RadioMining

Stefan Magerstedt
anginf.de

Chaos Communication Congress – 38c3
28. Dezember 2024, Hamburg

RadioMining

Stefan „starcalc“ Magerstedt

Chaostreff Dortmund

Head of Group IT

RadioMining?

Radio

Playlisten

Data Mining

„Immer die gleichen Lieder“

„Wie schnell sind die Songs eigentlich“

„Es werden nur Lieder aus den Charts gespielt “

→ Dann mal Daten sammeln gehen…

Ein wenig Theorie: ETL

Extract

Scraping der Webseiten oder der dahinterliegenden APIs

Transform

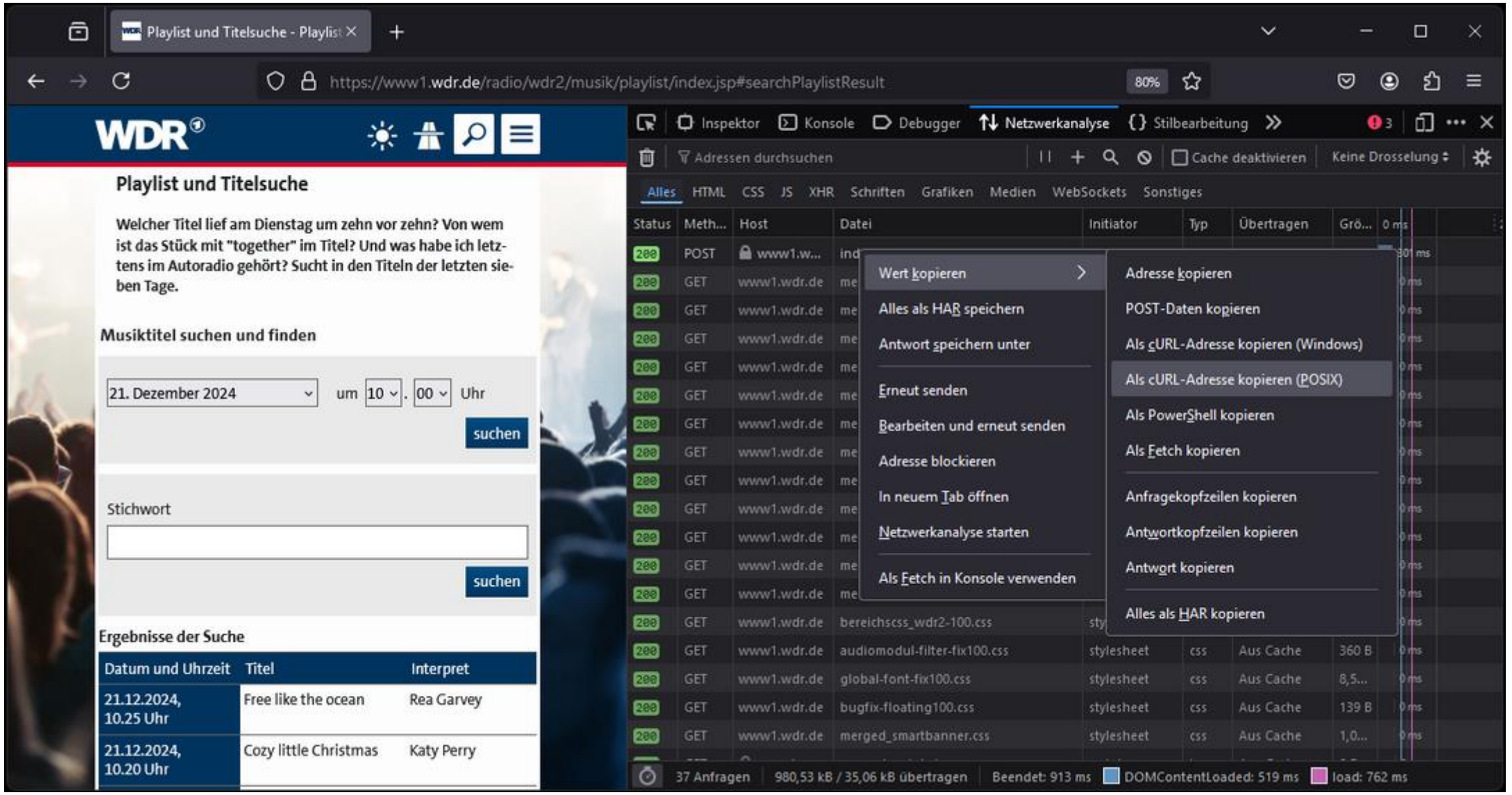
Alle Daten in dieselbe Form überführen

→ Insbesondere Datum

Load

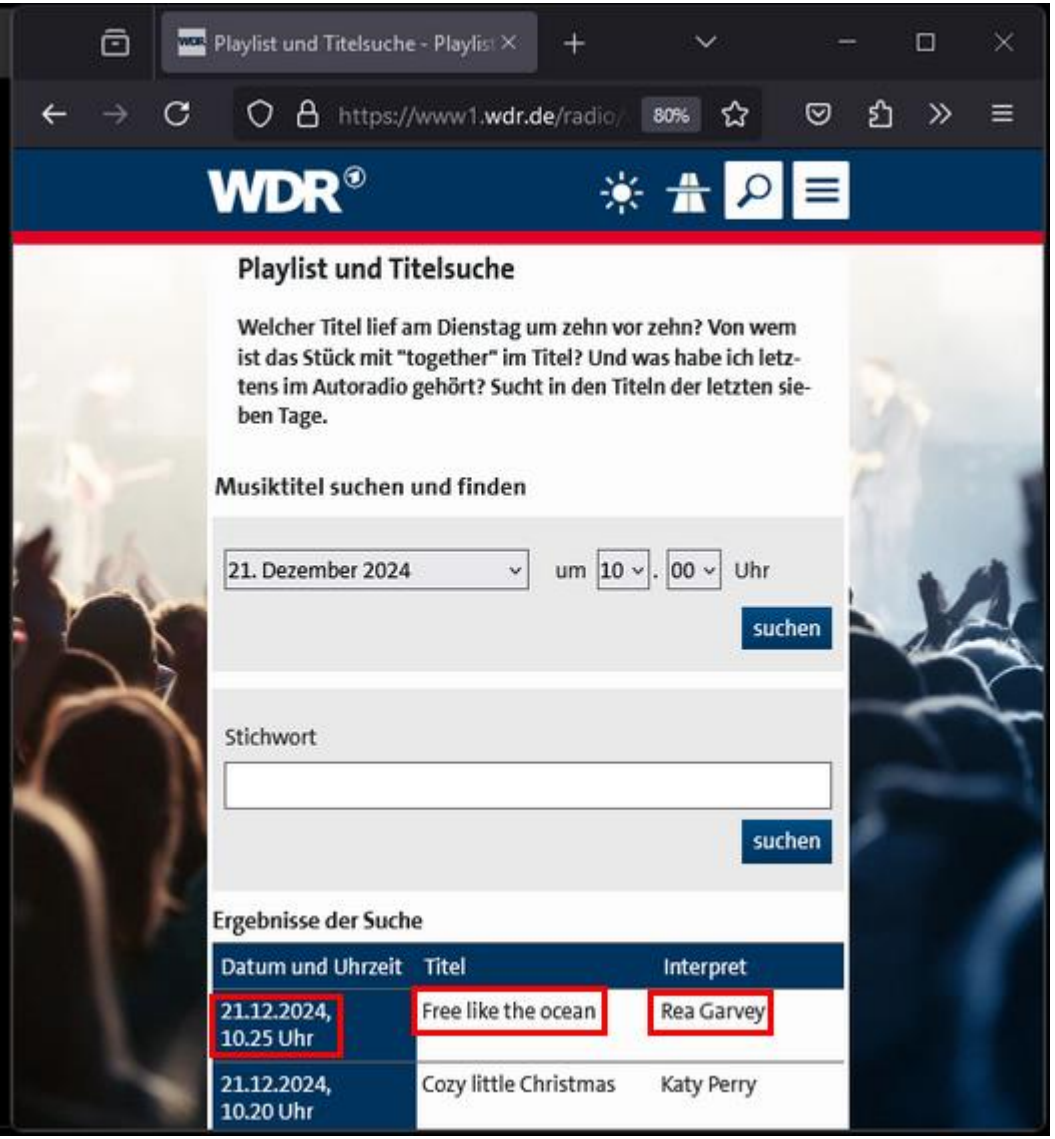
Import in eine Datenbank

Extract: Beispiel „WDR2“



Extract: Beispiel „WDR2“

```
<tr class="data">
<th scope="row" class="entry datetime">
21.12.2024,<br>10.25 Uhr
</th>
<td class="entry title">
Free like the ocean
</td>
<td class="entry performer">
Rea Garvey
</td>
</tr>
<tr class="data">
<th scope="row" class="entry datetime">
21.12.2024,<br>10.20 Uhr
</th>
<td class="entry title">
Cozy little Christmas
</td>
```



Extract: Beispiel „WDR2“

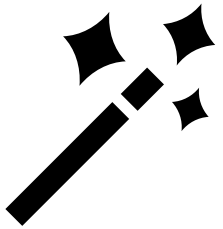
starcalc@radio: ~

```
<tr class="data">
<th scope="row" class="entry datetime">
21.12.2024,<br>10.25 Uhr
</th>
<td class="entry title">
Free like the ocean
</td>
<td class="entry performer">
Rea Garvey
</td>
</tr>
<tr class="data">
<th scope="row" class="entry datetime">
21.12.2024,<br>10.20 Uhr
</th>
<td class="entry title">
Cozy little Christmas
</td>
```

Extract: Beispiel „WDR2“

starcalc@radio: ~

```
<tr class="data">
<th scope="row" class="entry datetime">
21.12.2024,<br>10.25 Uhr
</th>
<td class="entry title">
Free like the ocean
</td>
<td class="entry performer">
Rea Garvey
```



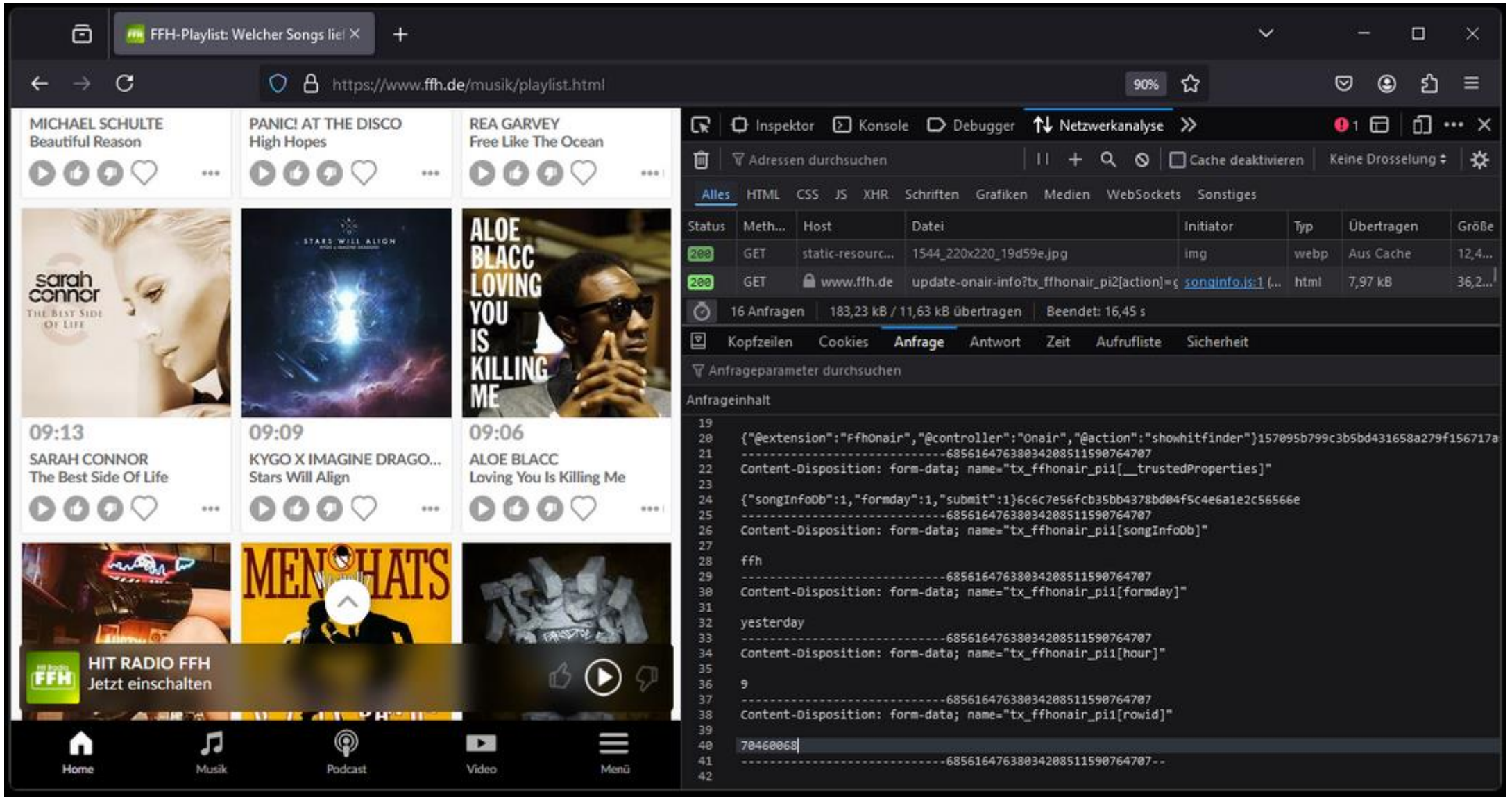
grep/sed magic

```
grep -A 7 '<th scope="row" class="entry datetime">' | \
tr -d '\n' | sed 's/--/\n/g; s/\ class="entry datetime"//g; s/<br>//g' | \
sed -e '$a\' | he --decode | grep -v "^$" | sed 's#;# #g' | \
sed 's#<th.*row">##; s#,# #; s#</th><td.*title">#;#; s#<td class.*performer">#;#; s#</td>##' | \
sed 's/\.\([0-9][0-9]\)\ Uhr/:\1/; s/\([0-9]\{1,2\}\).\([0-9]\{2\}\).\([0-9]\{4\}\)/\3-\2-\1/'
```

