002	Rap/Hip-Hop
Jimmy Cliff	1996	Reggae	Missy Elliott	2002	Rap/Hip-Hop
Willie Nelson	1997	Country	The Chemical Brothers	2002	Electronica
John Lee Hooker	1997	Blues	Sean Paul	2002	Reggae
Solomon Burke	1997	RnB/Soul	The Rolling Stones	2003	RocknRoll
Sepultura	1997	HM/HR	Justin Timberlake	2003	Pop

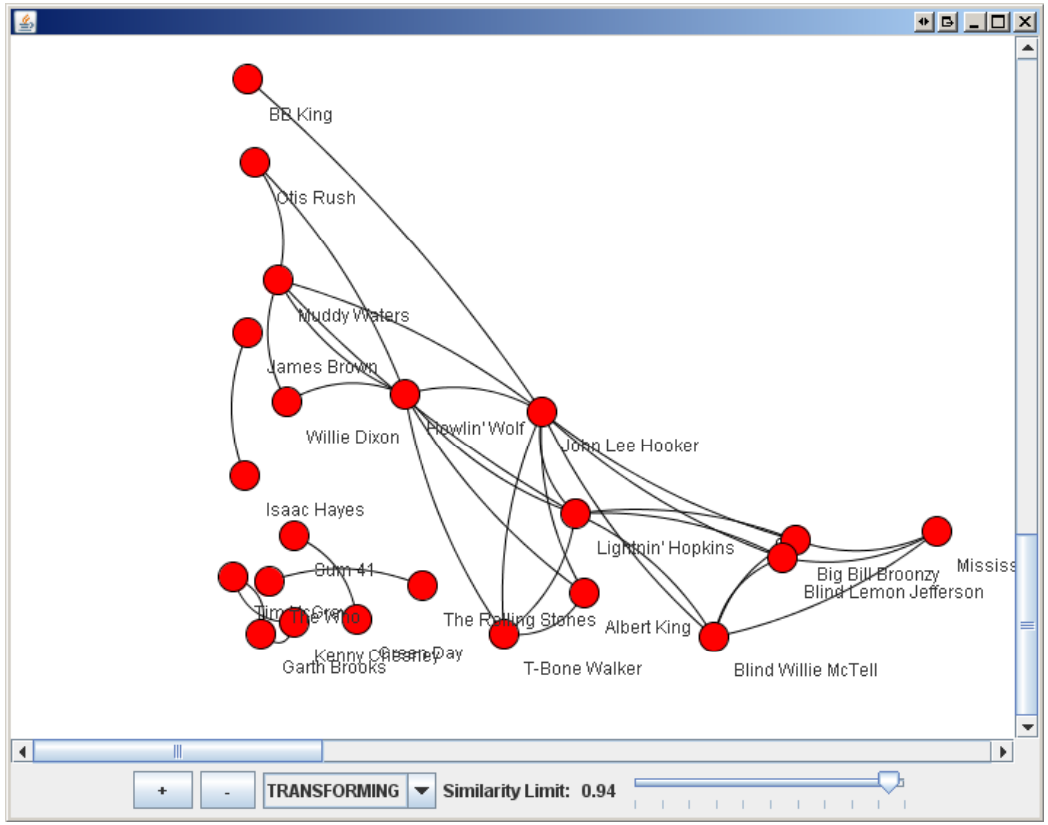


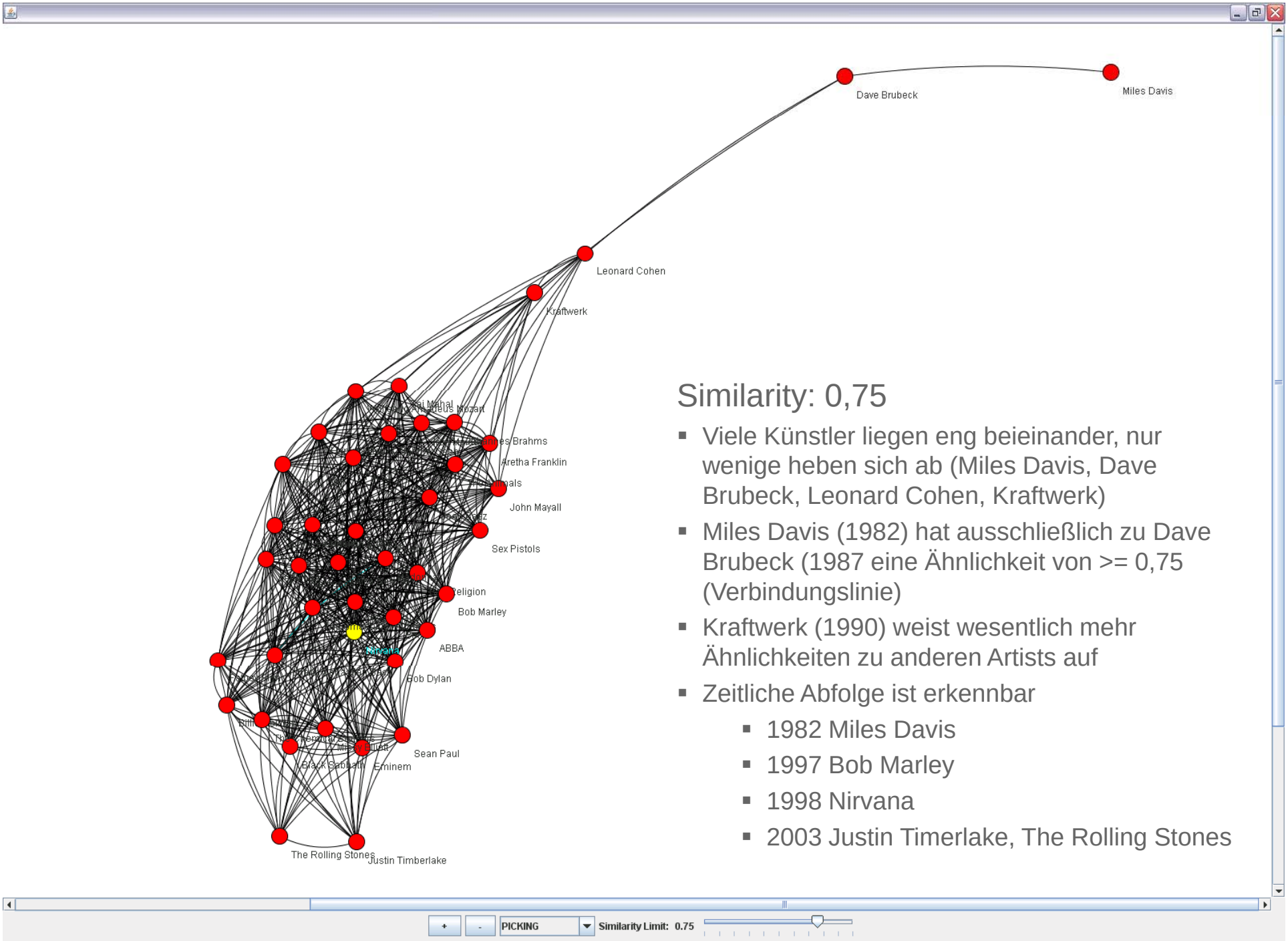
