Rap Recommender

Building A Hybrid Recommendation Engine From Scratch

Owen Curtis DSI-9

Goals

Build Rap Recommender + Execute Learning Agenda + Publish Dataset

The Data

800 Rappers/Artists

4MM+ Words

25,000 Tracks

211 Hours of Audio

17,000 Album Reviews

151,000 Users

The Data

800 Rappers/Artists

MM+ Words

25,00

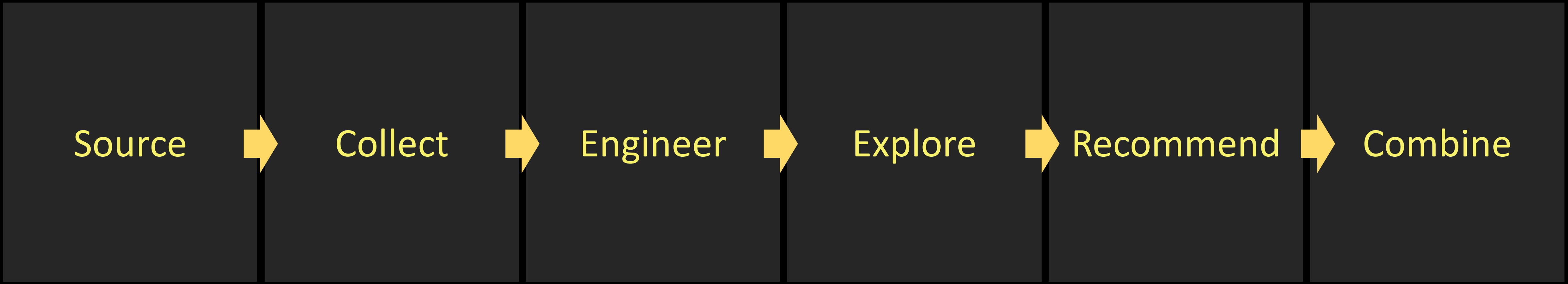
Audio

17,

151,000 Users

80% of tracks fucking profane!

The Process



Source

Collect

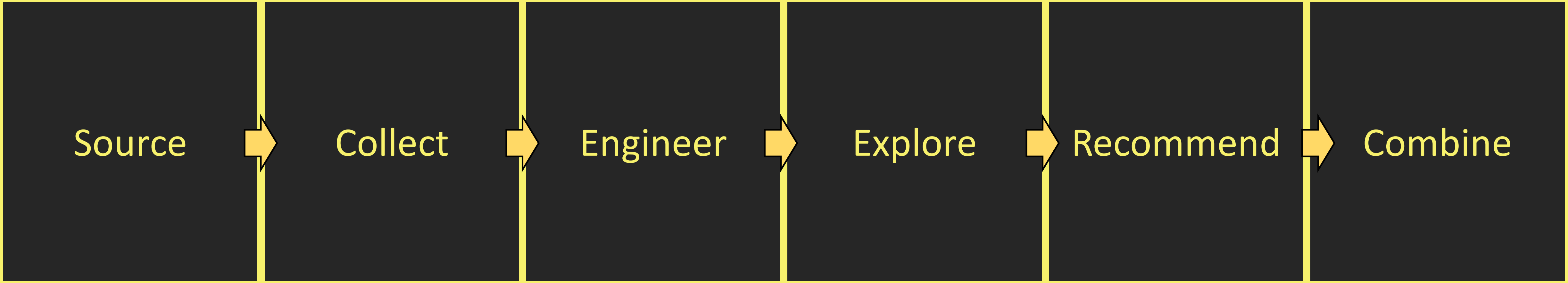
Engineer

Explore

Recommend

Combine

The Learning Agenda



Scraping: Is Selenium an effective library for more bespoke web scraping needs?

Matching: What are best practices for fuzzy matching on similar strings?

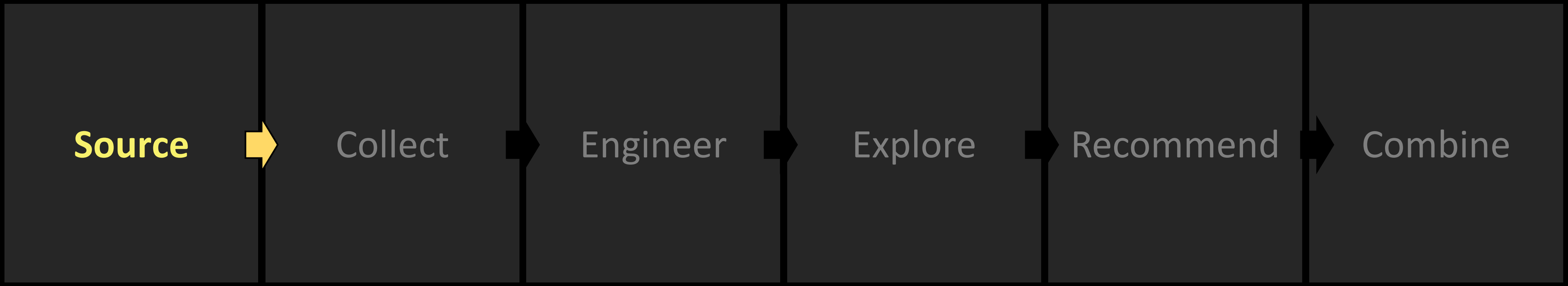
Insights: How has the genre changed? Is hip hop dead as an art form?

Topics: What is the most effective way to extract topics from text data?

Production: Are audio features an effective means of predicting song producers?

Recommender:

- How does the output differ between approaches (collab v content)?
- What are the pitfalls of recommender systems? Which do we encounter?
- What is the appropriate approach for a Hybrid recommender?