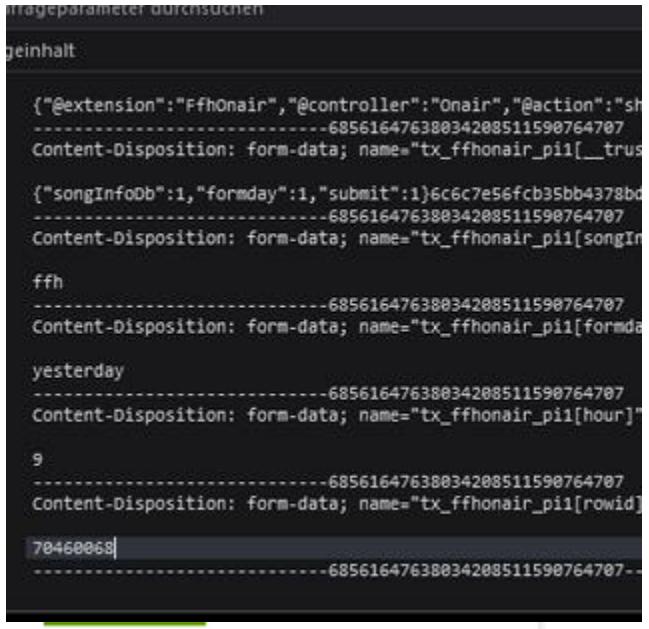
```
</th> 2024-12-21 10:35;Wonders;Michael Patrick Kelly
<td cl 2024-12-21 10:25;Free Like The Ocean;Rea Garvey
Cozy l 2024-12-21 10:20;Cozy Little Christmas;Katy Perry
</td>
```

RadioMining

Extract Beispiel „Radio FFH“



Extract Beispiel „Radio FFH“



Nur Heute / Gestern / Vorgestern

Anfrage lang und konfus

Man muss auf „Weitere Titel davor“ klicken

In der Anfrage steht kein Datum oder Uhrzeit sondern was von rowId

```
Content-Disposition: form-data; name="tx_ffhonair_pi1[rowid]"

70460068
```

Extract Beispiel „Radio FFH“

```
Content-Disposition: form-data; name=\"tx_ffhonair_pi1[songInfoDb]\"

ffh
--------------------------[redacted]
Content-Disposition: form-data; name=\"tx_ffhonair_pi1[formday]\"

yesterday
--------------------------[redacted]
Content-Disposition: form-data; name=\"tx_ffhonair_pi1[hour]\"

1
--------------------------[redacted]
Content-Disposition: form-data; name=\"tx_ffhonair_pi1[rowid]\"

${ROWID}
--------------------------[redacted]--
```

Eigene Werte für ROWID eingeben

yesterday etc. wird ignoriert

Was ist dann RowID = 1?

```
1;11:42;Bon Jovi;Runaway
```

```
2018-11-13 11:42;Runaway;Bon Jovi
```

Rückwärts laufen und Uhrzeit als Datumsindikator nutzen

Danke FFH

Extract: Hinweise

- Teilweise identische APIs
- Glaubt nicht den Angaben der Webseiten
  (Die Anfragen liefern teilweise mehr Daten als die Webseite behauptet)
- Manchmal werden nur 15 Minuten gezeigt statt einer Stunde (mehr Abfragen erforderlich)
- Glaubt nicht direkt den Daten:
  - Artist „SALT“, Titel „AVA MAX“?
  - Artist „Collins“, Titel „Phil“?

```
"23:56;Neelix feat. The Gardener & The Tree;Waterfall"
"24:00;CamelPhat & Jake Bugg;Be Someone"
"24:03;Yves Larock;Rise up"
"24:06;Endor;Pump it up"
"24:08;Phil Fuldner;Miami Pop"
"24:12;Yves V feat. Afrojack & Icona Pop;We Got That Cool"
"24:17;Kungs feat. Olly Murs & Coely;More Mess (Hugel Remix)"
```

(„UnserDing“, Silvester 2019)

RadioMining

Extract: Übersicht der Abfragen

Daten ab...	Format: JSON	Format: HTML	rowId
2 Tage	Antenne Niedersachsen, HITRADIO RTL, JAM FM, RTL Radio, 105'5 Spreeradio, UNSERDING	SR1	
7 Tage	Radio NRJ, Radio 91.2		
14-15 Tage	bigFM, bigFMsaar, KISS FM, RPR1., R.SH	Antenne MV, ffn, hr3, YOU FM	
26-31 Tage	MDRSputnik	Absolut HOT,Bayern3, PULS	
60-68 Tage		1LIVE, NDR2, N-JOY, WDR2	
90 Tage		DASDING, egoFM, SWR3	
1 Jahr	MDR Jump		
1,5 Jahre		Bremen NEXT, Bremen Vier	
08.07.2021	radio TOP 40		
13.11.2018			HIT RADIO FFH, planet radio

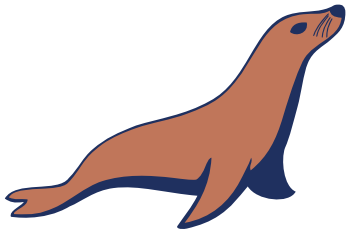
Transform: Insbesondere Datumsform

2024-12-21T10:03:16+01:00
2024-12-21T10:03+0100
2024-12-21T10:03
2024-12-21 10:03
2024-12-21 10:03:16
10:03 Uhr;21. Dez 2024
10:03 Uhr, am 21.12.2024
10:03
1734771796

Load: In eine Datenbank

Vorüberlegungen:

Wir wollen viele Abfragen ausführen, die auf Aggregation basieren
Kommerzielle Anbieter sind unattraktiv



MariaDB

ClickHouse

Langsam
Keine regex etc.

Zeilenorientiert
Langsam bei Abfragen
Schnell bei UPDATEs

Spaltenorientiert
Sehr schnell bei Aggregation

Bereinigen der Daten, Ergänzen der Daten

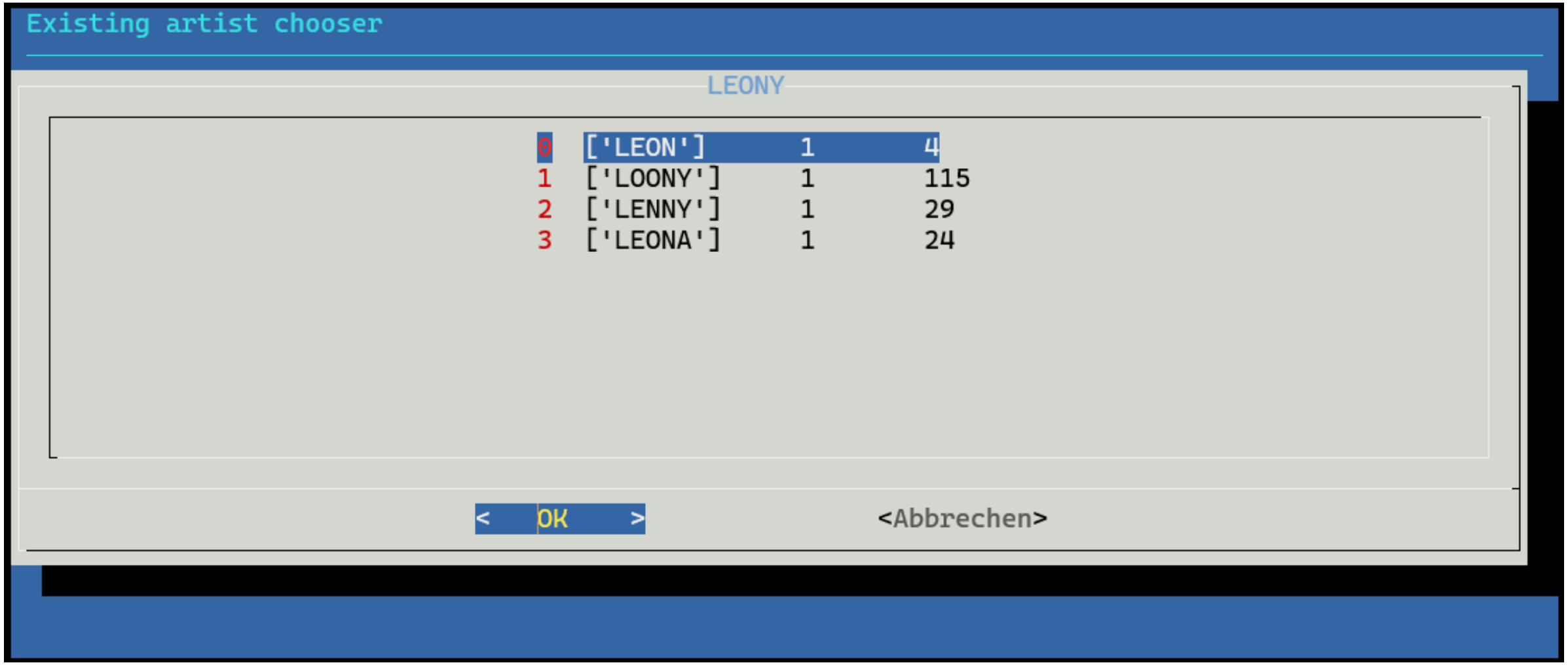
Besonderheiten erkennen oder wegfiltern:

(Remaster 2016)
(2016 Remaster)
(Radio Edit) (Radio Mix)
Klammerarten () [] {}
feat. vs. x , ; / & -
!!! NEU !!!

Hamming-Distanz (1 Zeichen unterschiedlich mit ncurses-Menu)

```
starcalc@radio:~$ ./wrong_spelling.sh
I\'ll Be There
I\'ll be there
I\'LL BE THERE
ILL BE THERE
I\'ll Be There*
I\'ll Be There**
I\\\\"ll Be There
I’ll Be There
I‘ll Be There
I‘ll Be There**
(B) I\'LL BE THERE
I\'Ll Be There
I’Ll Be There
I‘LL BE THERE
(Hook) I\'Ll Be There
```

Bereinigen der Daten, Ergänzen der Daten



Exkurs: spotify

Warum nicht jemanden fragen, der sich damit auskennt?