Je 3 Artists aus 14 Genres

Visualisierung

Visualisierung – Clustering

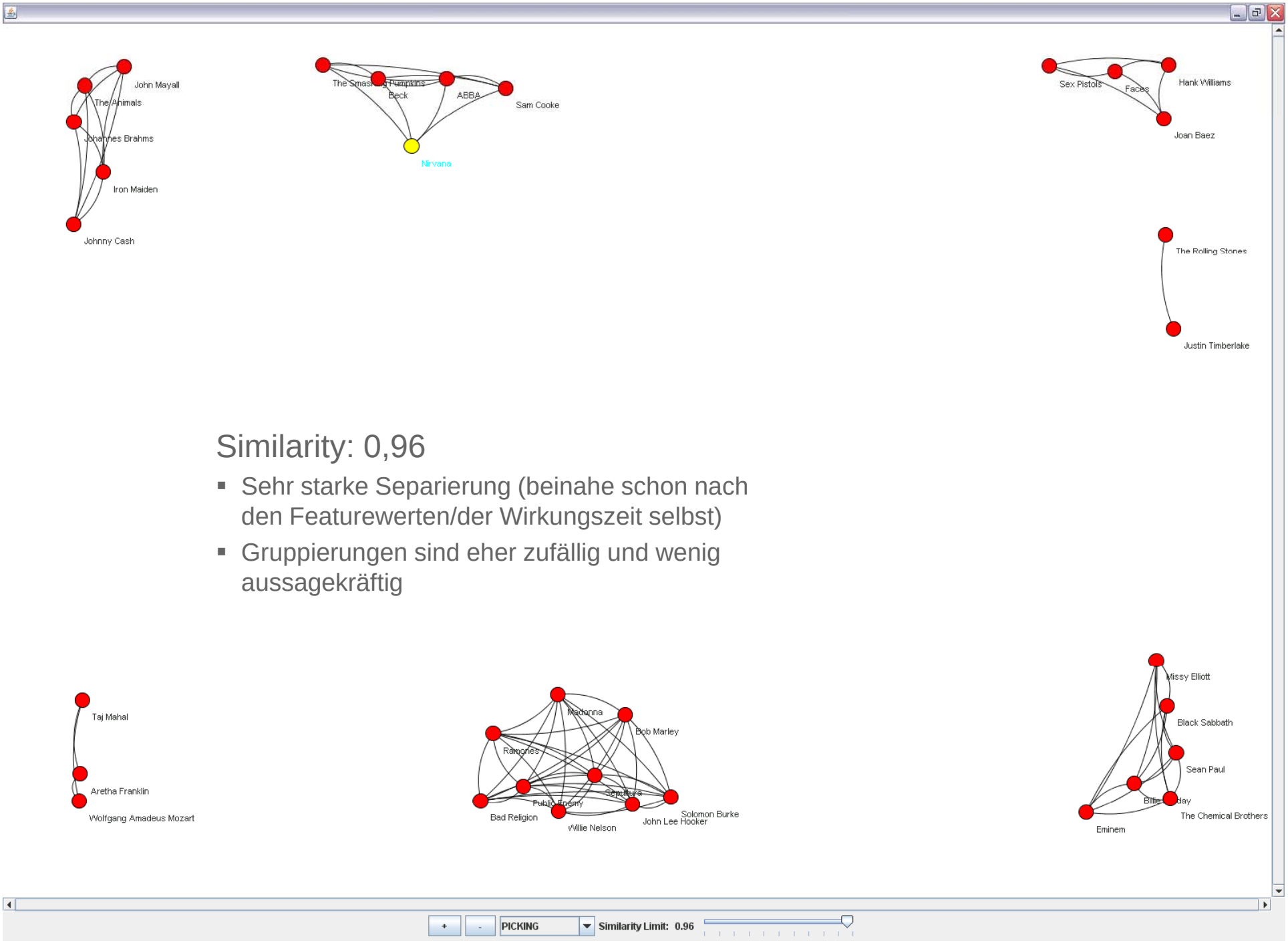
- Jung-basiertes Visualisierungstool zur Darstellung von Ähnlichkeitsclustern
- **Achtung**
  ClusterAbstand und Kantenlänge haben keine Bedeutung (Einschränkung nicht sehr elaboriert)
- Slider für die Wahl des Similarity-Thresholds [0.0-1.0]
- **Demo**