Source

Collect

Engineer

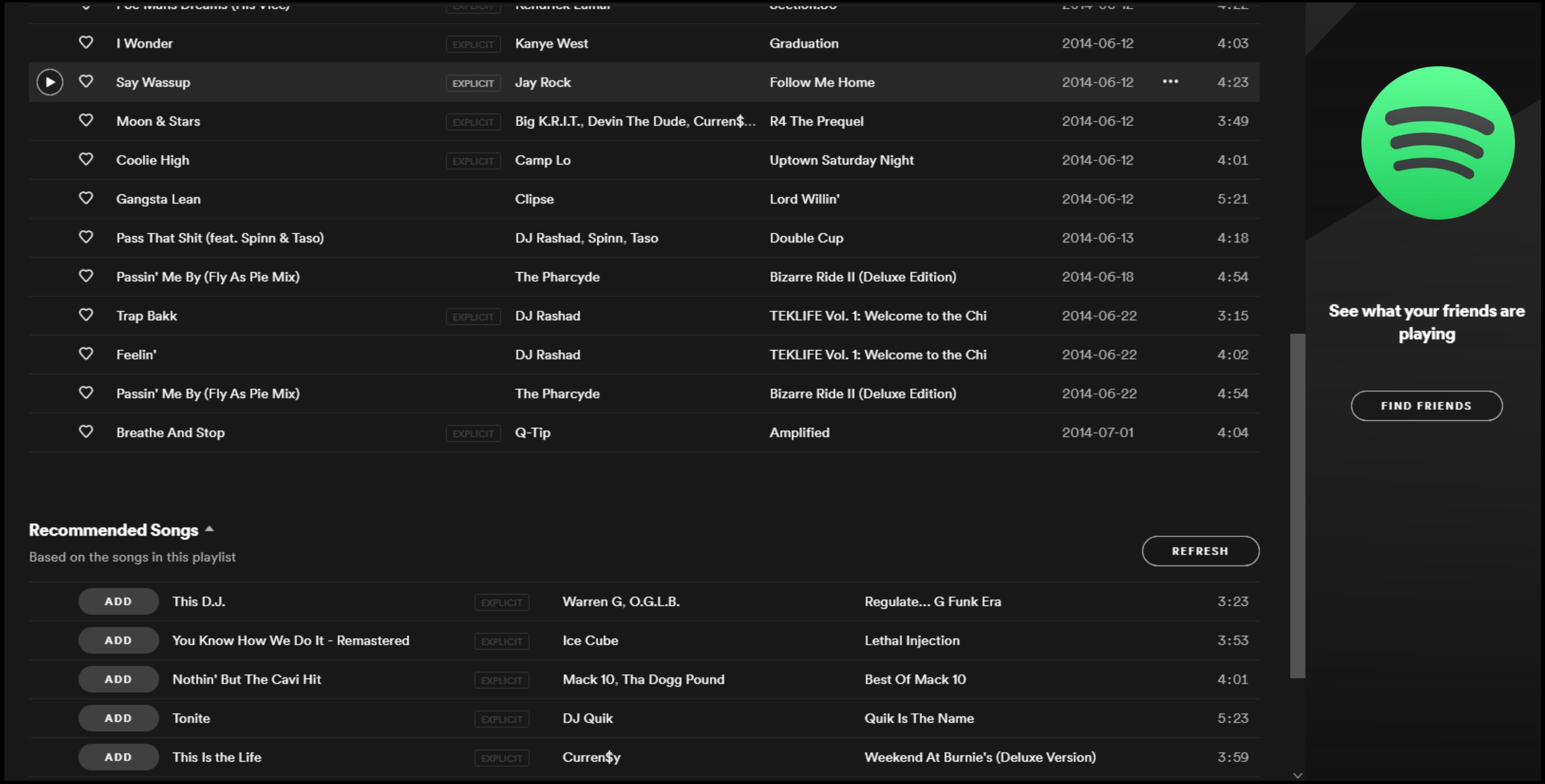
Explore

Recommend

Combine

- Audio
- Reviews
- Lyrics
- User Data

Audio Data



Poe Mans Dreams (His Vice)	EXPLICIT	Kendrick Lamar	Section.80	2014-06-12	4:22
I Wonder	EXPLICIT	Kanye West	Graduation	2014-06-12	4:03
Say Wassup	EXPLICIT	Jay Rock	Follow Me Home	2014-06-12	4:23
Moon & Stars	EXPLICIT	Big K.R.I.T., Devin The Dude, Curren$...	R4 The Prequel	2014-06-12	3:49
Coolie High	EXPLICIT	Camp Lo	Uptown Saturday Night	2014-06-12	4:01
Gangsta Lean		Clipse	Lord Willin'	2014-06-12	5:21
Pass That Shit (feat. Spinn & Taso)		DJ Rashad, Spinn, Taso	Double Cup	2014-06-13	4:18
Passin' Me By (Fly As Pie Mix)		The Pharcyde	Bizarre Ride II (Deluxe Edition)	2014-06-18	4:54
Trap Bakk	EXPLICIT	DJ Rashad	TEKLIFE Vol. 1: Welcome to the Chi	2014-06-22	3:15
Feelin'		DJ Rashad	TEKLIFE Vol. 1: Welcome to the Chi	2014-06-22	4:02
Passin' Me By (Fly As Pie Mix)		The Pharcyde	Bizarre Ride II (Deluxe Edition)	2014-06-22	4:54
Breathe And Stop	EXPLICIT	Q-Tip	Amplified	2014-07-01	4:04

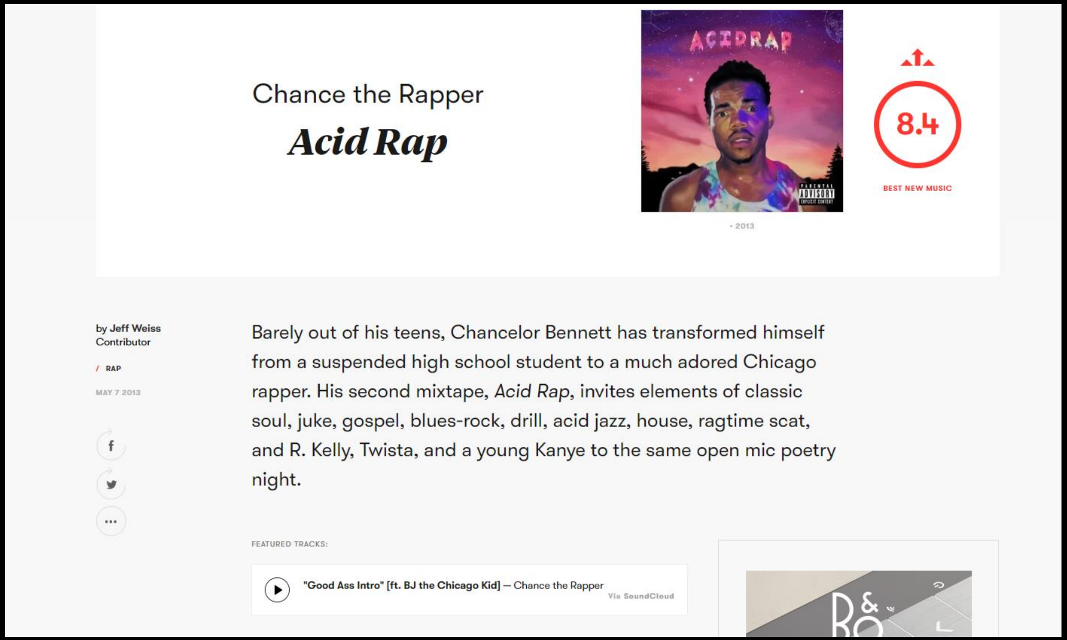
Recommended Songs ▲

Based on the songs in this playlist

ADD	This D.J.	EXPLICIT	Warren G, O.G.L.B.	Regulate... G Funk Era	3:23
ADD	You Know How We Do It - Remastered	EXPLICIT	Ice Cube	Lethal Injection	3:53
ADD	Nothin' But The Cavi Hit	EXPLICIT	Mack 10, Tha Dogg Pound	Best Of Mack 10	4:01
ADD	Tonite	EXPLICIT	DJ Quik	Quik Is The Name	5:23
ADD	This Is the Life	EXPLICIT	Curren$y	Weekend At Burnie's (Deluxe Version)	3:59

See what your friends are playing

Review Data



Chance the Rapper

Acid Rap

• 2013

8.4

BEST NEW MUSIC

by **Jeff Weiss**
Contributor

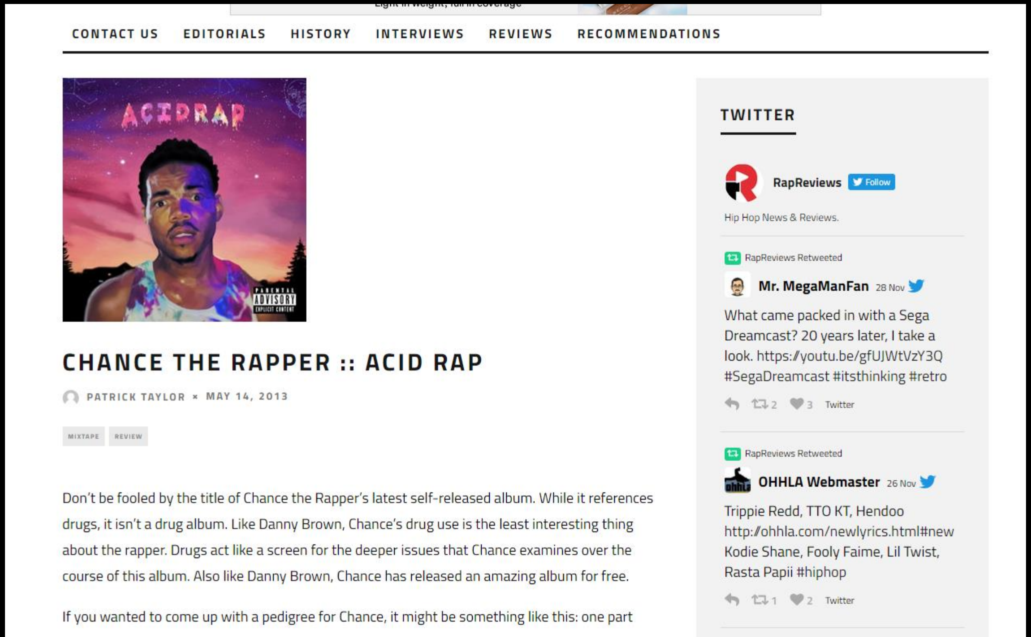
/ RAP

MAY 7 2013

Barely out of his teens, Chancelor Bennett has transformed himself from a suspended high school student to a much adored Chicago rapper. His second mixtape, *Acid Rap*, invites elements of classic soul, juke, gospel, blues-rock, drill, acid jazz, house, ragtime scat, and R. Kelly, Twista, and a young Kanye to the same open mic poetry night.

FEATURED TRACKS:

"Good Ass Intro" [ft. BJ the Chicago Kid] — Chance the Rapper
Via SoundCloud



CHANCE THE RAPPER :: ACID RAP

PATRICK TAYLOR × MAY 14, 2013

MIXTAPE REVIEW

Don't be fooled by the title of Chance the Rapper's latest self-released album. While it references drugs, it isn't a drug album. Like Danny Brown, Chance's drug use is the least interesting thing about the rapper. Drugs act like a screen for the deeper issues that Chance examines over the course of this album. Also like Danny Brown, Chance has released an amazing album for free.

If you wanted to come up with a pedigree for Chance, it might be something like this: one part

TWITTER

RapReviews Follow

Hip Hop News & Reviews.

RapReviews Retweeted

Mr. MegaManFan 28 Nov

What came packed in with a Sega Dreamcast? 20 years later, I take a look. https://youtu.be/gfUJWtVzY3Q #SegaDreamcast #itsthinking #retro

2 3 Twitter

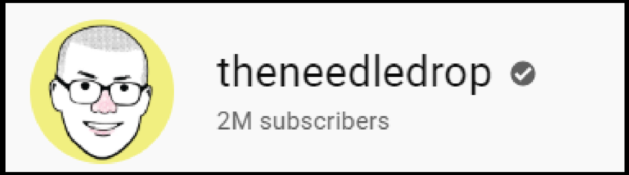
RapReviews Retweeted

OHHLA Webmaster 26 Nov

Trippie Redd, TTO KT, Hendoo http://ohhla.com/newlyrics.html#new Kodie Shane, Fooly Faime, Lil Twist, Rasta Papii #hiphop