API liefert meistens einen Song zurück, sogar die BPM

Web API • References / Tracks / Get Track's Audio Features

Get Track's Audio Features OAuth 2.0 Deprecated

Get audio feature information for a single track identified by its unique Spotify ID.

https://developer.spotify.com/blog/2024-11-27-changes-to-the-web-api

spotify hat immer noch ein 30 Sekunden preview verfügbar...

preview laden, lokal an eine Tempo-Erkennungs-Software

Schei Encoding

```
TI?STO
TIESTO
TISTO
TIËSTO
TI◆STO
```

```
HERBERT GRÖNEMEYER
HERBERT GROENEMEYER
HERBERT GR∄NEMEYER
HERBERT GRNEMEYER
HERBERT GR?NEMEYER
```

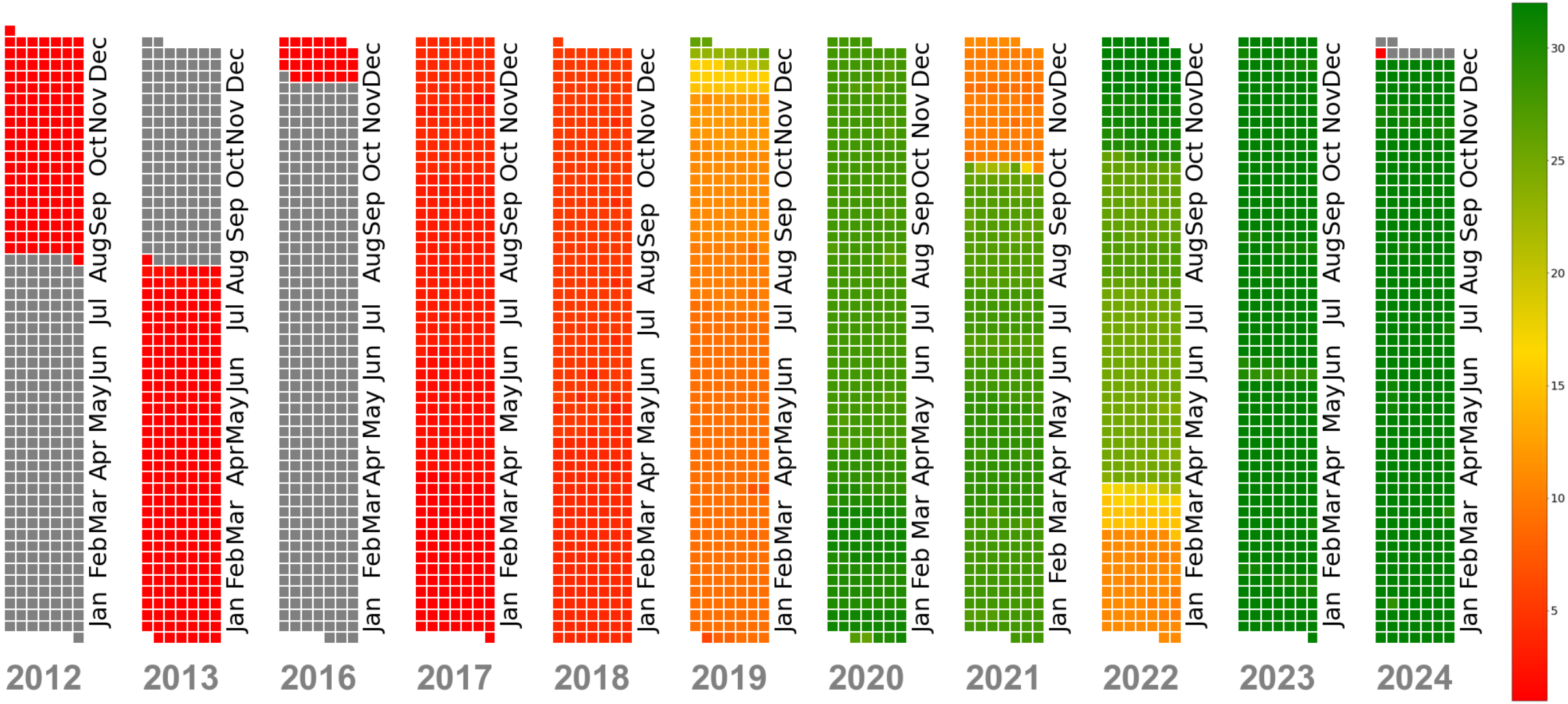
```
SEKUNDENGLüCK
SEKUNDENGLüCK*
SEKUNDENGLÜCK
SEKUNDENGLUECK
SEKUNDENGL∄7CK
SEKUNDENGL◆CK
```

```
AERZTE
DIE ÄRZTE
DIE ∄ᚾRZTE
DIE ◆RZTE
ÄRZTE
∄ᚾRZTE
◆RZTE
```

```
PEDRO CAP
PEDRO CAPO
PEDRO CAPó
PEDRO CAPÓ
PEDRO CAP∄
```

```
AGOSTINO, GIGI D'
AGOSTINO, GIGI D`
D'AGOSTINO, GIGI
D`AGOSTINO, GIGI
D´AGOSTINO, GIGI
DAGOSTINO, GIGI
D\'AGOSTINO, GIGI
GIGI D'AGOSTINO
GIGI D' AGOSTINO
GIGI D'  AGOSTINO
GIGI D´AGOSTINO
GIGI D`AGOSTINO
GIGI DAGOSTINO
GIGI D AGOSTINO
GIGI D/AGOSTINO
GIGI D^AGOSTINO
GIGI D\'AGOSTINO
GIGI D\\\"'"AGOSTINO
GIGI DÁGOSTINO
```

Haben wir noch Daten? Zwei noch?



(Single) Charts

Verwendung der „Offiziellen Charts“, geführt seit 1959, Einträge ab 1977 online verfügbar

Ursprünglich „nur“ Plattenverkäufe
2004 Downloads → 2007 Umsatz statt Verkauf → 2014 Premium-Streams → 2020 Airplay → 2022 Ad-Streams

Downloads ähnlich wie die Playlisten der Radiosender, aber mit Platzierung.

```
starcalc@radio:~$ head -n 10 561038400
1;Rick Astley;Never Gonna Give You Up
2;Sabrina;Boys
3;Bee Gees;You Win Again
4;Desireless;Voyage Voyage
5;Michael Jackson;Bad
6;Depeche Mode;Never Let Me Down Again
7;Sandra;Everlasting Love
8;Francesco Napoli;Balla... Balla!
9;Pet Shop Boys With Dusty Springfield;What Have I Done To Deserve This?
10;New Order;True Faith
```

Chart vs. Gespielt pro Woche & Sender (2023)

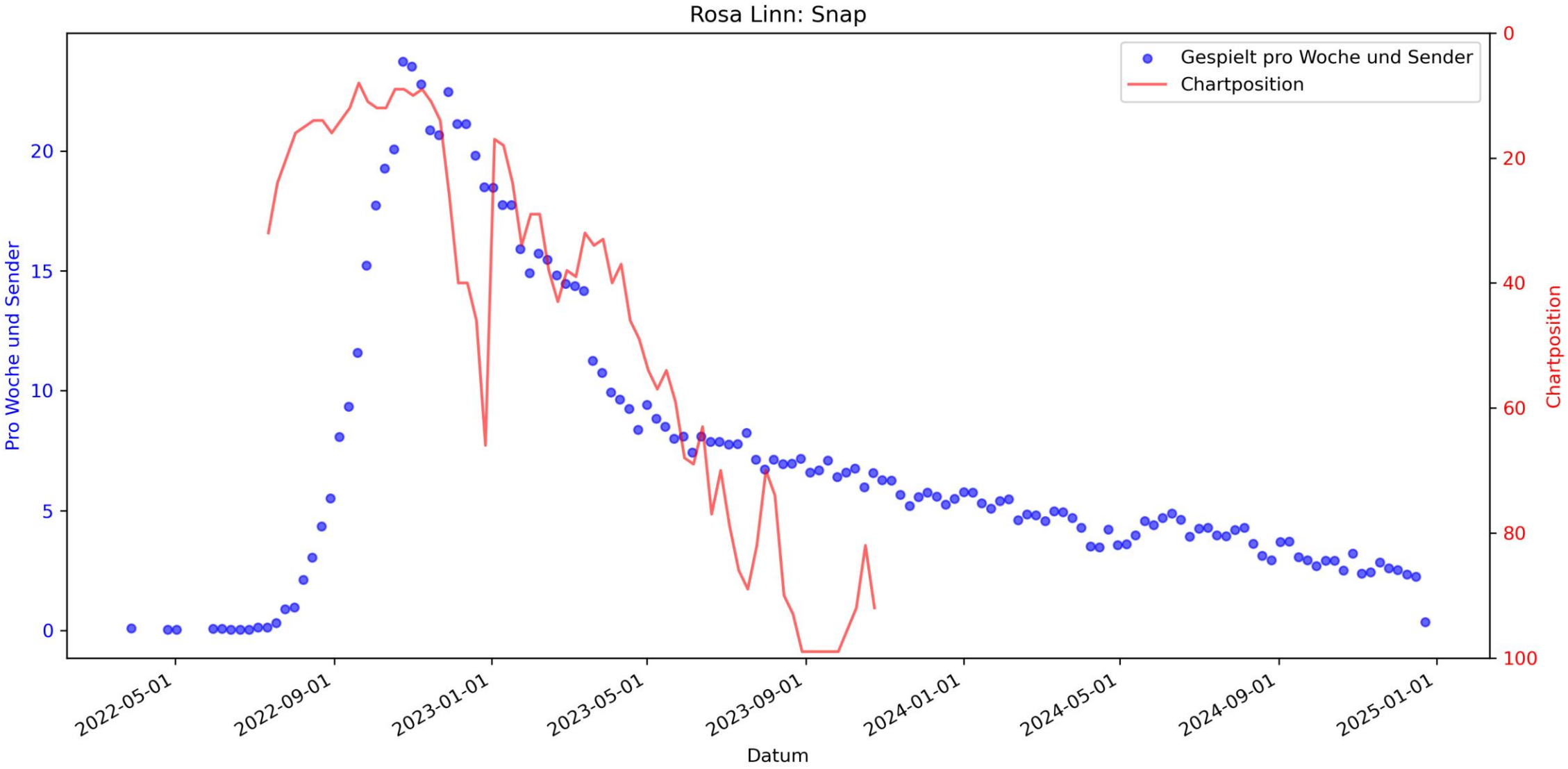


Chart vs. Gespielt pro Woche & Sender (2013)