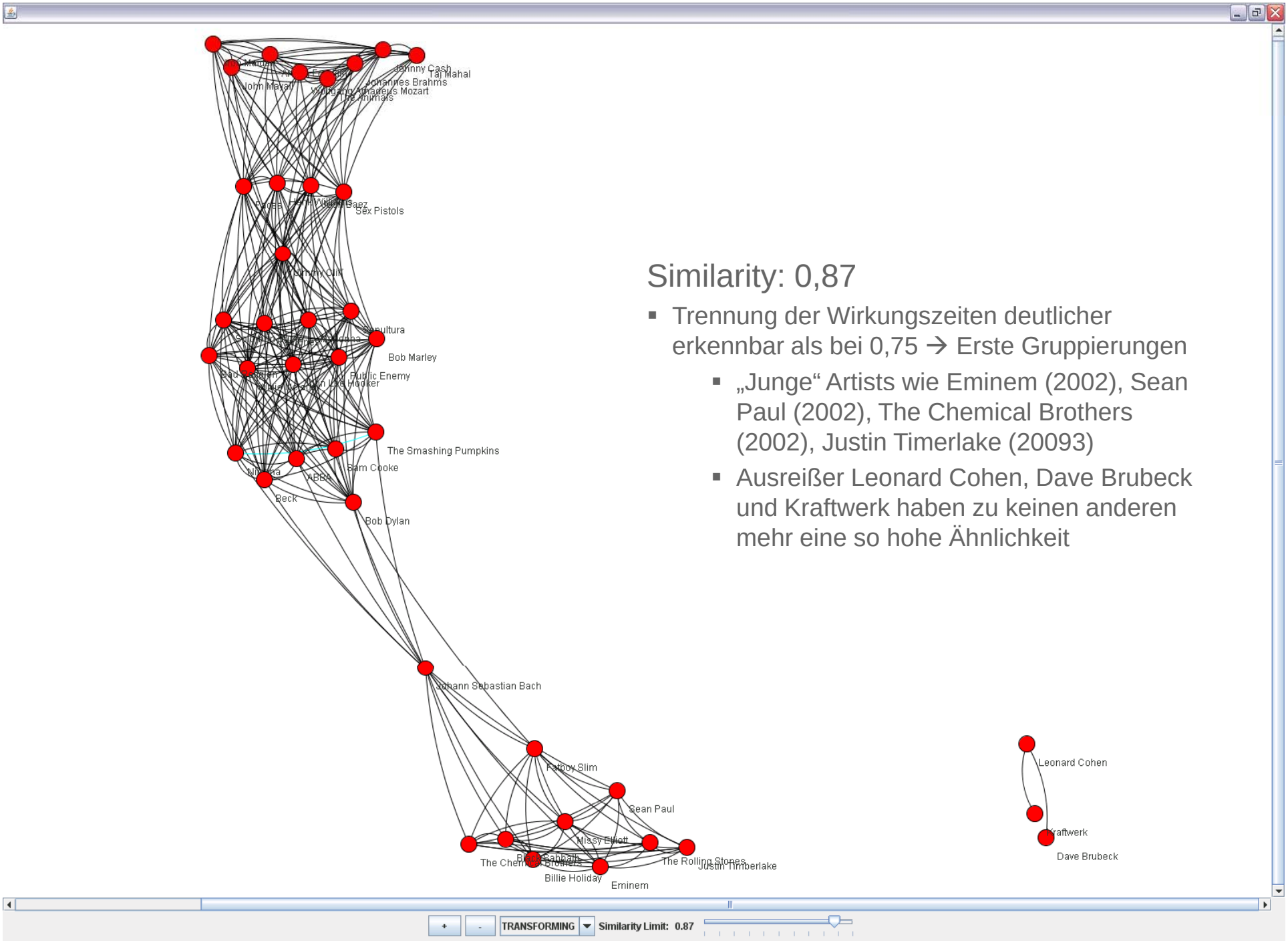
Similarity: 0,75

- Viele Künstler liegen eng beieinander, nur wenige heben sich ab (Miles Davis, Dave Brubeck, Leonard Cohen, Kraftwerk)
- Miles Davis (1982) hat ausschließlich zu Dave Brubeck (1987 eine Ähnlichkeit von >= 0,75 (Verbindungslinie)
- Kraftwerk (1990) weist wesentlich mehr Ähnlichkeiten zu anderen Artists auf
- Zeitliche Abfolge ist erkennbar
  - 1982 Miles Davis
  - 1997 Bob Marley
  - 1998 Nirvana
  - 2003 Justin Timerlake, The Rolling Stones



Similarity: 0,96

- Sehr starke Separierung (beinahe schon nach den Featurewerten/der Wirkungszeit selbst)
- Gruppierungen sind eher zufällig und wenig aussagekräftig