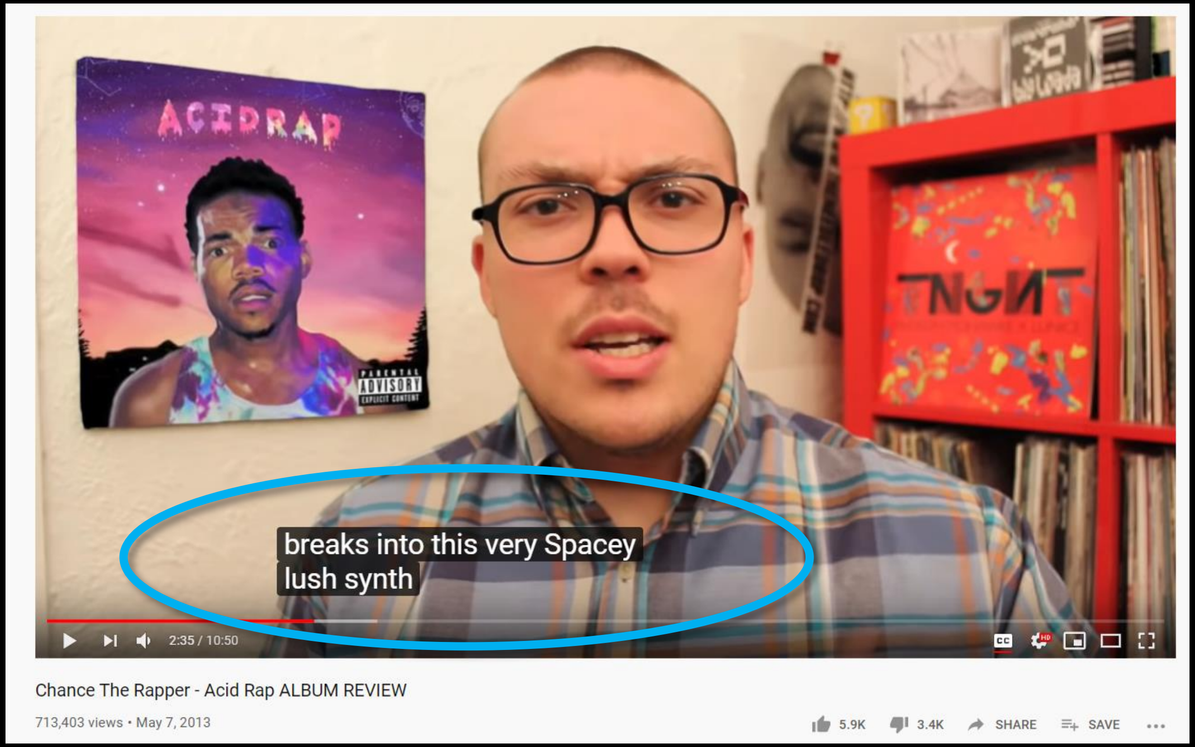
1 2 Twitter

Review Data



theneedledrop

2M subscribers



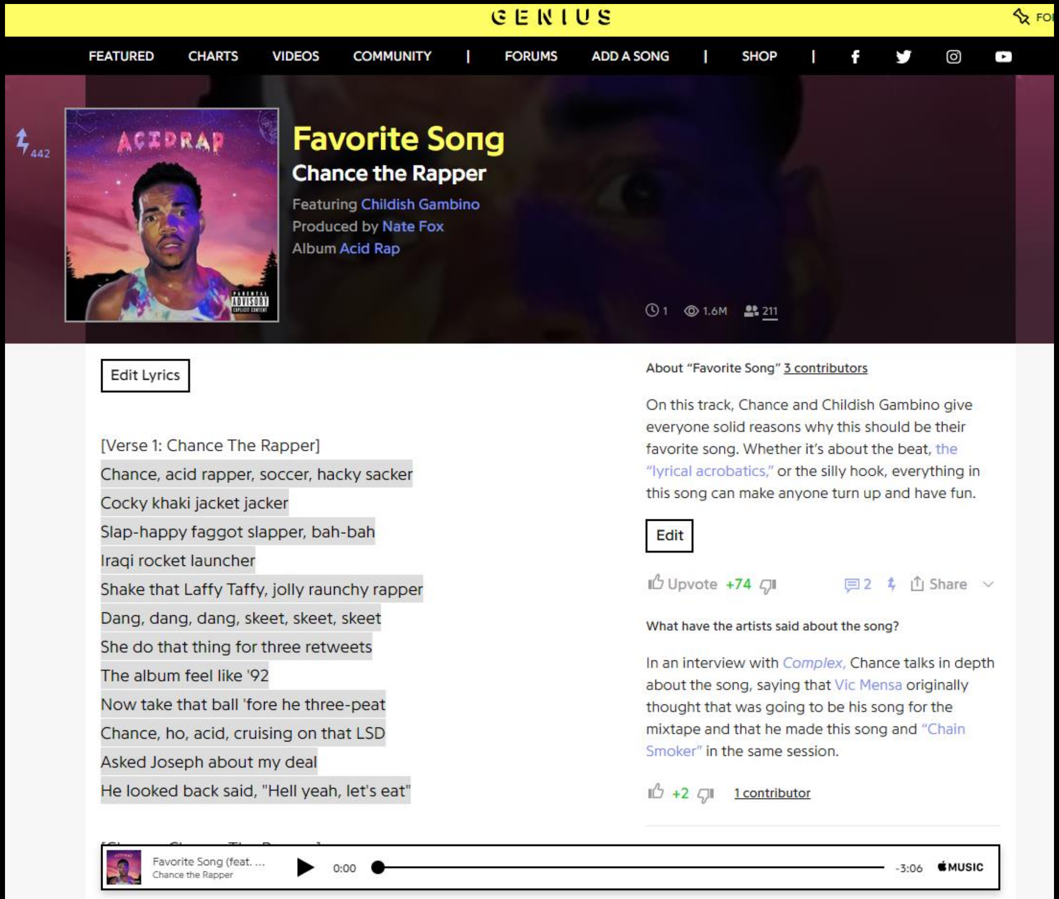
breaks into this very Spacey lush synth

Chance The Rapper - Acid Rap ALBUM REVIEW

713,403 views • May 7, 2013

Lyrics & Users

GENIUS



Favorite Song

Chance the Rapper

Featuring Childish Gambino
Produced by Nate Fox
Album Acid Rap

[Verse 1: Chance The Rapper]
Chance, acid rapper, soccer, hacky sacker
Cocky khaki jacket jacker
Slap-happy faggot slapper, bah-bah
Iraqi rocket launcher
Shake that Laffy Taffy, jolly raunchy rapper
Dang, dang, dang, skeet, skeet, skeet
She do that thing for three retweets
The album feel like '92
Now take that ball 'fore he three-peat
Chance, ho, acid, cruising on that LSD
Asked Joseph about my deal
He looked back said, "Hell yeah, let's eat"

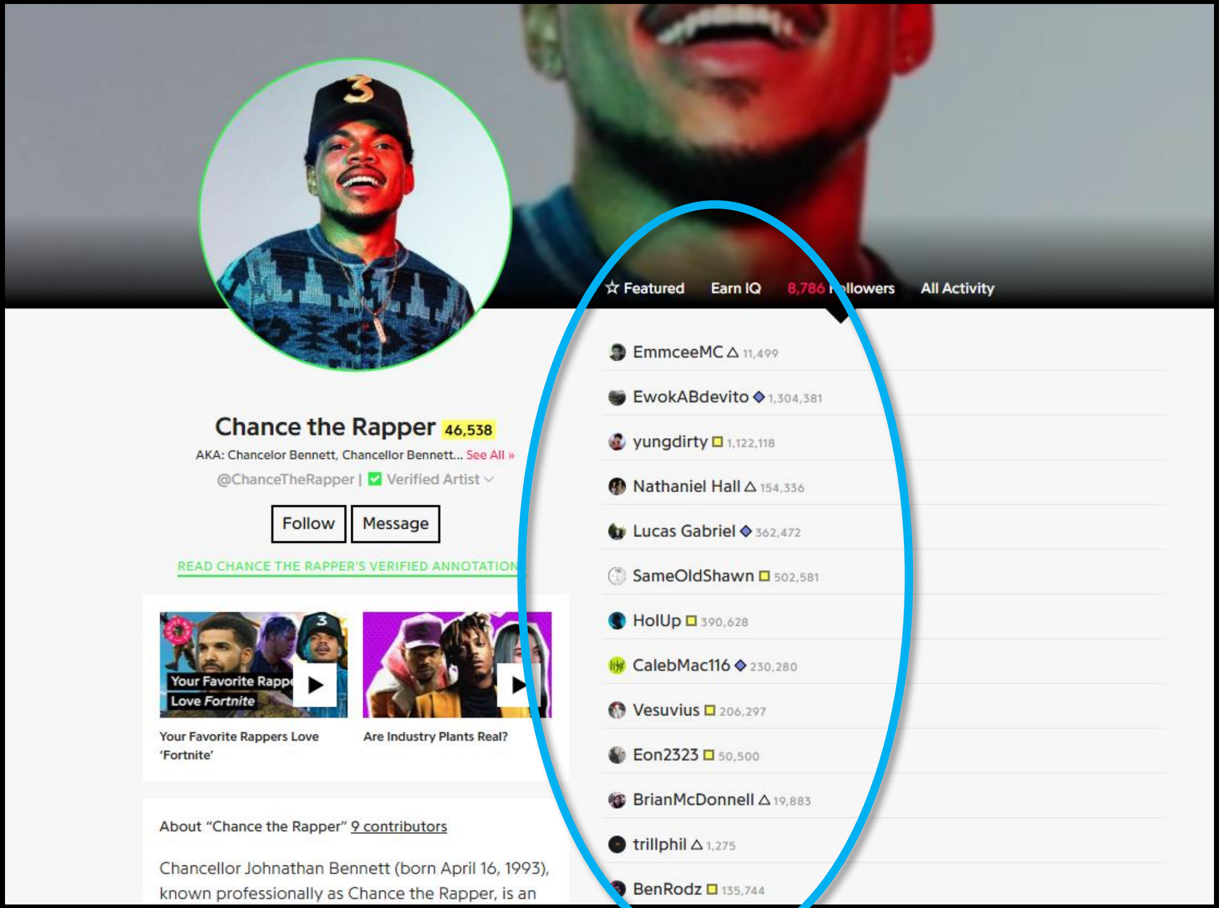
About "Favorite Song" 3 contributors

On this track, Chance and Childish Gambino give everyone solid reasons why this should be their favorite song. Whether it's about the beat, the "lyrical acrobatics," or the silly hook, everything in this song can make anyone turn up and have fun.