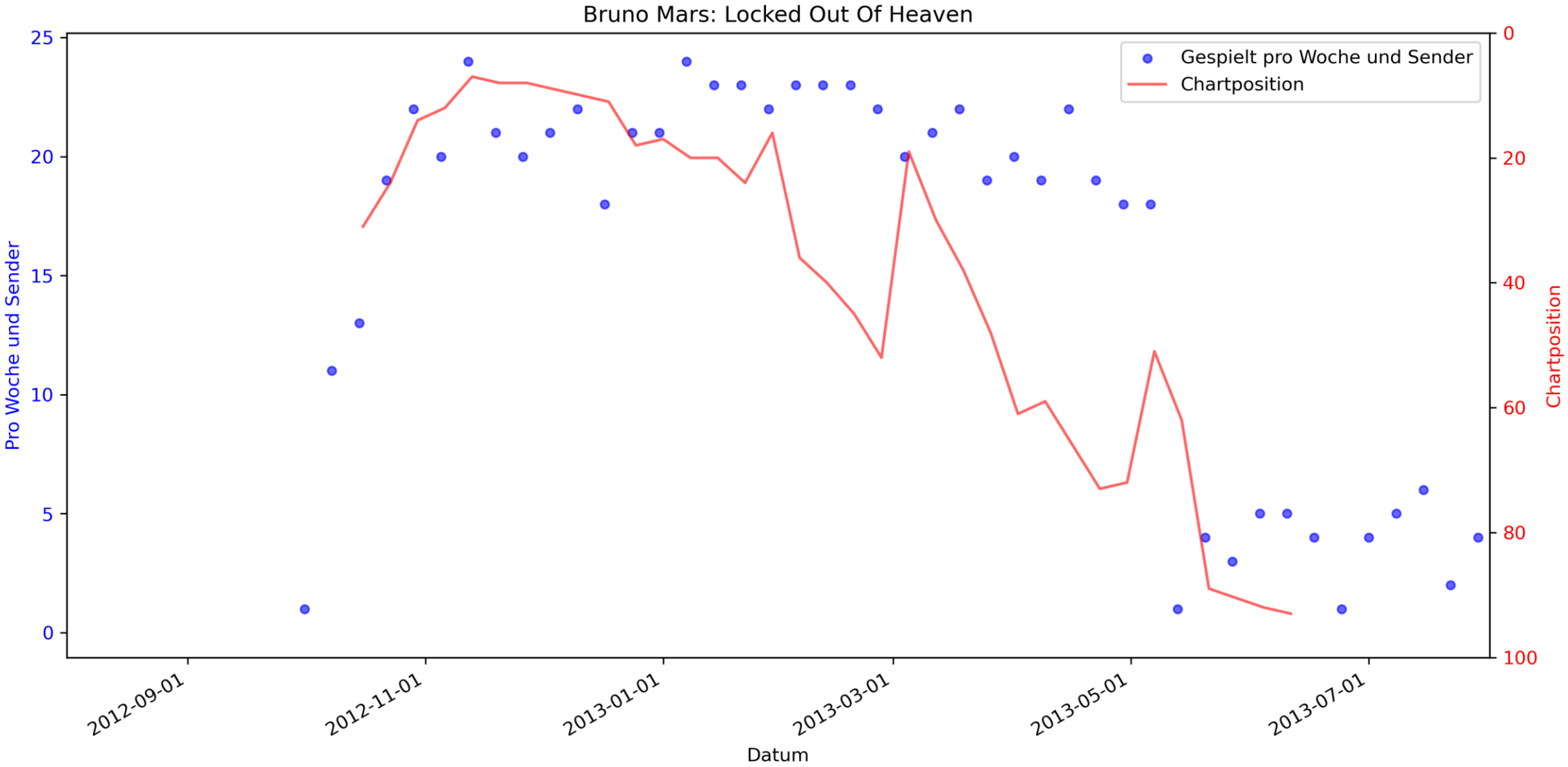
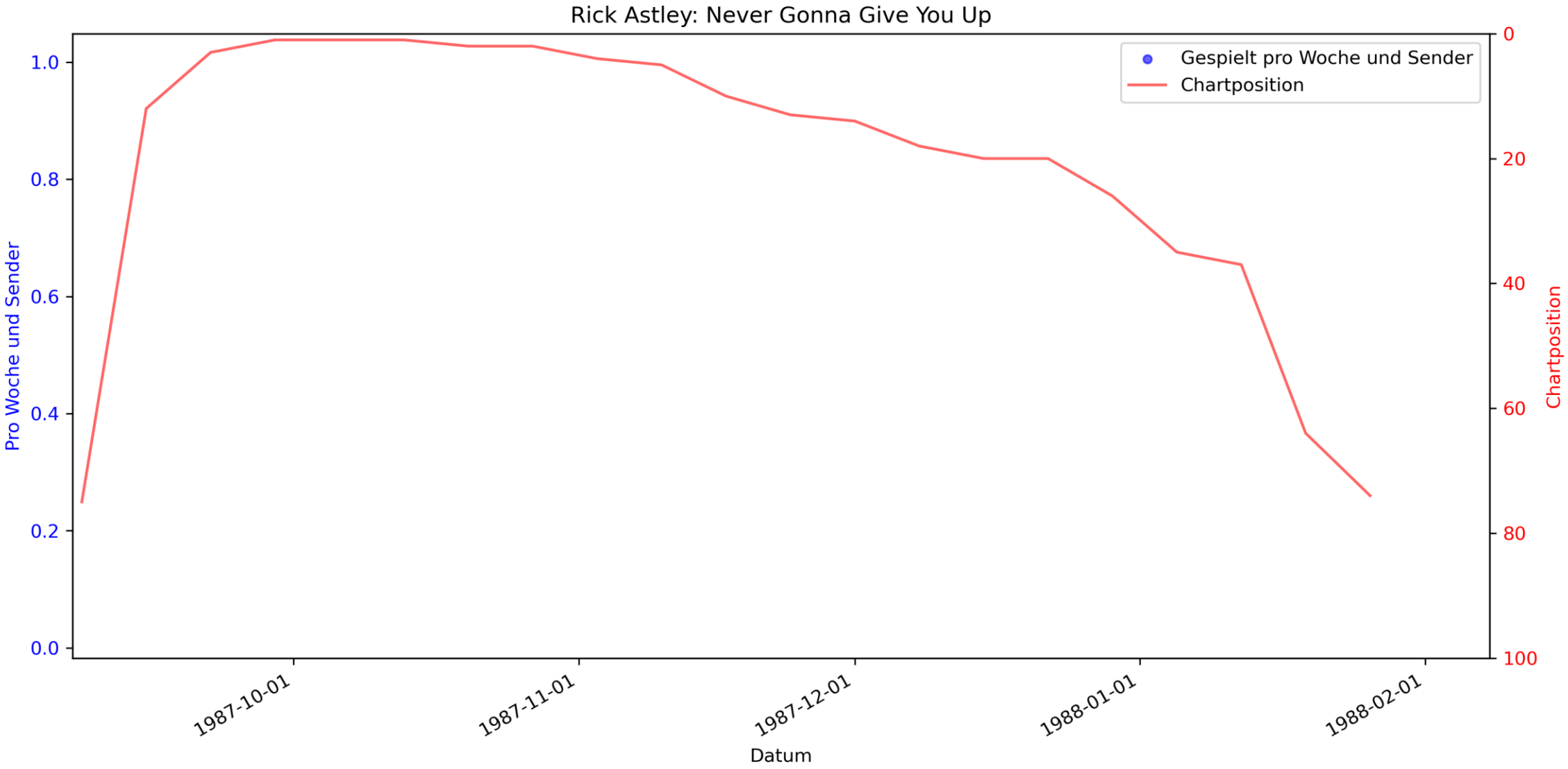
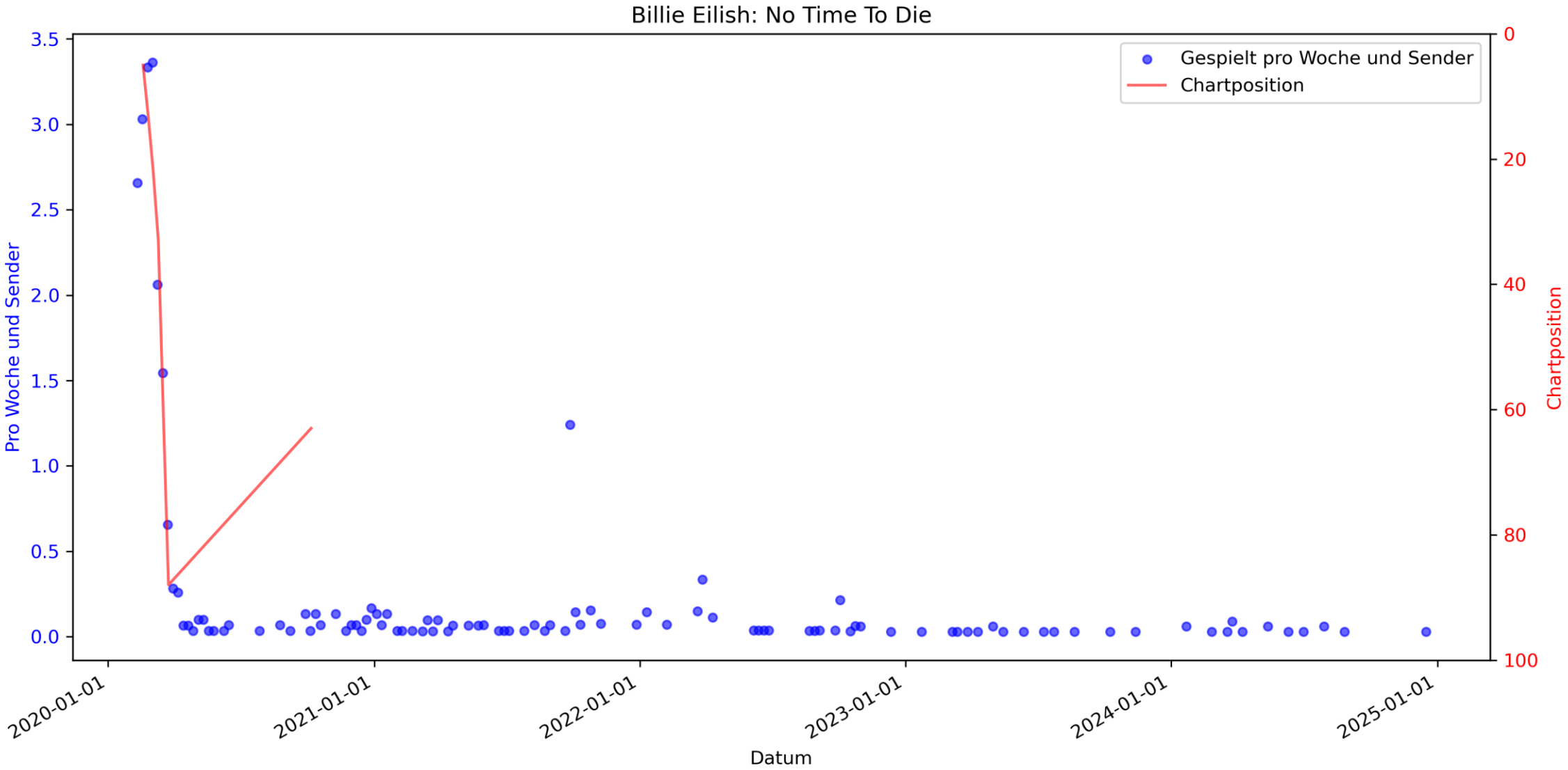


Chart (1987)


Rick Astley: Never Gonna Give You Up
Gespielt pro Woche und Sender
Chartposition
Pro Woche und Sender
Chartposition
Datum
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
1987-10-01
1987-11-01
1987-12-01
1988-01-01
1988-02-01

Und bei Kinofilmen?



Es gibt kein Entrinnen...

27.10.2022 01:31
Glockenbach feat ÁSDÍS: Dirty Dancing (Platz 24 Single Charts)
12 Sender: Absolute HOT, RTL Radio, JAM FM, RPR1, Radio ffn, bigFM
WDR2, SWR3, NDR2, HR3, N-JOY, MDR JUMP

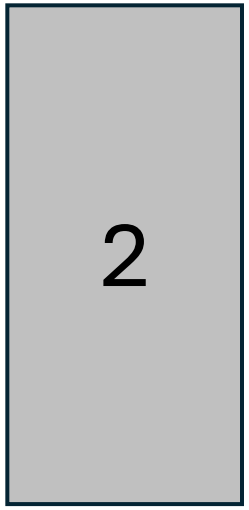
19.07.2024 10:04
Eminem: Houdini (Platz 4 Single Charts)
9 Sender: Bayern 3, radio Top40, youFM, UNSERDING, KISS FM, RTL Radio, JAM FM, bigFM, Bremen Vier

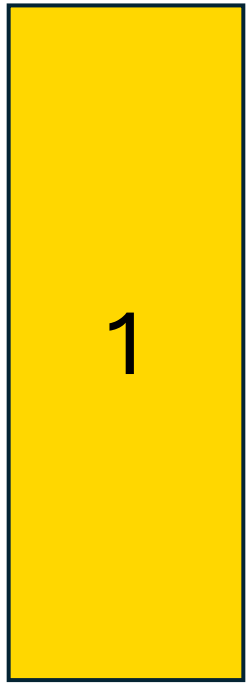
Am häufigsten passiert das mit (mindestens 4 Sender, nicht ARD Popnacht):
Miley Cyrus: Flowers (39 Mal!)
Am meisten betroffen sind dabei die Sender:
youFM (20), bigFM/bigFMsaar (19), SR1 (17)

Spielt denselben Song nochmal!