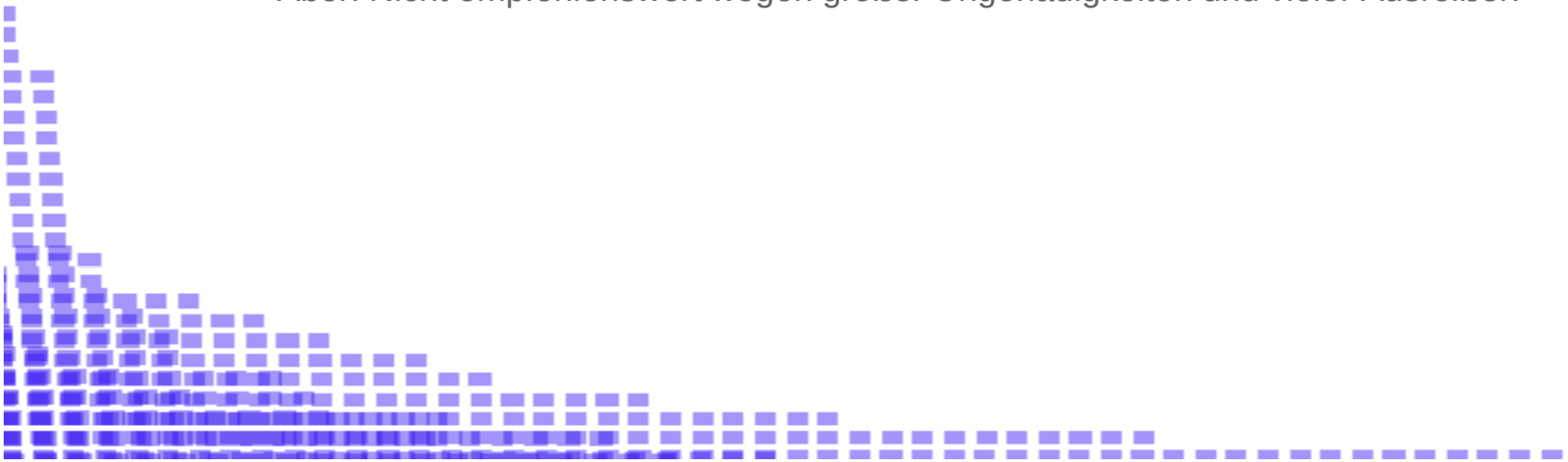
Similarity: 0,87

- Trennung der Wirkungszeiten deutlicher erkennbar als bei 0,75 → Erste Gruppierungen
  - „Junge“ Artists wie Eminem (2002), Sean Paul (2002), The Chemical Brothers (2002), Justin Timerlake (20093)
  - Ausreißer Leonard Cohen, Dave Brubeck und Kraftwerk haben zu keinen anderen mehr eine so hohe Ähnlichkeit

Ergebnisse : Albumbasierte Wirkungszeit

- Mängel
  - Bei älteren oder schon verstorbenen Künstler hat die berechnete nichts mit der tatsächlichen Wirkungszeit zu tun, v.a. im Genre Klassik: Wolfgang Amadeus Mozart (1993), J. S. Bach (2000)
  - Gründe: Alben später veröffentlicht und teilweise in lastFM nicht so gut abgebildet

- Genreähnlichkeiten
  - Bei 0,96 Ähnlichkeit keine aussagekräftigen Ergebnisse (zu kleine Zeitintervalle)
  - Bei 0,87 bessere Abbildung der Genres