What have the artists said about the song?

In an interview with *Complex*, Chance talks in depth about the song, saying that Vic Mensa originally thought that was going to be his song for the mixtape and that he made this song and "Chain Smoker" in the same session.

1 contributor


☆ Featured
Earn IQ
8,786 Followers
All Activity
Follow
Message
READ CHANCE THE RAPPER'S VERIFIED ANNOTATIONS
Your Favorite Rappers Love Fortnite

Chance the Rapper 46,538

AKA: Chancelor Bennett, Chancellor Bennett... See All »

@ChanceTheRapper | ☑ Verified Artist

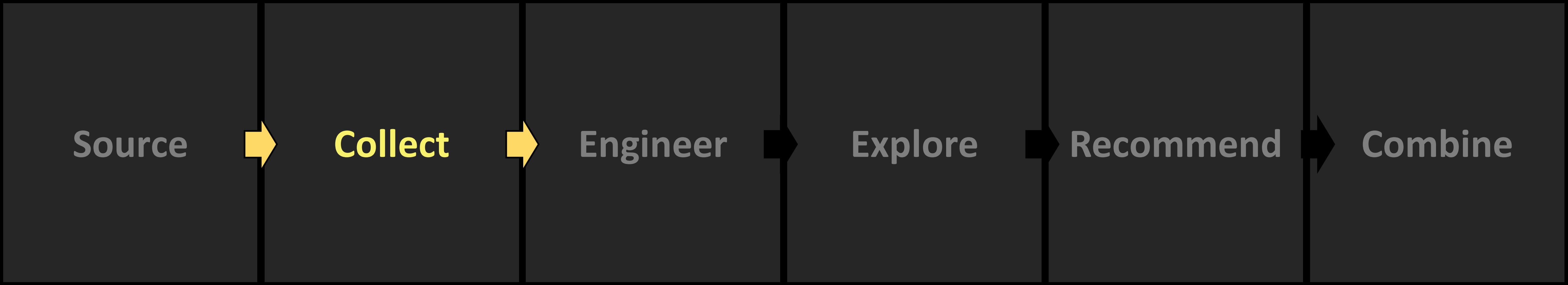
Your Favorite Rappers Love 'Fortnite'

Are Industry Plants Real?

About "Chance the Rapper" 9 contributors

Chancellor Johnathan Bennett (born April 16, 1993), known professionally as Chance the Rapper, is an

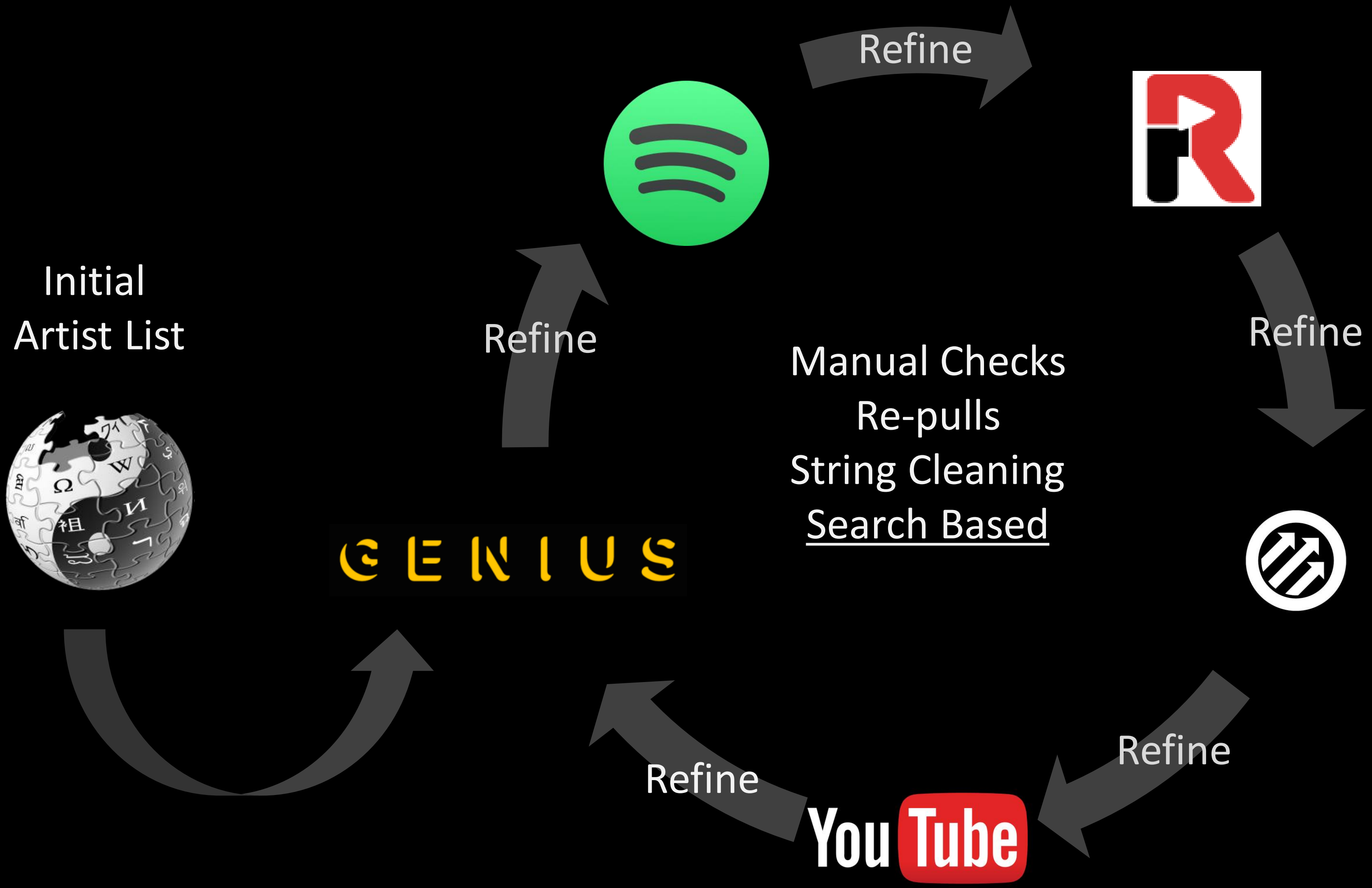
EmmceeMC △	11,499
EwokABdevito ◆	1,304,381
yungdirty □	1,122,118
Nathaniel Hall △	154,336
Lucas Gabriel ◆	362,472
SameOldShawn □	502,581
HolUp □	390,628
CalebMac116 ◆	230,280
Vesuvius □	206,297
Eon2323 □	50,500
BrianMcDonnell △	19,883
trillphil △	1,275
BenRodz □	135,744