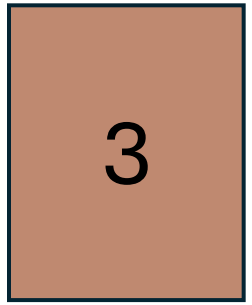
11.04.2022 -- 20 Mal
Iann Dior: Let you
radio Top 40

28.11.2022 -- 16 Mal
Kim Petras: Unholy
KISS FM

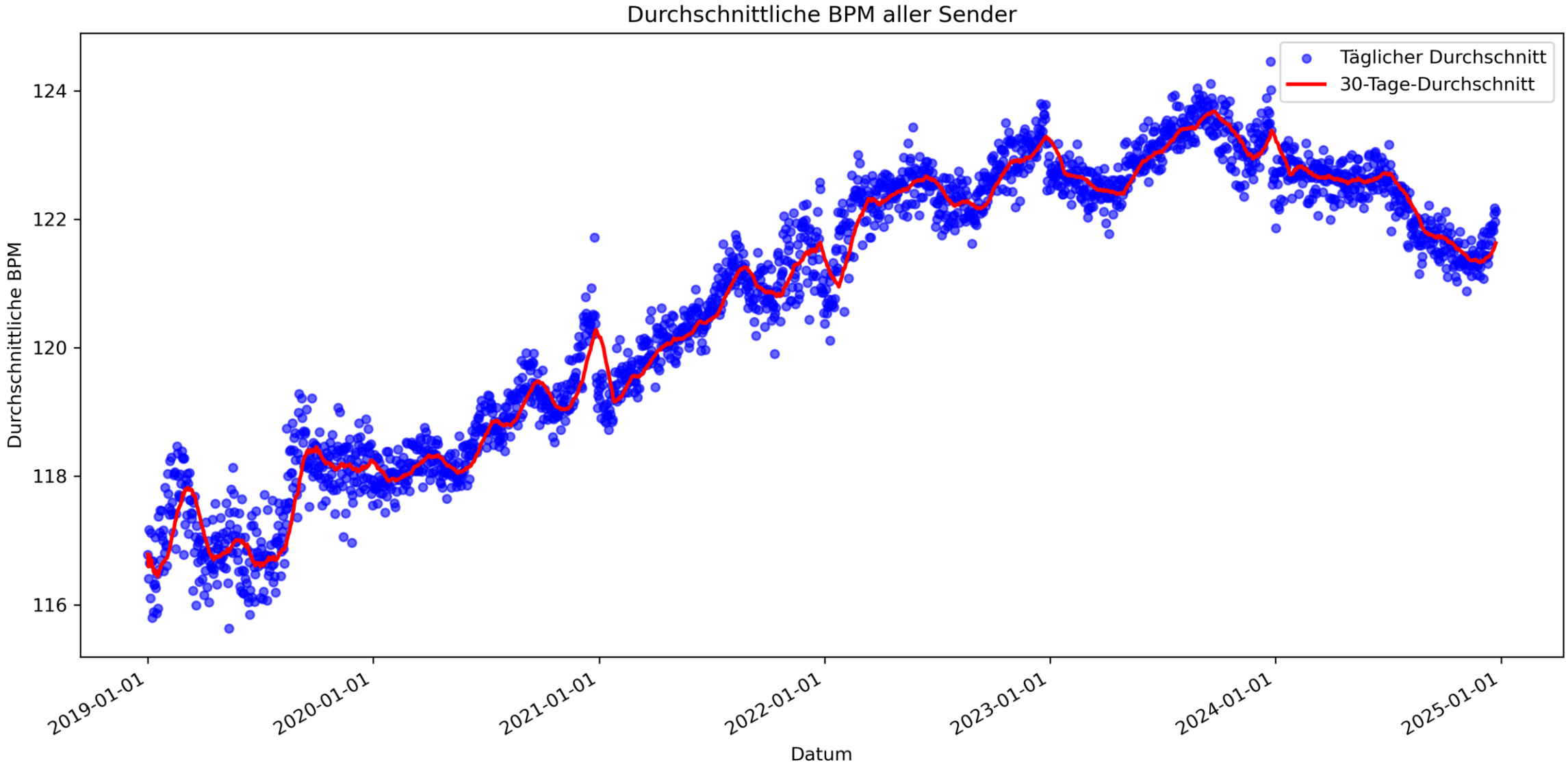




12.04.2024 -- 16 Mal
Nina Chuba: NINA
bigFM



BPM im Radio: Speedrun any%



Songs, die öfter laufen als ich gehe

Was wurde (in meinen vorliegenden Daten) am häufigsten gespielt?

66969: THE WEEKND - BLINDING LIGHTS
44848: REGARD - RIDE IT
39617: MILEY CYRUS - FLOWERS
38021: PURPLE DISCO MACHINE & SOPHIE AND THE GIANTS - HYPNOTIZED
37662: HARRY STYLES - AS IT WAS
37093: A7S & TOPIC - BREAKING ME
36904: TAYLOR SWIFT - ANTI-HERO
36220: KAMRAD - I BELIEVE
36108: MILEY CYRUS - MIDNIGHT SKY
36106: DUA LIPA - PHYSICAL

Songs, die öfter laufen als ich gehe (XMAS)

Was wurde (in meinen vorliegenden Daten) am häufigsten vom 01.-24.12. gespielt?

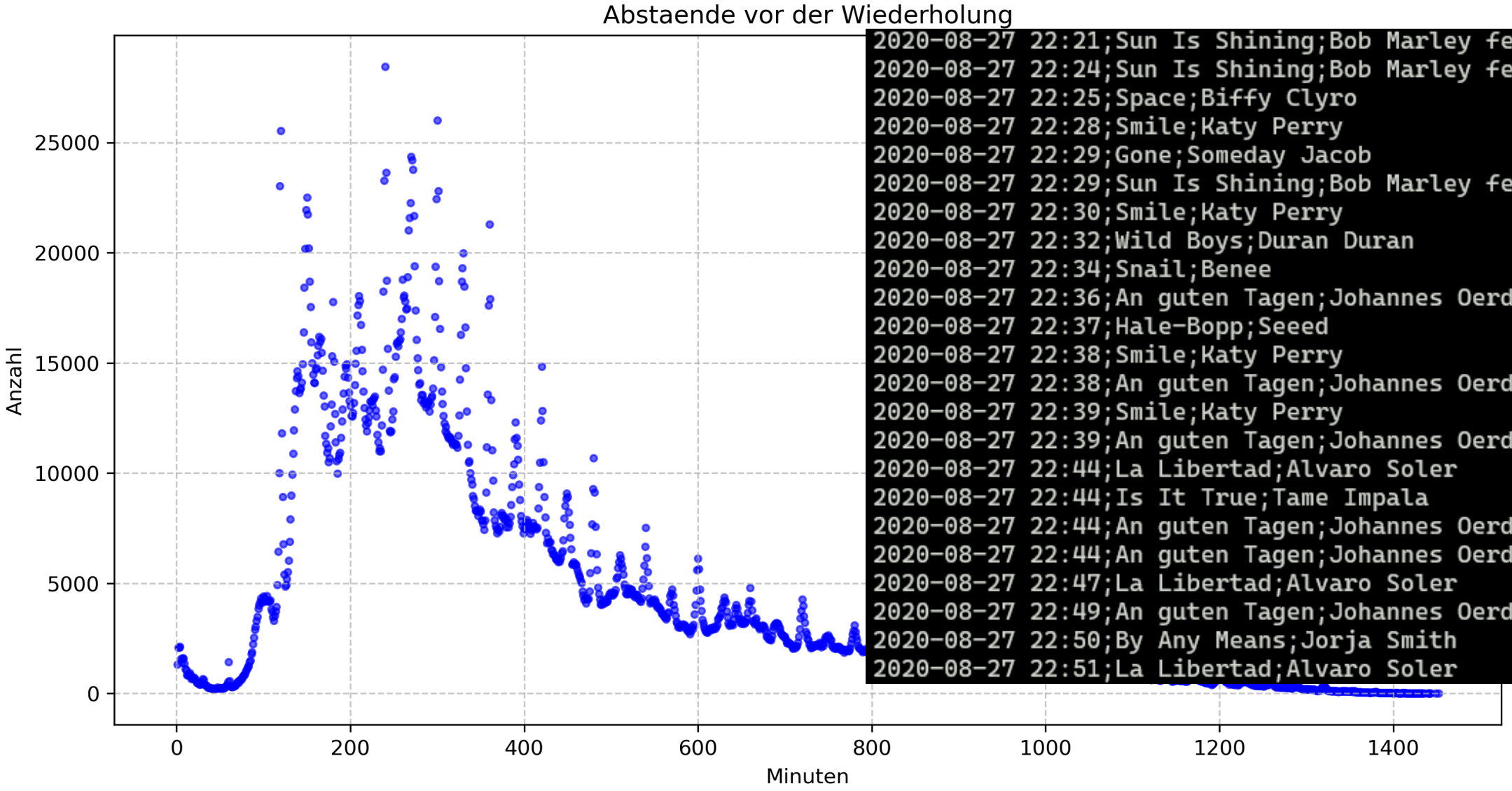
3641: TAYLOR SWIFT - ANTI-HERO
3565: THE WEEKND - BLINDING LIGHTS
3464: ROSA LINN - SNAP
3313: CLOCKCLOCK - SOMEONE ELSE
3231: KENYA GRACE - STRANGERS
3188: MILEY CYRUS - MIDNIGHT SKY
3116: FELIX JAEHN & RAY DALTON - CALL IT LOVE
3050: ED SHEERAN - CELESTIAL
3047: PURPLE DISCO MACHINE & SOPHIE AND THE GIANTS - HYPNOTIZED
3013: MARIAH CAREY - ALL I WANT FOR CHRISTMAS IS YOU

T*nder für Songs

Praktischer Service – die Radio-Redaktionen geben uns frei Haus Songs, die gut hintereinander zu spielen sind.

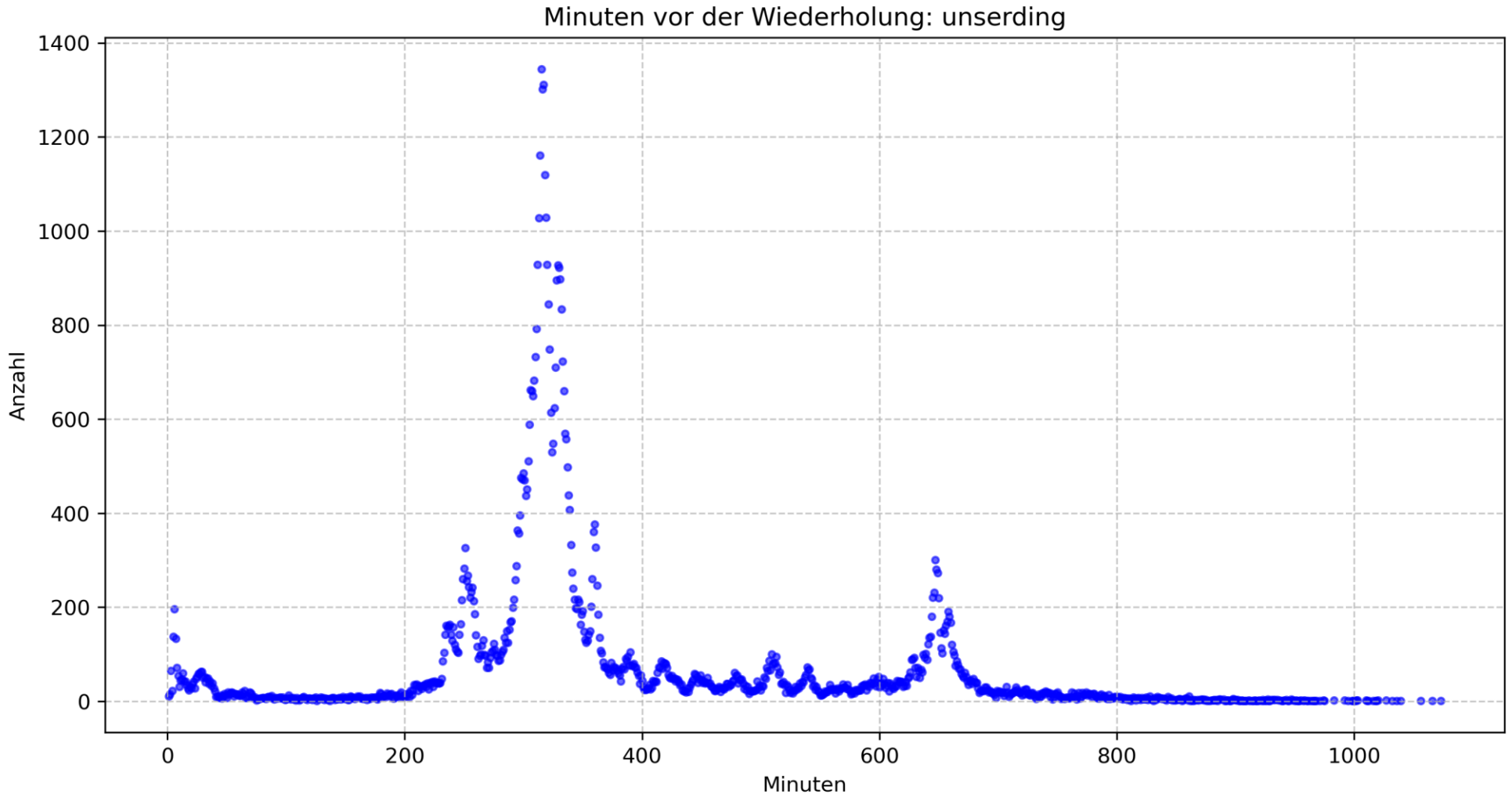
THE WEEKND - BLINDING LIGHTS
21 SAVAGE x METRO BOOMIN x THE WEEKND - CREEPIN'
MILEY CYRUS - FLOWERS
LOI - GOLD
DUA LIPA - HOUDINI
ARTEMAS - I LIKE THE WAY YOU KISS ME
ALIDA x ROBIN SCHULZ - IN YOUR EYES
TIËSTO - LAY LOW
BECKY HILL x GOODBOYS x MEDUZA - LOSE CONTROL
NORMA JEAN MARTINE x OFENBACH - OVERDRIVE
DERMOT KENNEDY x MEDUZA - PARADISE
ONEREPUBLIC - RUNAWAY
JULIAN PERRETTA x KUNGS x PURPLE DISCO MACHINE - SUBSTITUTION
MICHAEL SCHULTE x R3HAB - WATERFALL