    Rap/Hip-Hop: Eminem, Missy Elliott; Electronic: The Chemical Brothers, Fatboy Slim
  - Aber: Nicht empfehlenswert wegen großer Ungenauigkeiten und vieler Ausreißer!



Fazit & Optimierung : Albumbasierte Wirkungszeit

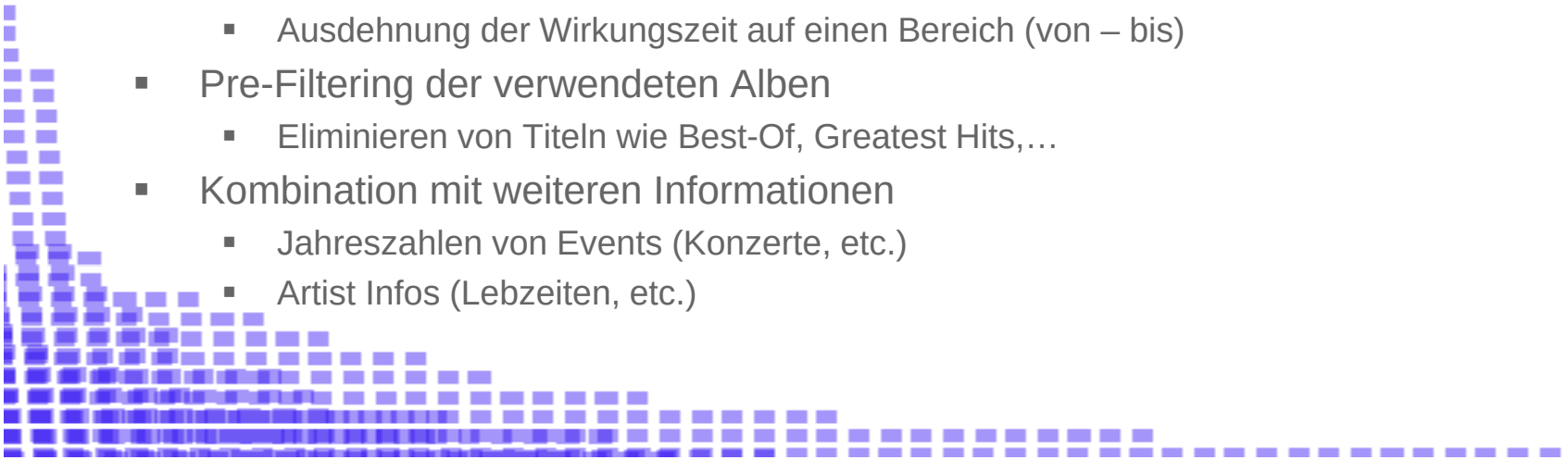
Fazit

Zum Vergleich der aktiven Wirkzeit von Artists der letzten 30 Jahre gut geeignet
→ gute Darstellung WANN die meisten Alben veröffentlicht wurden

- Keine Berücksichtigung ob One-Hit/Album-Wonder oder langjährig Veröffentlichungen

Optimierung

- Andere Berechnungsart des Features
  - Mittelwert ohne Werte außerhalb der Standardabweichung
  - Median statt arithmetischem Mittel
  - Ausdehnung der Wirkungszeit auf einen Bereich (von – bis)
- Pre-Filtering der verwendeten Alben
  - Eliminieren von Titeln wie Best-Of, Greatest Hits,…
- Kombination mit weiteren Informationen
  - Jahreszahlen von Events (Konzerte, etc.)
  - Artist Infos (Lebzeiten, etc.)