Source → Collect → Engineer → Explore → Recommend → Combine

- Querying/Scraping
- Challenges
- Cleaning
- Fuzzy Matching

Collect: An Iterative Process


Initial
Artist List

Manual Checks
Re-pulls
String Cleaning
Search Based

Collect: Approach

Audio

- Pre-existing API python wrappers

Reviews

- BeautifulSoup

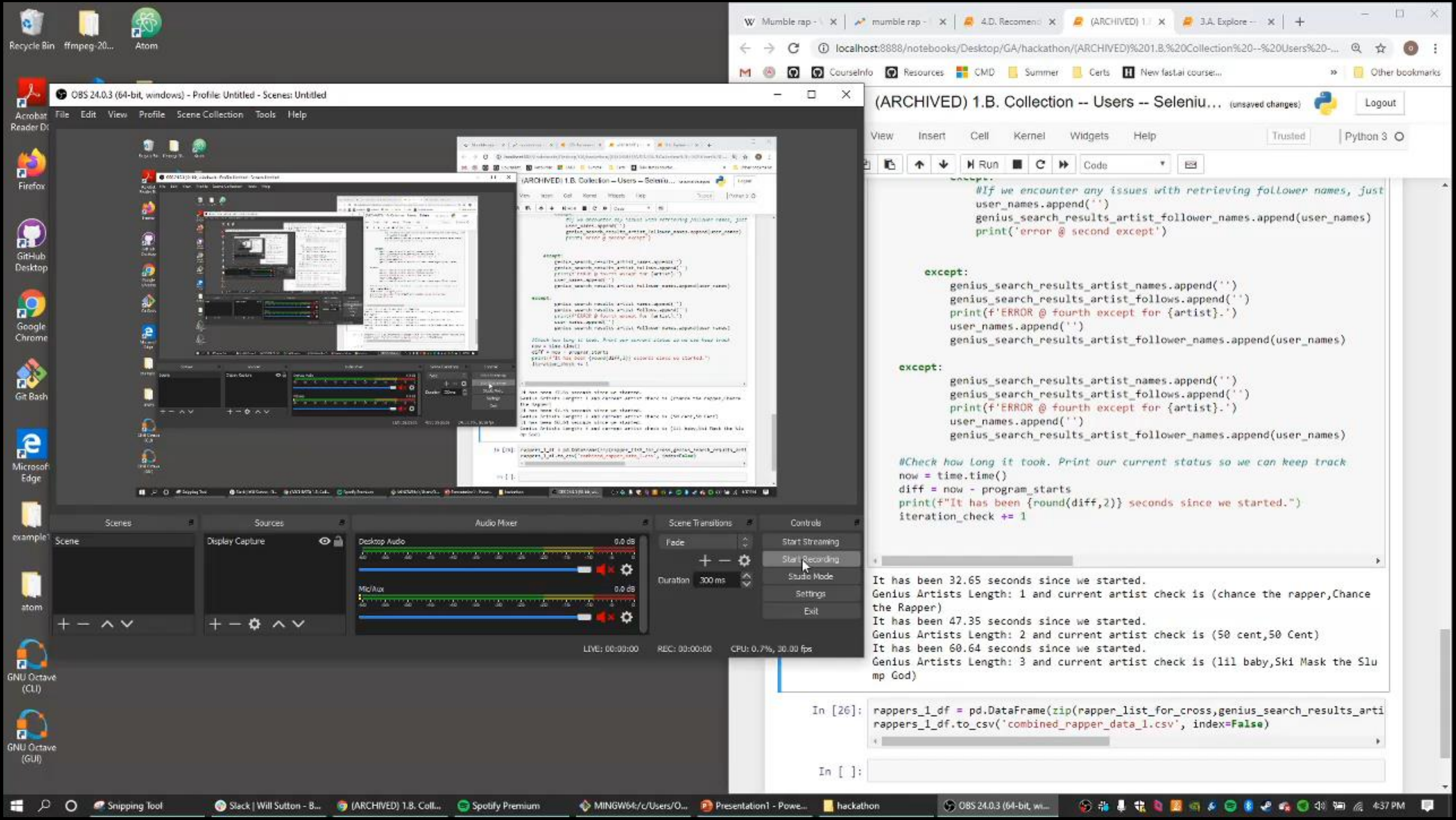
- Kaggle / Google Datasets
- Pre-existing API python wrappers

Lyrics & Users

- **Lyrics:** Pre-existing API python wrappers

- **Users:** a bit more complicated…
  - No API Endpoint
  - Infinite Scroll
  - Large scale scrape

Collect: User Data



```
        except:
            #If we encounter any issues with retrieving follower names, just
            user_names.append('')
            genius_search_results_artist_follower_names.append(user_names)
            print('error @ second except')


    except:
        genius_search_results_artist_names.append('')
        genius_search_results_artist_follows.append('')
        print(f'ERROR @ fourth except for {artist}.')
        user_names.append('')
        genius_search_results_artist_follower_names.append(user_names)

except:
        genius_search_results_artist_names.append('')
        genius_search_results_artist_follows.append('')
        print(f'ERROR @ fourth except for {artist}.')
        user_names.append('')
        genius_search_results_artist_follower_names.append(user_names)

#Check how long it took. Print our current status so we can keep track
now = time.time()
diff = now - program_starts
print(f"It has been {round(diff,2)} seconds since we started.")
iteration_check += 1
```

```
It has been 32.65 seconds since we started.
Genius Artists Length: 1 and current artist check is (chance the rapper,Chance
the Rapper)
It has been 47.35 seconds since we started.
Genius Artists Length: 2 and current artist check is (50 cent,50 Cent)
It has been 60.64 seconds since we started.
Genius Artists Length: 3 and current artist check is (lil baby,Ski Mask the Slu
mp God)
```

```
In [26]: rappers_1_df = pd.DataFrame(zip(rapper_list_for_cross,genius_search_results_arti
         rappers_1_df.to_csv('combined_rapper_data_1.csv', index=False)

In [ ]:
```

Collect: User Data

Question: Is Selenium an effective tool for web scraping?

Se

Pros

- High degree of customization
- If it’s in the HTML, you can get it
- Safeguards in place for interruptions

Cons

- Scalability/Speed
- Ads / Scripts

VS

Requests

Pros

- Well structured data
- Speed/Scalability
- No issues with ads/scripts

Cons

- Limited to data provided

Learning

- Selenium is an excellent tool for things like automated site testing
- An effective tool for scraping when there are no other options (ex form filling, “clicking”)

We have our data. Let’s check it out!!!!

Source	Artists	Album	Track
	A$AP Rocky ft. Black Hippie	Long Live A$AP	U.E.N.O

months ago ap rocky career seemed mired purgatory year old harlem spectacular snagging still crazy million deal rca based early buzz silencing skeptics breakout livelovea ap mixtape throughout longlivea ap repeatedly failed materialize slated july th september th october eventually sometime december final pushback happened dead zone new year began feel like rca attempting quietly jettison investment corporate hull longlivea ap delivers even exceeds promise livelovea ap like mixtape album triumph craft curation preserving rocky immaculate taste smartly upgrading sound third record remains close style livelovea ap collaborations follow steps last

sp bir ega harlem

ro g hu song rhyme

patterns diplomats obvious primary color reaching back another harlemite deceptively catchy flow agreeable air penchant colorful clothing find another kindred spirit e like e rocky tucks steely edges behind goofy smile

ur op si so ne ar city

so bea err li nce blowing

ye asap atr ed mpletely

er uest cu vi ocky pretty

m arisi pla ad king

w rap y f rd duction

believe people say stuff like like yeah course asap rocky dependent upon production beats chooses mean beats chooses makes projects worthwhile however post post beautiful dark twisted fantasy land mc dependent upon production yeah guys like action bronson big krit danny brown kendrick lamar well yay praised things say messages mean certainly contribute personalities also unique beats either produce choose mean instrumentals choose give musical edge makes stand mcs true asap rocky invent sound yes asap rocky incredibly influenced styles kind amalgamation everything east coast style boom bap southern houston stuff chopped screwed well darker atmospheric rappers space ghost purp people seem really angry asap gotten way famous game style know first guy planet kind get influenced growing style make huge record true asap rocky sounds like asa brack asap rocky name sonically resemble asap rock care think hilarious people actually think try grab attention asap rock fans even really good marketing scheme considering one rock makes entirely different kind think lot crossover fandom going really reason compare even think one others name mentioned asinine way well aware asap rocky flaws sure come around bite butt mean think saw dropped asap mob mixtape little back last year many weaknesses piled top one another listening crap honestly pretty surprised impressed came album one track another asap rocky makes right beat choices gets good producers good writers current features writes co writes right hooks top maintains style still coming album feels bit commercial nice radio friendly singles sure people may even interpret album bit improvement live love asap variety detailed production overall kind expensive sort new improved revamped feel get wrong think live love asap great tape consistently slow hazy feel makes pretty potent listen beginning end asap rocky album instead brings variety little something everybody something different mood almost every track even though yeah really step lyrically anything like think proves pretty versatile versatile think people assume could album opens bass rumbling title track disjointed imagery thrown together cash registers ringing guns cocking cocaine horses kind got gangsta vibe also posed sort weird reverb artsy way like lot love spacious falsetto vocals course soar beat track switches verse nice transition one prettiest moments album track quite producers credited name really shows track goldy comes shortly pretty much like haters anthem hilarious one saying people talking shit get lockjaw asap rocky throws corny lines like think shit tell stink another line saying think king ask coretta scott honestly serious kind semi joke yeah asap rocky put forth really cool personality composed much time cocky end day album fun appealing asap rocky manages still really kind cavernous sleek sort dark vibe right track pmw comes course stands pussy money weed expect song know another extremely charismatic stunt anthem nice atmospheric synthesizers lot clickety clack rhythms schoolboy q comes nice feature moving ahead actually enjoy song fucking problems quite bit drake track well chainz hook kendrick lamar follows back actually think kendrick weakest moment bad lyricist anything like feel like party track especially like jiwan feels slightly place much fitting song train little bit later huge long posse track really lyrical really hard hitting love epic cinematic beat thing mentioned song previous video still enamored huge names thing like action bronson danny brown big krit yellow wolf disappoint asap rocky goes kind weird suicidal song phoenix gets kind personal history growing fan closing track thing also dubstep reggae influenced wild night actually features skrillex production believe kind flip burdine name production french dj outfit definitely check album surprising left turns messages cohesive narratives say though album low points one asap rocky singing toward end phoenix yuck also song pain really one nondescript beats entire lp course asap rocky lyricism really make song feel distinct either also album features santee gould decent verses know interesting personal quirks asap rocky saying used rugged boots getting tailor suits santee gould hook flat lifeless even feel like trying put emotion vocals fashion killer incredibly saccharine sickening love song seems like asap rocky woman interested clothes wearing really kind making fall mean know could throw clothes mannequin maybe would look attractive point know song really bonus track could gotten switched ghetto symphony would nice ironically track credits way writers songs entire lp oh still say overall enjoyed listening album lot mean thought really really good commercial ever asap rocky came head held high bass rumbling atmosphere cockiness mean know really needed change much make commercial album live love asap really kind plays lot hedonistic ideas commercial usually still say asap rock angle style sort ushered atmospheric production mainstay moment going stop ranting going stop ranting feeling strong like thing given listen think loved review next fantana asap rocky forever

Oh….Oh god….