Radio-aktive Halbwertszeit



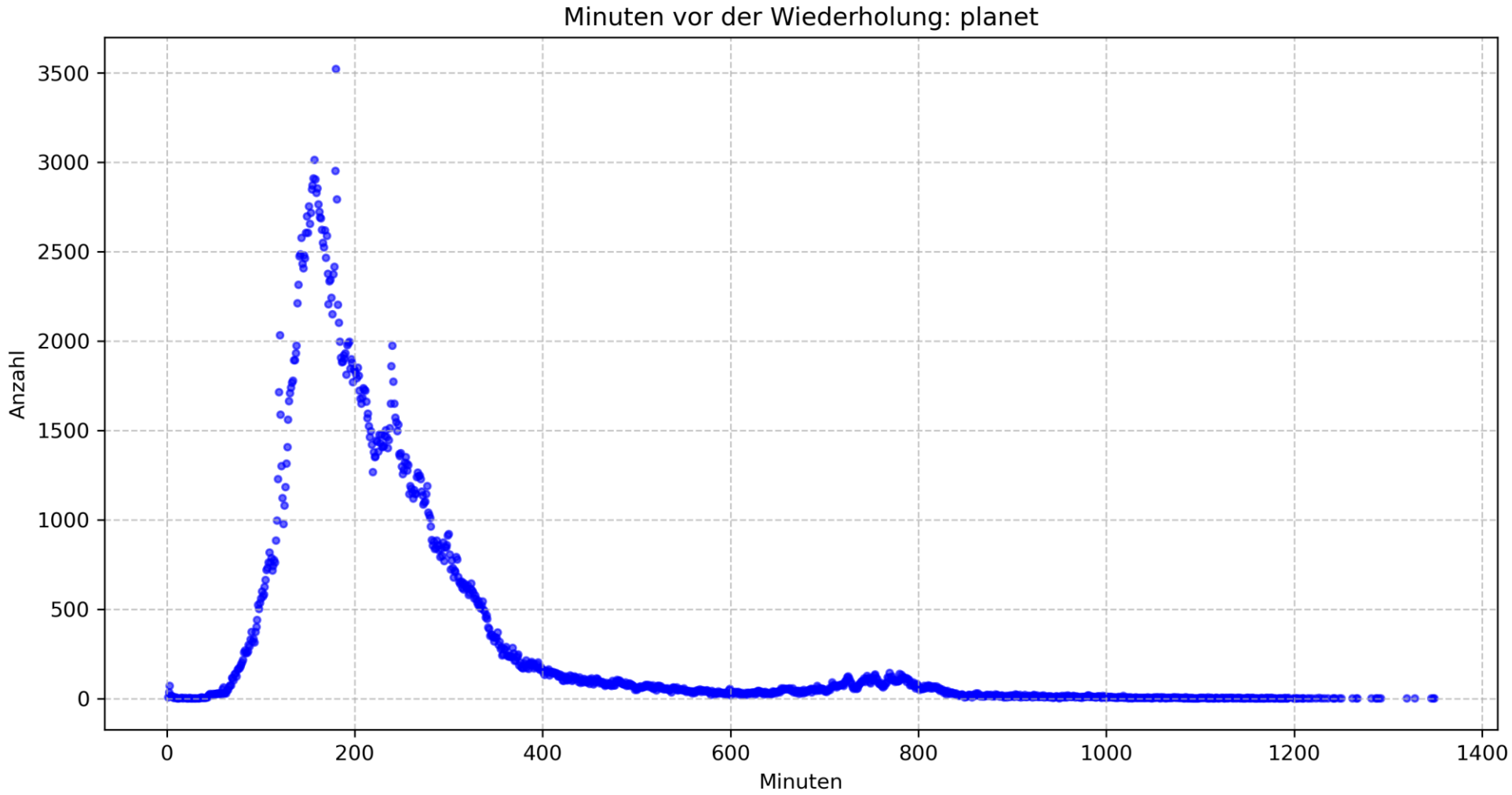

```
2020-08-27 22:21;Sun Is Shining;Bob Marley feat. Robin
2020-08-27 22:24;Sun Is Shining;Bob Marley feat. Robin
2020-08-27 22:25;Space;Biffy Clyro
2020-08-27 22:28;Smile;Katy Perry
2020-08-27 22:29;Gone;Someday Jacob
2020-08-27 22:29;Sun Is Shining;Bob Marley feat. Robin
2020-08-27 22:30;Smile;Katy Perry
2020-08-27 22:32;Wild Boys;Duran Duran
2020-08-27 22:34;Snail;Benee
2020-08-27 22:36;An guten Tagen;Johannes Oerding
2020-08-27 22:37;Hale-Bopp;Seeed
2020-08-27 22:38;Smile;Katy Perry
2020-08-27 22:38;An guten Tagen;Johannes Oerding
2020-08-27 22:39;Smile;Katy Perry
2020-08-27 22:39;An guten Tagen;Johannes Oerding
2020-08-27 22:44;La Libertad;Alvaro Soler
2020-08-27 22:44;Is It True;Tame Impala
2020-08-27 22:44;An guten Tagen;Johannes Oerding
2020-08-27 22:44;An guten Tagen;Johannes Oerding
2020-08-27 22:47;La Libertad;Alvaro Soler
2020-08-27 22:49;An guten Tagen;Johannes Oerding
2020-08-27 22:50;By Any Means;Jorja Smith
2020-08-27 22:51;La Libertad;Alvaro Soler
```

Radio-aktive Halbwertszeit: UNSERDING - 5h



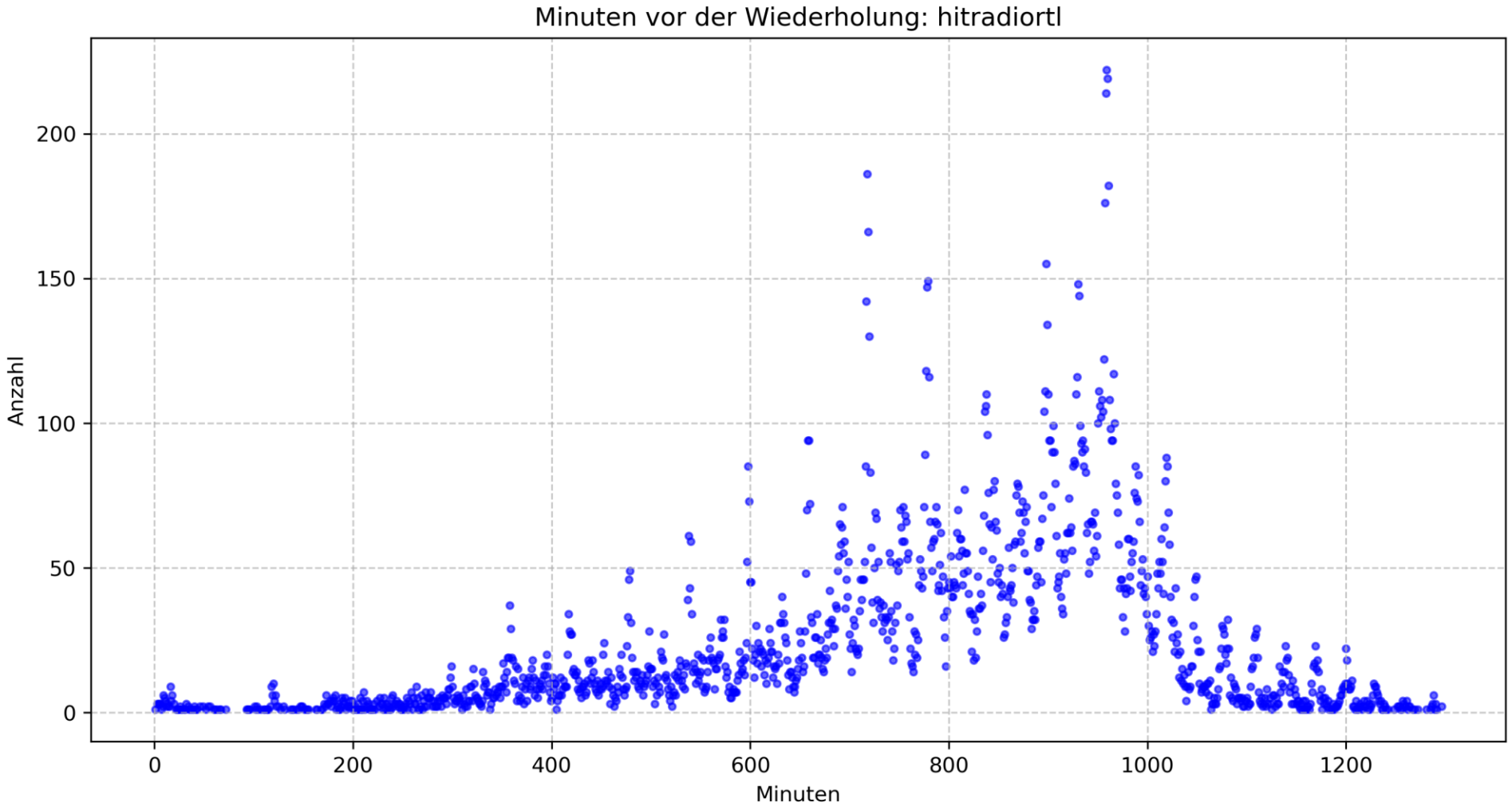
Im Schnitt 1037 unterschiedliche Songs / Woche