Informationsgewinnung : Tag Cloud

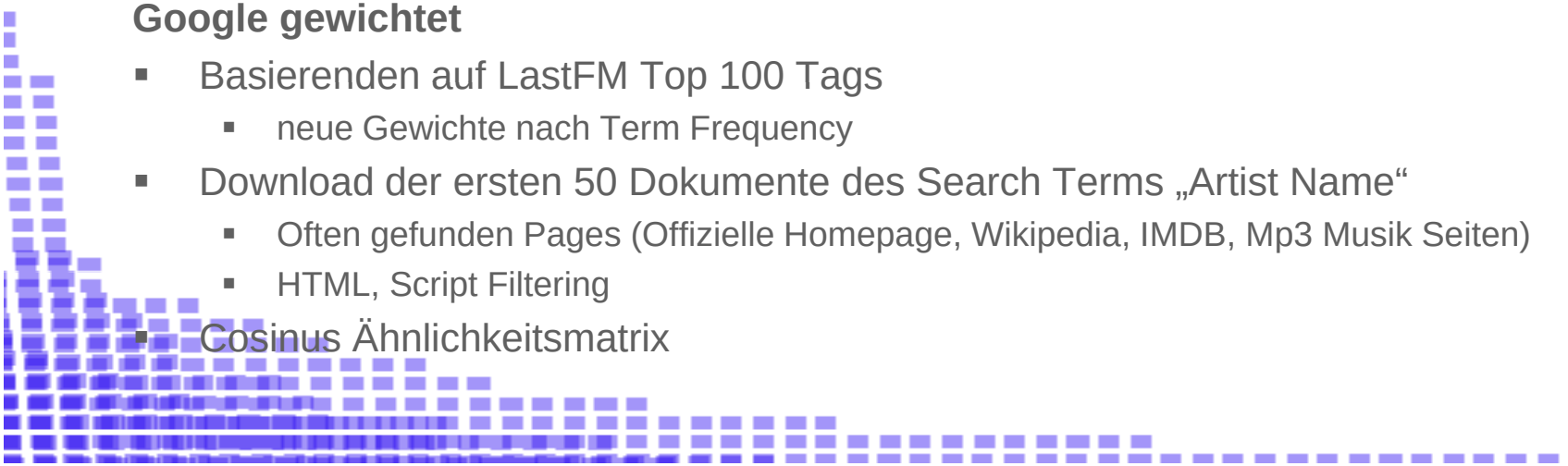
Feature Extraktion und Ähnlichkeitsmaß (II) –Tag Cloud

LastFM gewichtet

- Extraktion der Top 100 Tags für einen Artist
- Tag Filtering
  - Substrings der Artists („billy Joel“ → billy joel billy-joel)
- Normierung der Tags [100,0]
- Cosinus Ähnlichkeitsmatrix