Source	Artists	Album	Track
	A$AP Rocky ft. Black Hippie	Long Live A$AP	U.E.N.O
	A$AP Rocky	Long Live A$AP	U.E.N.O. (ft. Jay Rock, Ab Soul, Schoolboy Q)
GENIUS	A$AP Rocky ft Jay Rock, Ab Soul, Schoolboy Q	Long.live.A$AP	U.E.N.O.
R	A$AP Rocky	Â°Ã¬âLongÂ°Ã¬âliveÂ°Ã¬âA$ AP	

months ago ap rocky career seemed mired purgatory year old harlem spectacular snagging still crazy million deal rca based early buzz silencing skeptics breakout livelovea ap mixtape throughout longlivea ap repeatedly failed materialize slated july th september th october eventually sometime december final pushback happened dead zone new year began feel like rca attempting quietly jettison investment corporate hull longlivea ap delivers even exceeds promise livelovea ap like mixtape album triumph craft curation preserving rocky immaculate taste smartly upgrading sound third record remains close style livelovea ap collaborations follow steps last

patterns diplomats obvious primary color reaching back another harlemite deceptively catchy flow agreeable air benchant colorful clothing find another kindred spirit e like e rocky tucks steely edges behind goofy smile

believe people say stuff like like yeah course asap rocky dependent upon production beats chooses mean beats chooses makes projects worthwhile however post post beautiful dark twisted fantasy land mc dependent upon

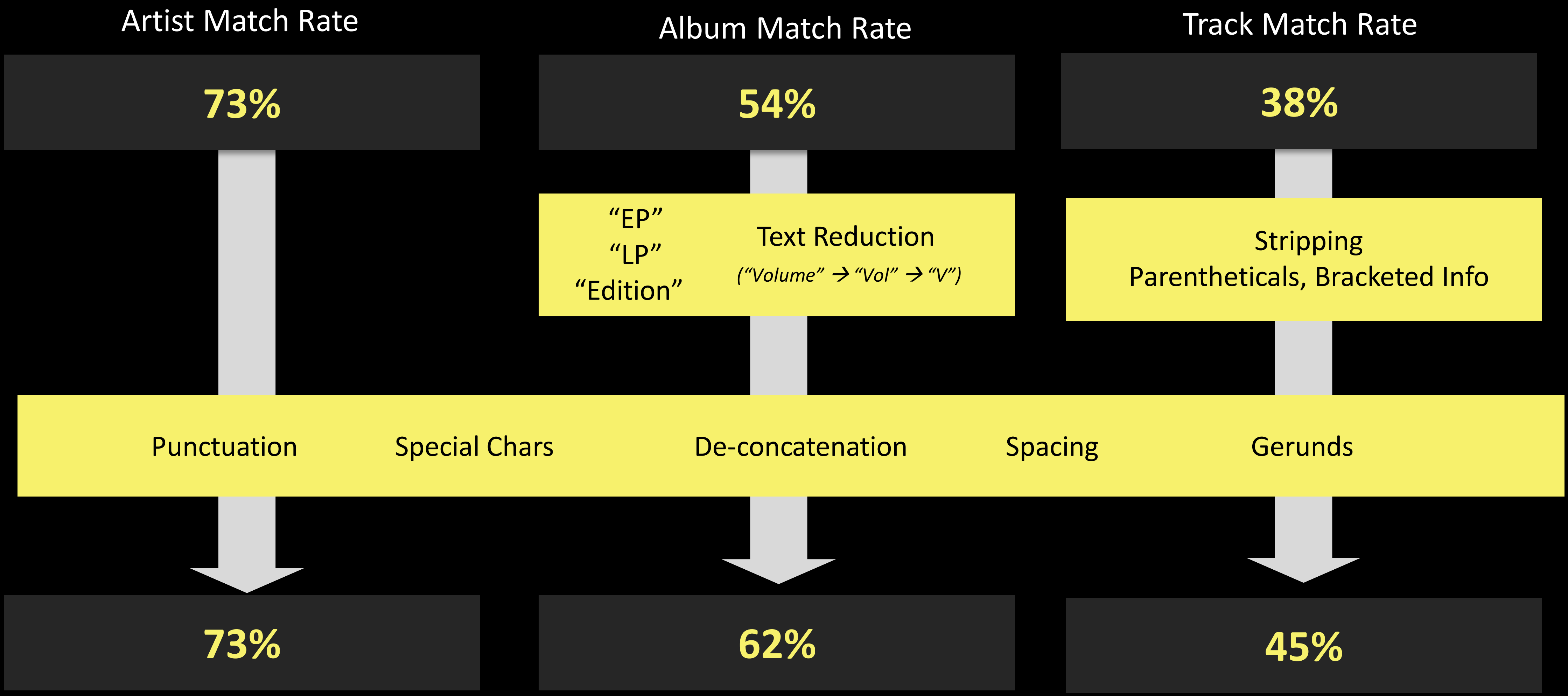
writes right hooks top maintains style still coming album feels bit commercial nice radio friendly singles sure people may even interpret album bit improvement live love asap variety detailed production overall kind expensive

track especially like jiwan feels slightly place much fitting song train little bit later huge long posse track really lyrical really hard hitting love epic cinematic beat thing mentioned song previous video still enamored huge

needed change much make commercial album live love asap really kind plays lot hedonistic ideas commercial usually still say asap rock angle style sort ushered atmospheric production mainstay moment going stop ranting going stop ranting feeling strong like thing given listen think loved review next fantana asap rocky forever

Time to switch projects?

Collect: Preprocessing



Artist Match Rate

73%

Album Match Rate

54%

Track Match Rate

38%

“EP”
“LP”
“Edition”

Text Reduction
(“Volume” → “Vol” → “V”)

Stripping
Parentheticals, Bracketed Info

Punctuation

Special Chars

De-concatenation

Spacing

Gerunds

73%

62%

45%

Collect: Cleaning

Missing Data

7,000 Dates	1,000 Albums	6,000 Producers
Imputed with mean	*Imputed where possible*	*TBD (See Next Section)*

Collect: Fuzzy Matching?

Matching: What are best practices for fuzzy matching on similar strings?

Jaro-Winkler Distance

Hamming Distance

Dice Distance

Levenshtein Distance

q - gram Distance

Levenshtein distance (or edit distance) between two strings is the number of deletions, insertions, or substitutions required to transform source string into target string.

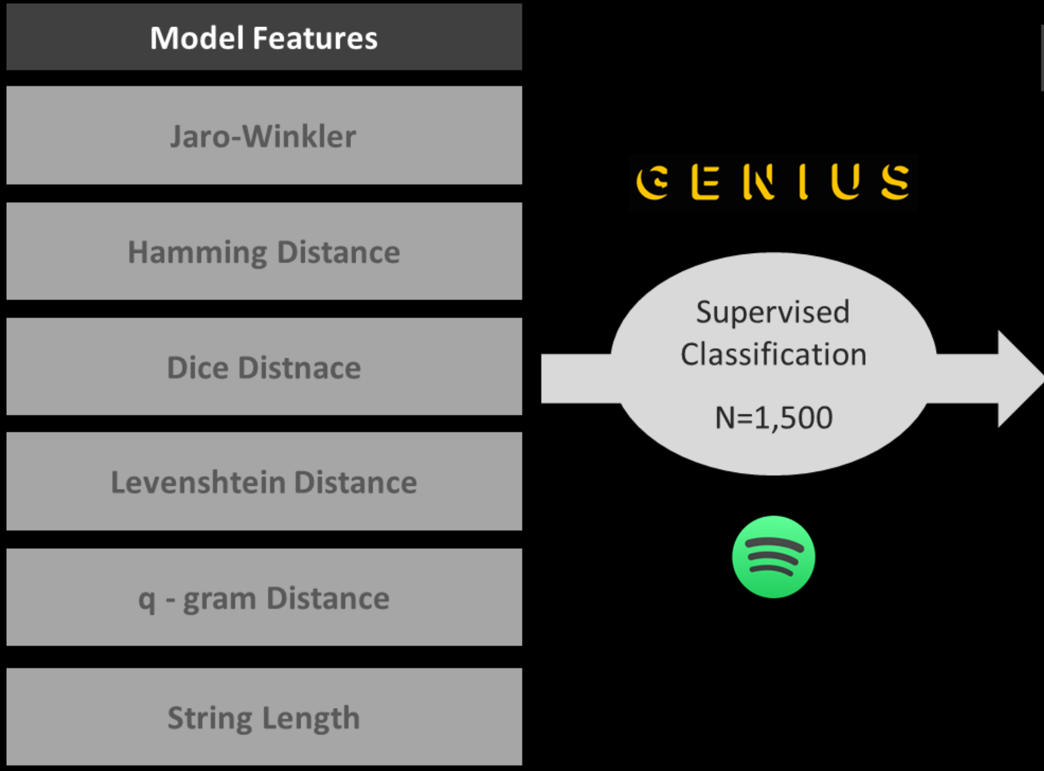
A$AP Rockey | U.E.N.O

ASAP Rocky | U.E.N.O.

Distance: 3

Collect: Fuzzy Matching?

Matching: What are best practices for fuzzy matching on similar strings?