Radio-aktive Halbwertszeit: Planet - 3h



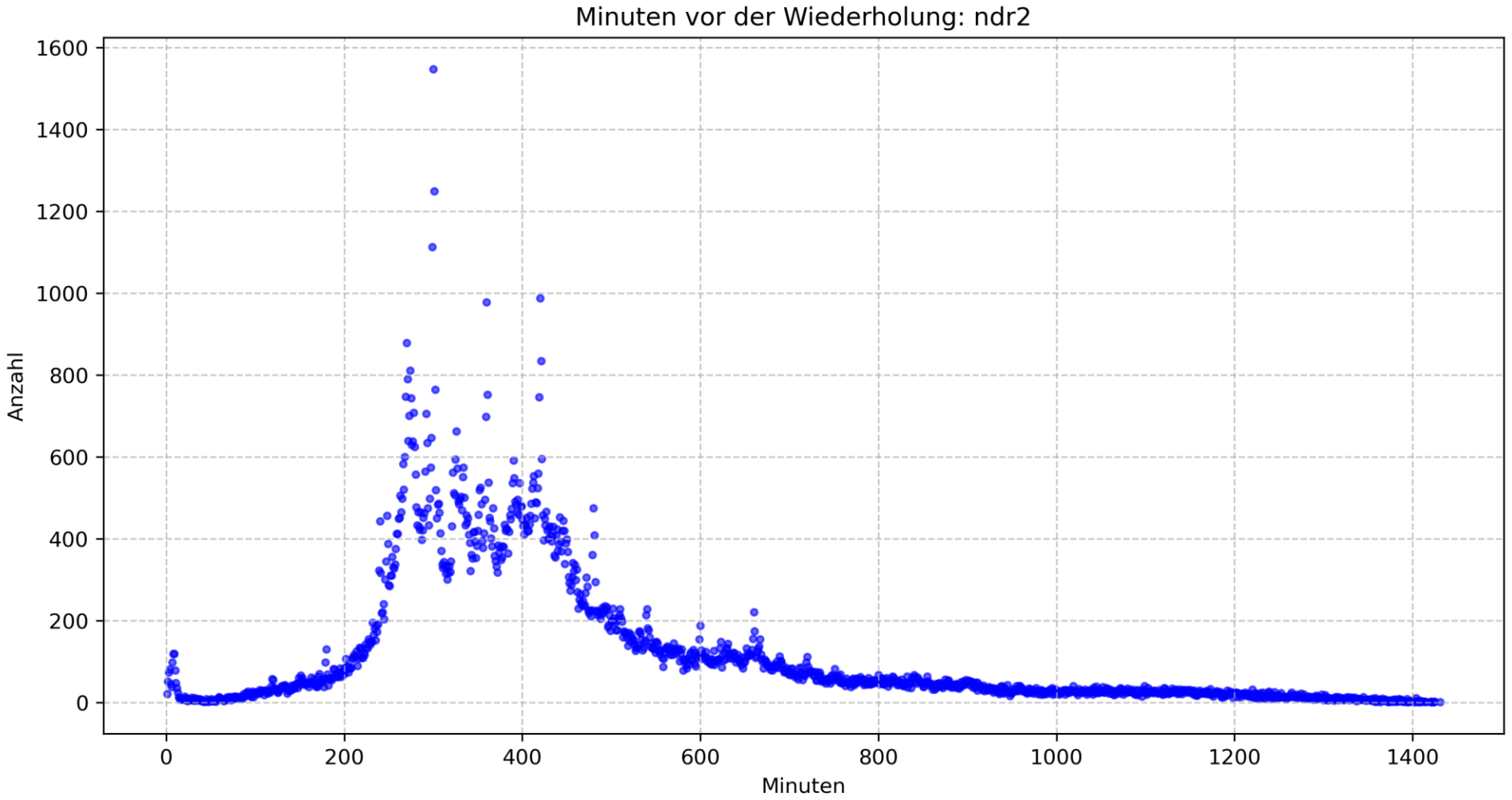
Im Schnitt 1032 unterschiedliche Songs / Woche

Radio-aktive Halbwertszeit: Hitradio RTL



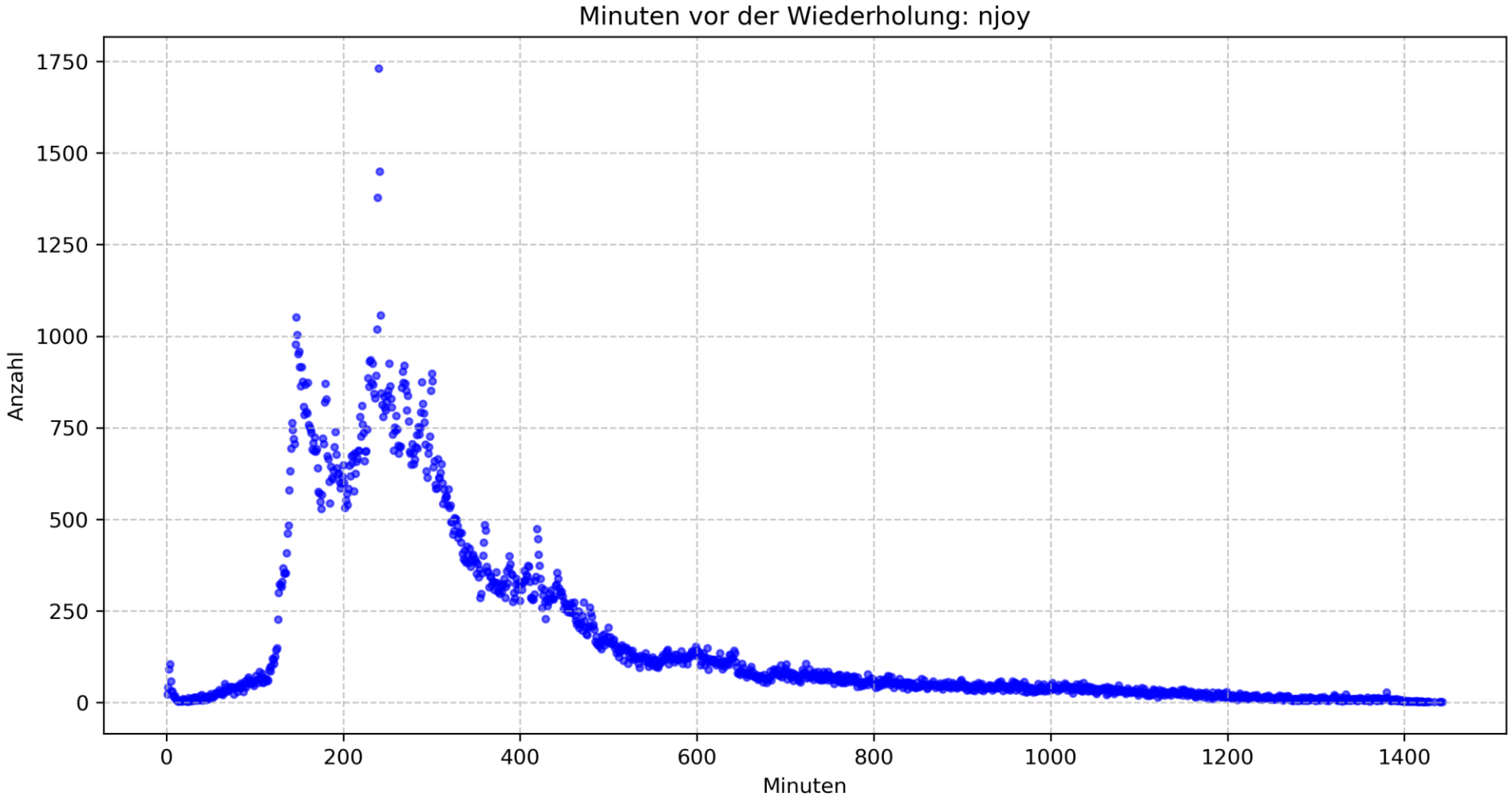
Im Schnitt 925 unterschiedliche Songs / Woche

Radio-aktive Halbwertszeit: Nord: NDR2 - 4h



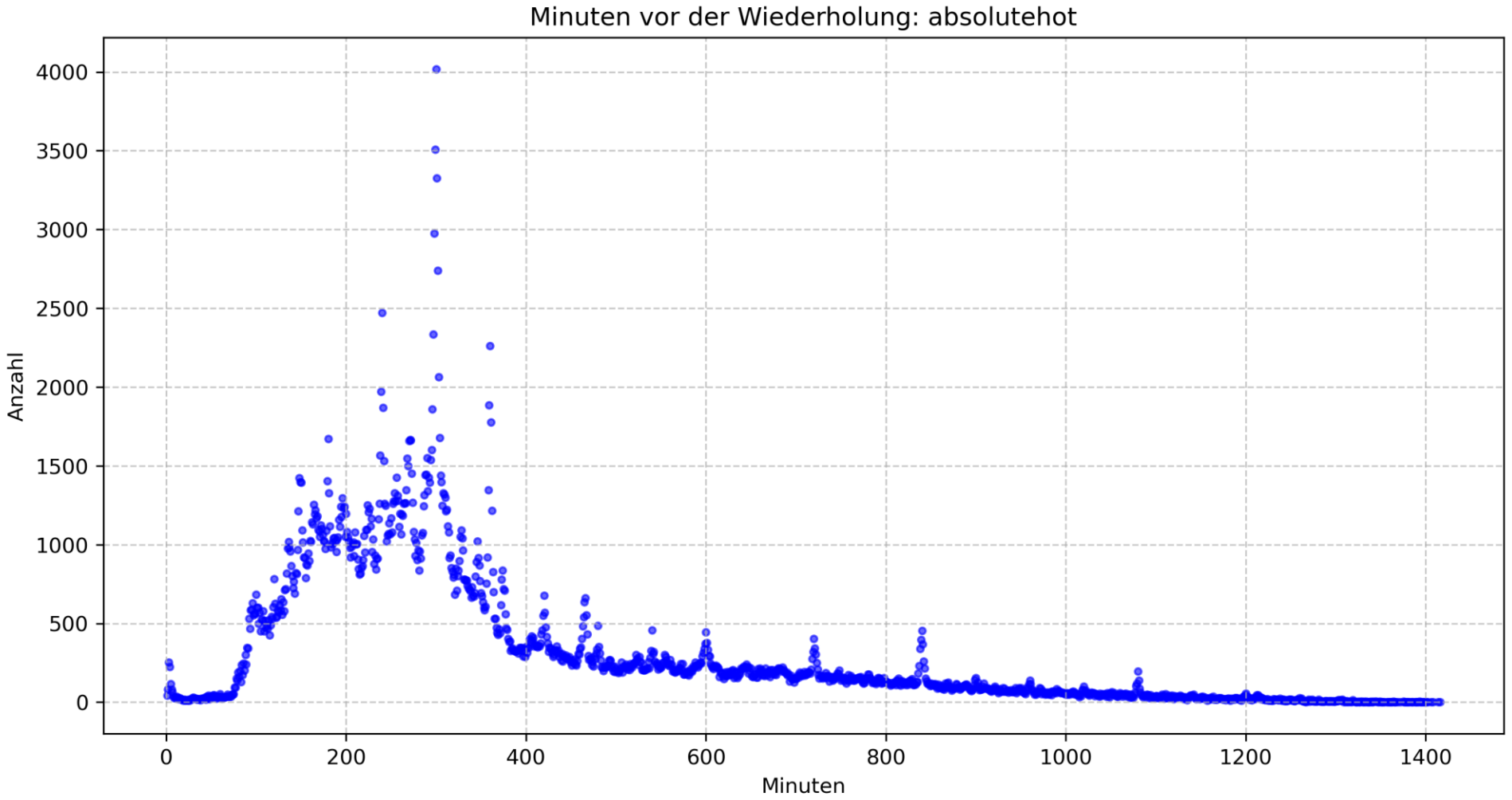
Im Schnitt 1224 unterschiedliche Songs / Woche