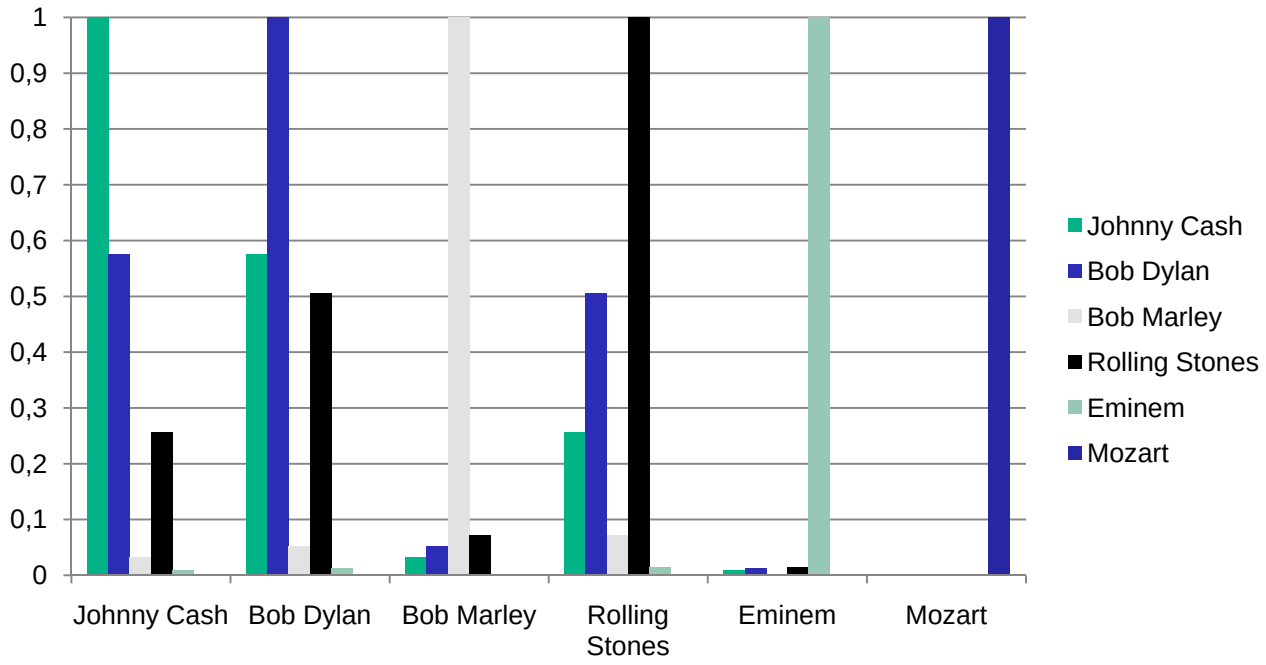
Google gewichtet

- Basierenden auf LastFM Top 100 Tags
  - neue Gewichte nach Term Frequency
- Download der ersten 50 Dokumente des Search Terms „Artist Name“
  - Often gefunden Pages (Offizielle Homepage, Wikipedia, IMDB, Mp3 Musik Seiten)
  - HTML, Script Filtering
- Cosinus Ähnlichkeitsmatrix



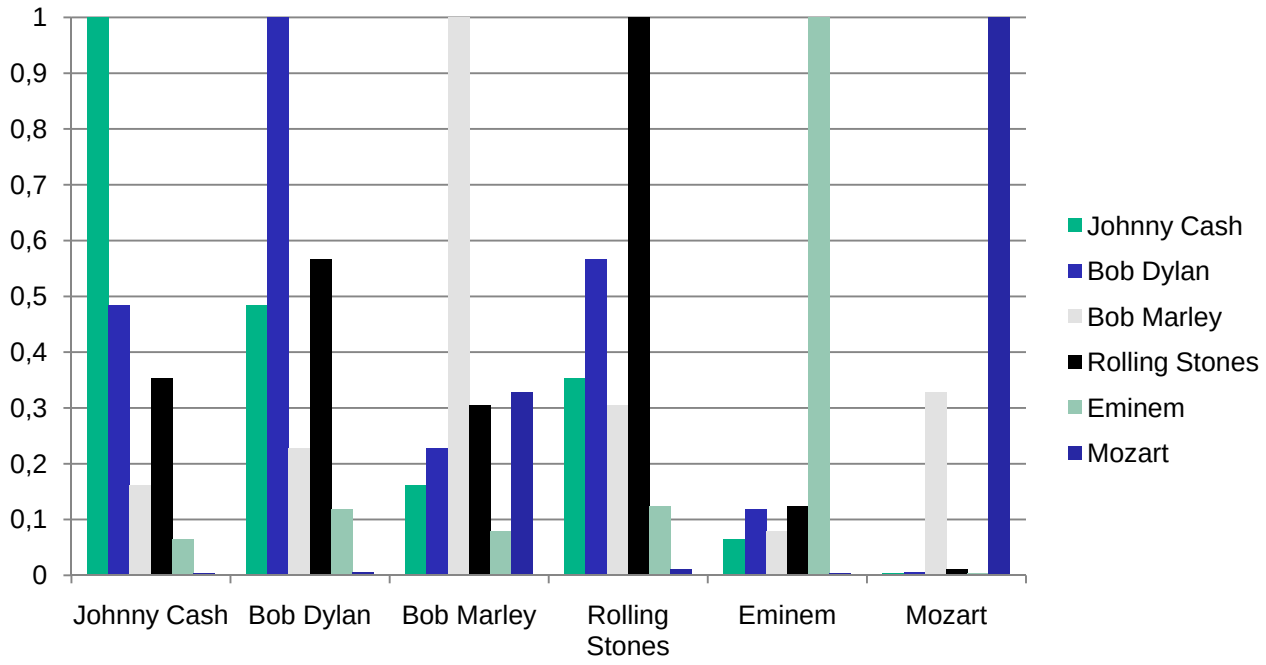
Ergebnisse : Tag Cloud

Ähnlichkeiten –Tag Cloud LastFM gewichtete Ähnlichkeiten (I)