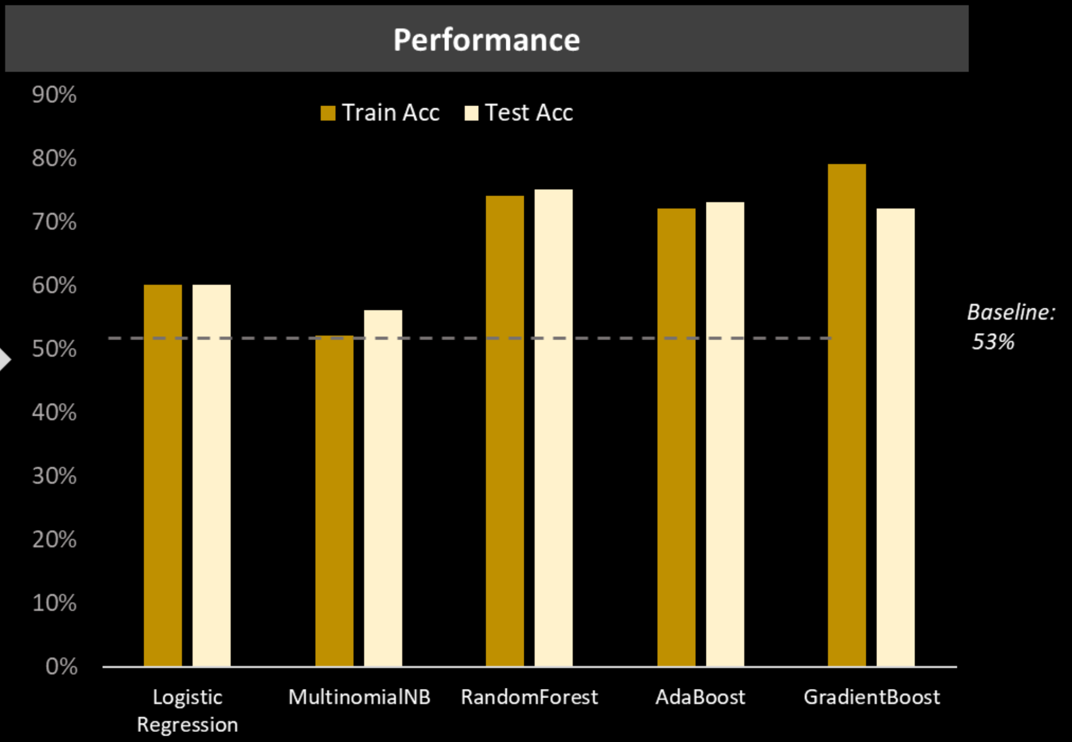


Model Features

- Jaro-Winkler
- Hamming Distance
- Dice Distnace
- Levenshtein Distance
- q - gram Distance
- String Length

Supervised Classification

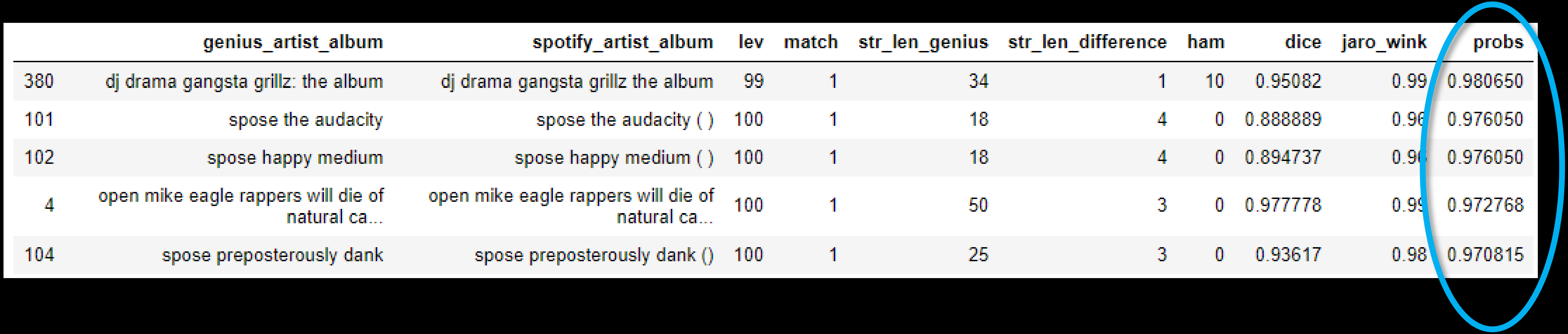
N=1,500



Performance

Collect: Fuzzy Matching?

Matching: What are best practices for fuzzy matching on similar strings?



	genius_artist_album	spotify_artist_album	lev	match	str_len_genius	str_len_difference	ham	dice	jaro_wink	probs
380	dj drama gangsta grillz: the album	dj drama gangsta grillz the album	99	1	34	1	10	0.95082	0.99	0.980650
101	spose the audacity	spose the audacity ( )	100	1	18	4	0	0.888889	0.96	0.976050
102	spose happy medium	spose happy medium ( )	100	1	18	4	0	0.894737	0.96	0.976050
4	open mike eagle rappers will die of natural ca...	open mike eagle rappers will die of natural ca...	100	1	50	3	0	0.977778	0.99	0.972768
104	spose preposterously dank	spose preposterously dank ()	100	1	25	3	0	0.93617	0.98	0.970815

Learning

- There are several distance metrics which can be leveraged for text processing; selection context dependent
- These metrics can be fed in as features for supervised classification; high sample recommended

Collect: Fuzzy Matching?

Matching: What are best practices for fuzzy matching on similar strings?

Probable Matches

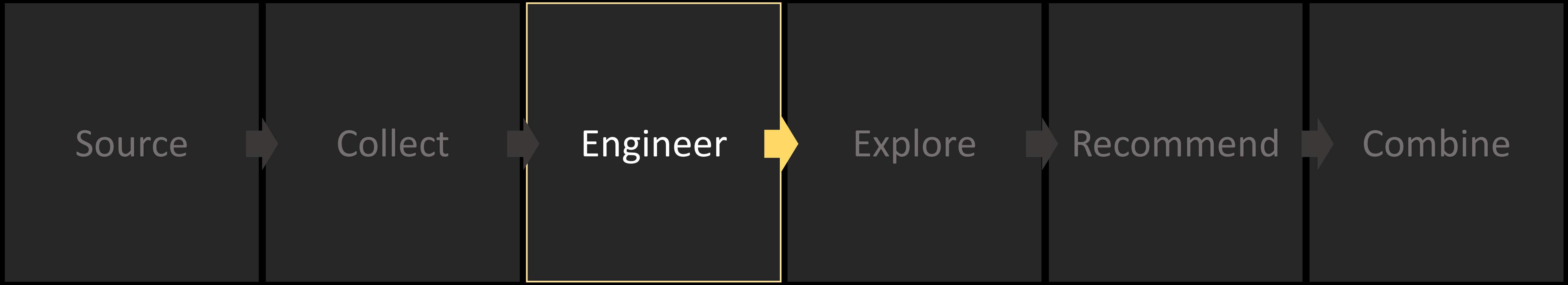
+

amazon mechanical turk™
Artificial Artificial Intelligence

Crowdsourced Validation

Learning

- There are several distance metrics which can be leveraged for text processing; selection context dependent
- These metrics can be fed in as features for supervised classification; high sample recommended



Source

Collect

Engineer

Explore

Recommend

Combine

- Reviews
- Audio
- Lyrics

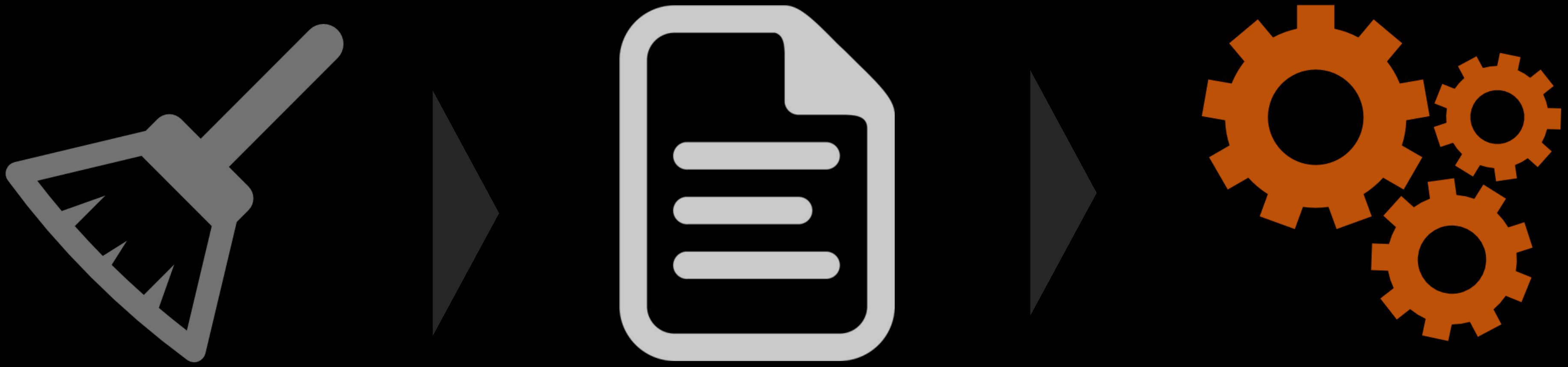
RAPREVIEWS

YouTube

Engineer: Reviews

“hey guys Anthony Fantana here Internet's busiest music nerd this review is a little late but you know it's because I spent all night last night compiling my favorite music videos EPS album's tracks singles I think I'm going to have a pretty good set of videos coming up toward the end of this year just a best-of thing alright we have a long long review this time around and it's because you know what we have to rap about rap there's a growing sentiment in some of my videos that I don't like rap that I don't like hip-hop and of course this foul stench is emanating mostly from my kanye and my big boy reviews people get hung up on the scores don't actually listen to what I say what can you expect but if you did hear what I said in those reviews you would know that some of my favorite songs of this year have come from those albums nevermind the fact that most people gloss over my roots review and tie early hip hop has really been kind of weird for me this year a lot of the major label releases the big ones have been somewhat decent for the most part but it's not like there haven't been hyped artists that I didn't really care for like all the weed wrap on currencies new LPS kind of makes it boring for me and big crits debut was alright and I have no clue why some of my videos are getting waka flocka flame requests there have been indie label releases that I liked for example gangrene or strong arms steady but for the most part Stones Throw rhyme Sayers Def Jux these indie hip hop labels have not been bringing the heat they haven't been bringing the competition to the majors….”