Radio-aktive Halbwertszeit: Nord: N-JOY - 3h



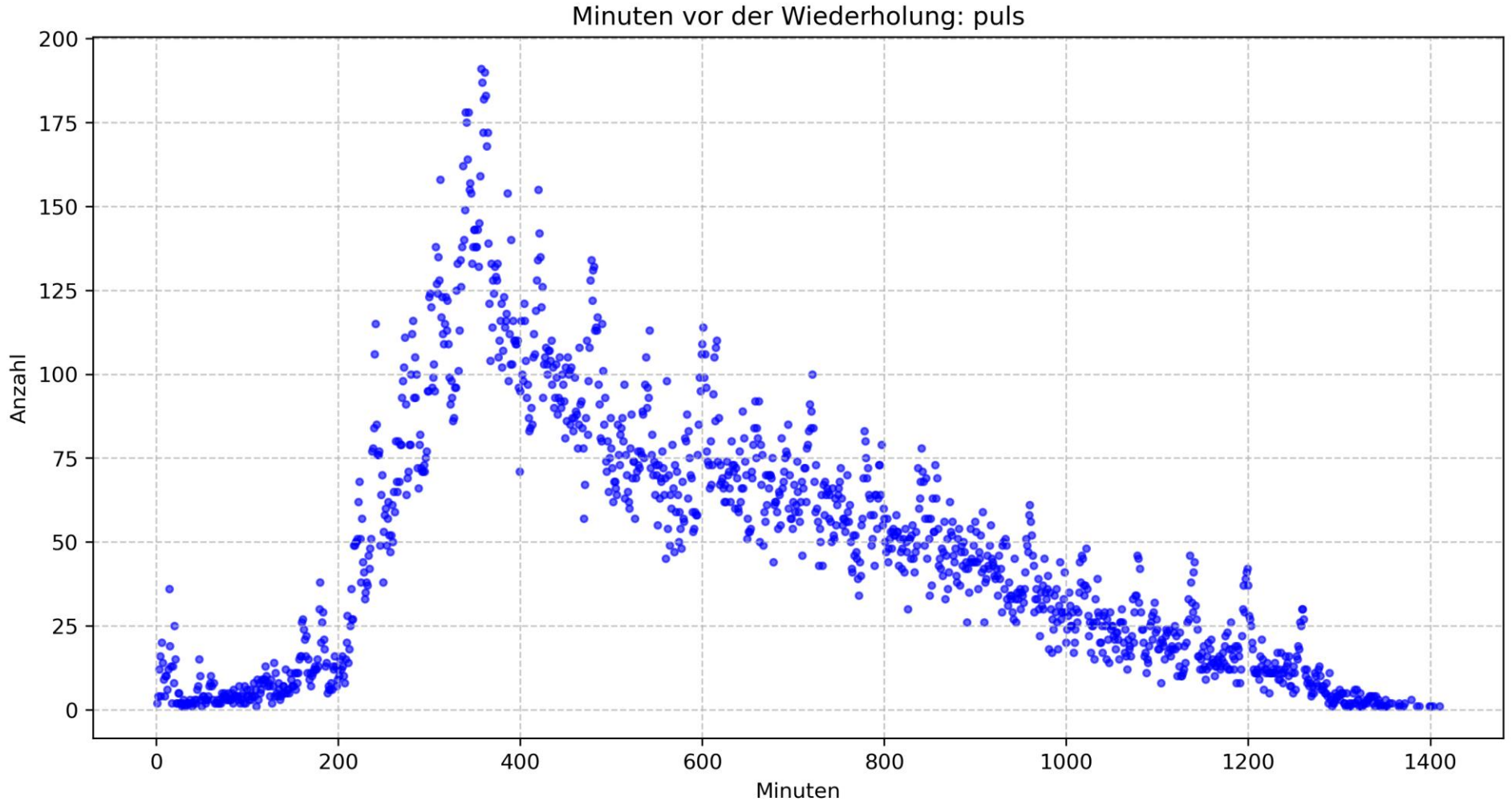
Im Schnitt 958 unterschiedliche Songs / Woche

Radio-aktive Halbwertszeit: Absolut HOT



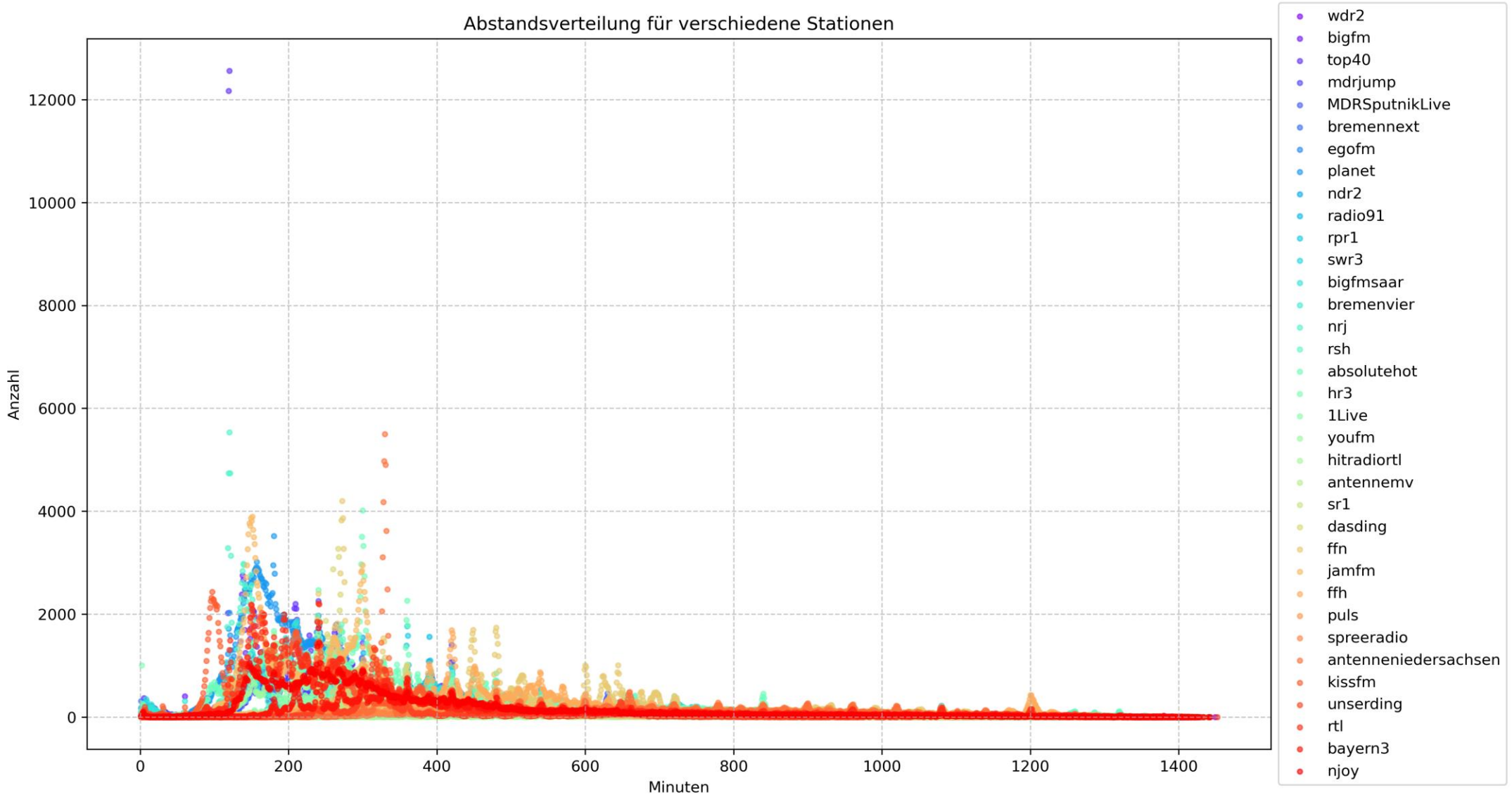
Im Schnitt 386 unterschiedliche Songs / Woche

Radio-aktive Halbwertszeit: PULS

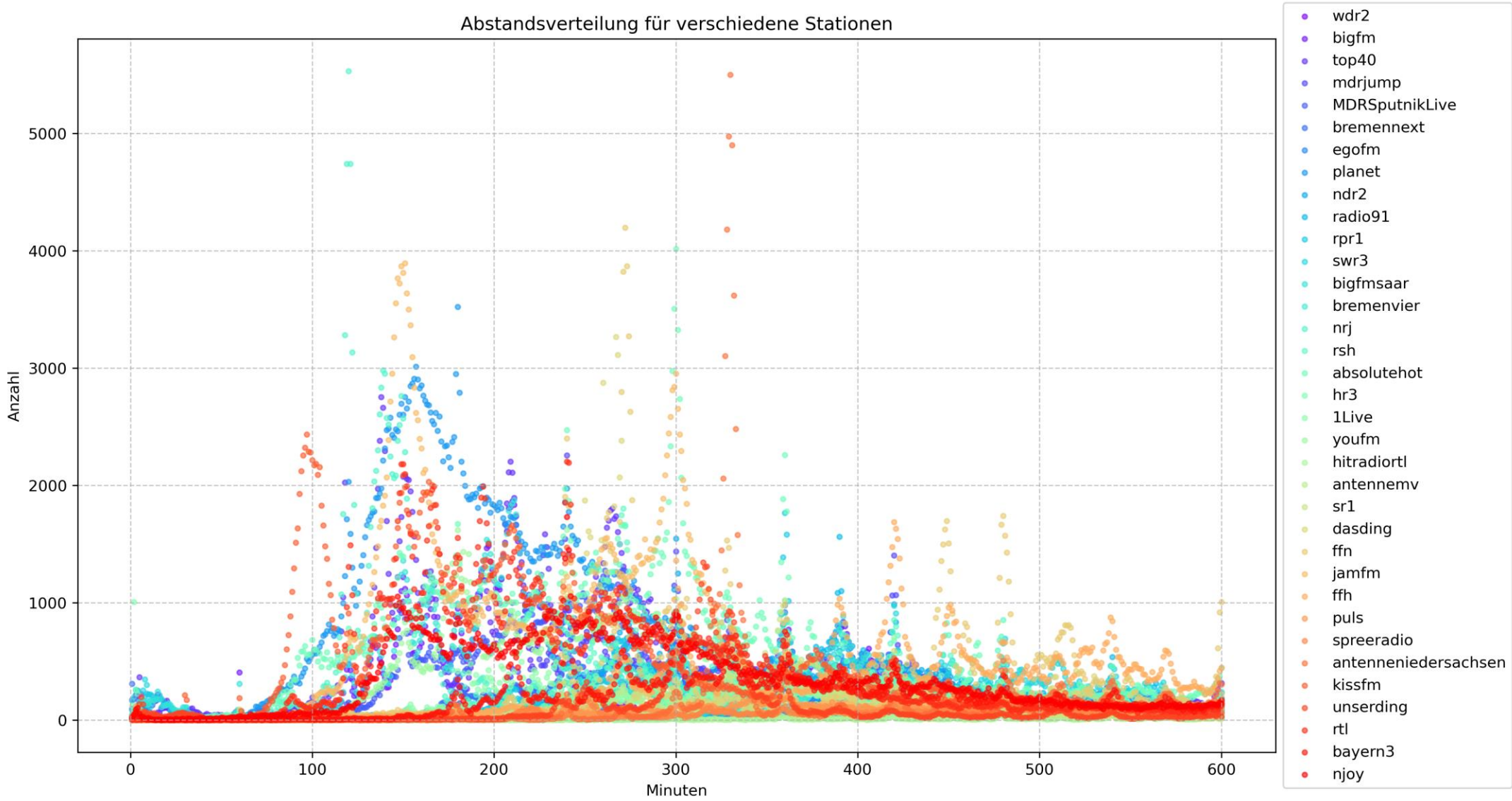


Im Schnitt 2012 unterschiedliche Songs / Woche

Radio-aktive Halbwertszeit: Alle gleichzeitig



Radio-aktive Halbwertszeit: (Zoom)



RadioMining

Exkurs: Darf man das?

Die Radiosender veröffentlichen die Daten ohne Einschränkung oder Passwort

Eine AGB oder anderweitige Einschränkung der Webseiten-Verwendung habe ich nicht gefunden (nur Copyrights für Logos, Cookies, etc)

Ich habe alle Radiosender angefragt. Insbesondere, ob sie noch ältere Daten zur Vervollständigung haben.

Die meisten haben nicht reagiert, aber:

- „haben die Daten nicht“
- „ist unser Betriebsgeheimnis“ (WTF?)
- Haben mir ein Telefoninterview angeboten (Danke Jan Kuhlmann @ NDR2!)
- „wollen die Daten gerne haben, die ich (von ihnen) gesammelt habe“
  – weil sie ihnen selbst fehlen (WTF?)

Vielen Dank!

Weitere Informationen und Folien:

https://anginf.de/

Stefan „starcalc“ Magerstedt

Kontakt: radio@anginf.de

Grafik „RadioMining“ von Thomas „AlSimm0ns“ Hille
Alle anderen Grafiken sind selbsterstellt, Logos oder Screenshots