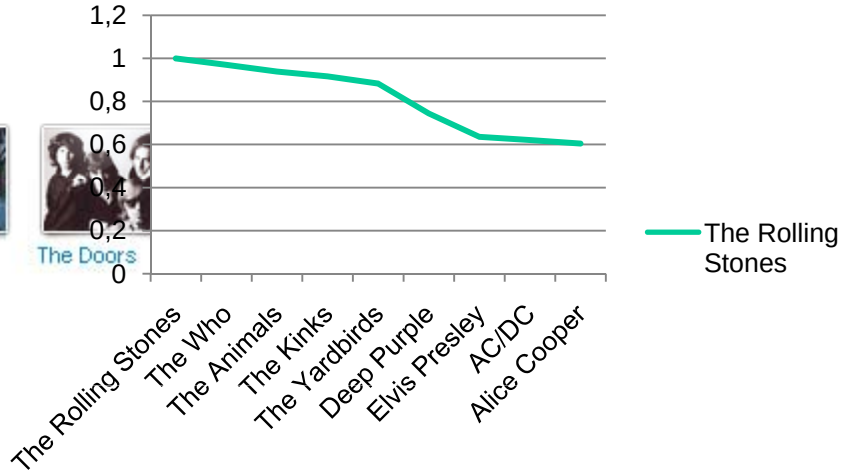
Ergebnisse : Tag Cloud

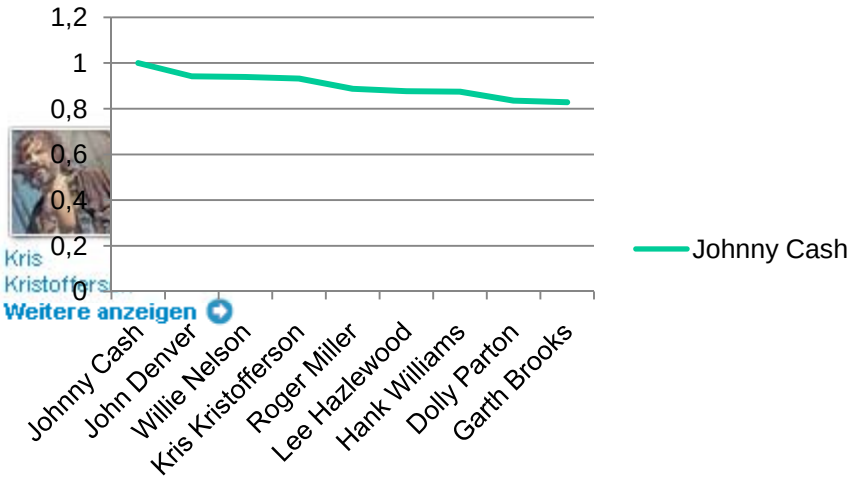
Ähnlichkeiten – Tag Cloud Google gewichtete Ähnlichkeiten (II)



Ergebnisse : Tag Cloud

Ähnlichste Künstler - TagCloudSim versus LastFM Website





Work in progress

Versuch Genre Klassifizierung

- Klassifikation der Labels des Genre<>Artist Mappings
- Gute Resultate bei Default Einstellungen , 10-fold CV **Lazy Ibk (KNN Klassifier) - 95 %** , Naive Bayes - 89% **Baseline ZeroR, 4,5%**
- **Aber:** Unabhängigkeit der Features<> Instances nicht gegeben. Jeder Artist ist ultimativ unterscheidbar durch eine Dimension (1.0) → Overfitting
- Geplant:Nur Ähnlichkeiten der Top 10 Artists eines Genres als Features, Alle überbleibenden Artists als Instances

Lazy IBk

TP Rate	FP Rate	Precision	Class
1	0	1	reggae
0.938	0	1	alt.rockindie
0.875	0.01	0.875	folk
1	0.005	0.941	jazz
1	0.01	0.889	pop
0.938	0	1	punk
1	0.005	0.941	electronica
0.875	0.005	0.933	country
1	0	1	classical
1	0	1	heavymetalhardrock
0.938	0.005	0.938	rocknroll
0.938	0.005	0.938	rnbsoul
1	0	1	raphiphop
0.938	0	1	blues
0.96	0.003	0.961	(Weighted Avg)

Work in progress

Versuch Genre Clustering

- 14 Genres → kMeans Clustering mit 14 Cluster?!

country	Reggae 2x
Folk	country
Jazz	electronica
Blues	raphiphop 3x
rnbsoul	heavymetalhardrock
hHeavymetalhardrock	punk 3x
Alt.rockindie	Classical
punk	rnbsoul
Raphiphop	jazz
electronica	
reggae	
RocknRoll	
pop	
classical