Engineer: Reviews



Pre-Processing

Combined Review Text

Tfidf Vectorization

Engineer: Audio

Spotify

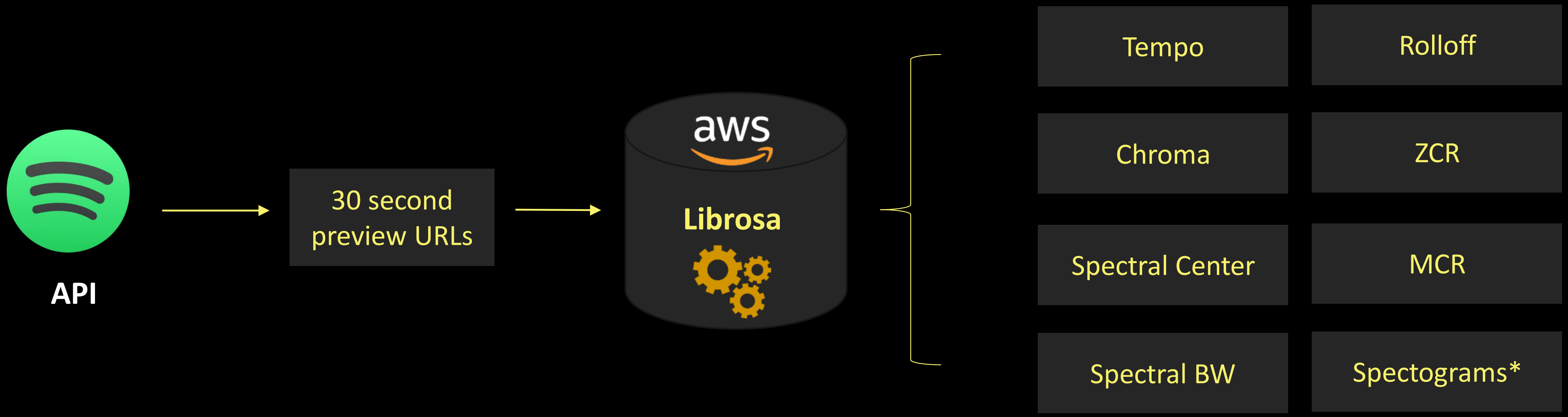
Custom

Engineer: Audio

Spotify

Acousticness	Danceability	Energy
Instrumentalness	Liveness	Loudness
Speechiness	Valence	Tempo

Engineer: Audio



*https://towardsdatascience.com/music-genre-classification-with-python-c714d032f0d8

Engineer: Audio

Let’s look at a few examples…

	Tempo	Chroma	Zero Crossing Rate
Description	Beats per Minute	Representation of Musical signals in terms of octaves.	Rate at which music signal changes from positive to negative, or vice versa
Use Case	High Pace, Low Pace	Identifying the tones of a song	Percussion

*https://towardsdatascience.com/music-genre-classification-with-python-c714d032f0d8

Engineer: Audio

We're still missing producer data on ~25% of tracks.

Can we use our audio features to identify them?

*https://towardsdatascience.com/music-genre-classification-with-python-c714d032f0d8

Engineer: Audio

Question: Are audio features an effective means of predicting the track producer?

Neural Network Setup

Input Layer (16)

Hidden Layer (32, L2 Reg = 0.1)

Hidden Layer (32, L2 Reg = 0.1)

Output (‘Softmax’)

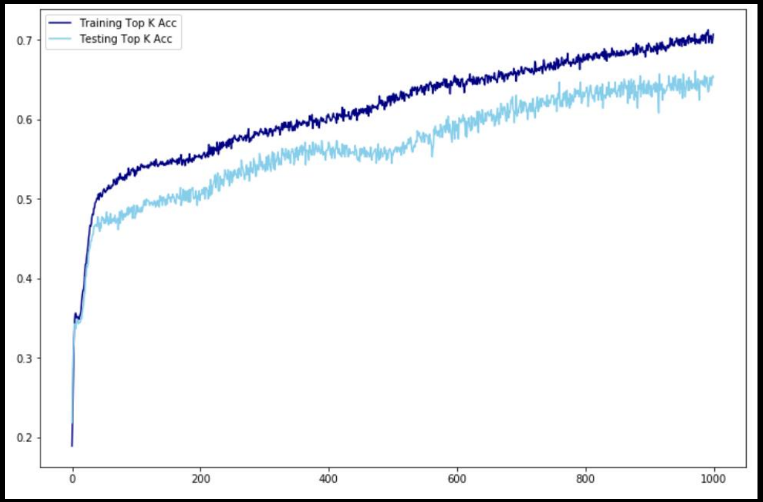
Epochs: 1000, Batch size: 128
N = 3,000

Accuracy

Train	Test
~33%	~28%

Baseline
4%

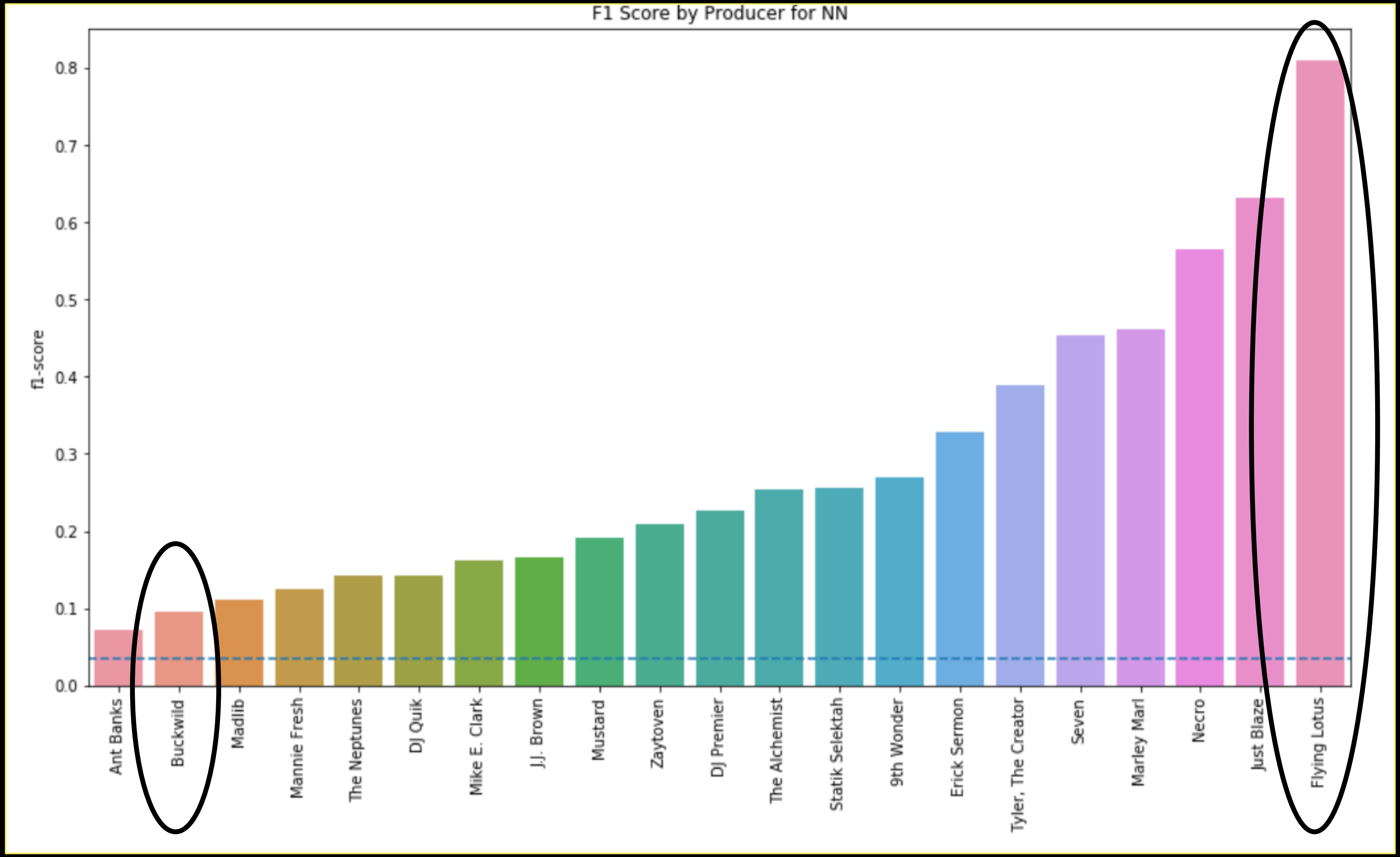
Top 5 Accuracy



*https://towardsdatascience.com/music-genre-classification-with-python-c714d032f0d8

Engineer: Audio

Question: Are audio features an effective means of predicting the track producer?



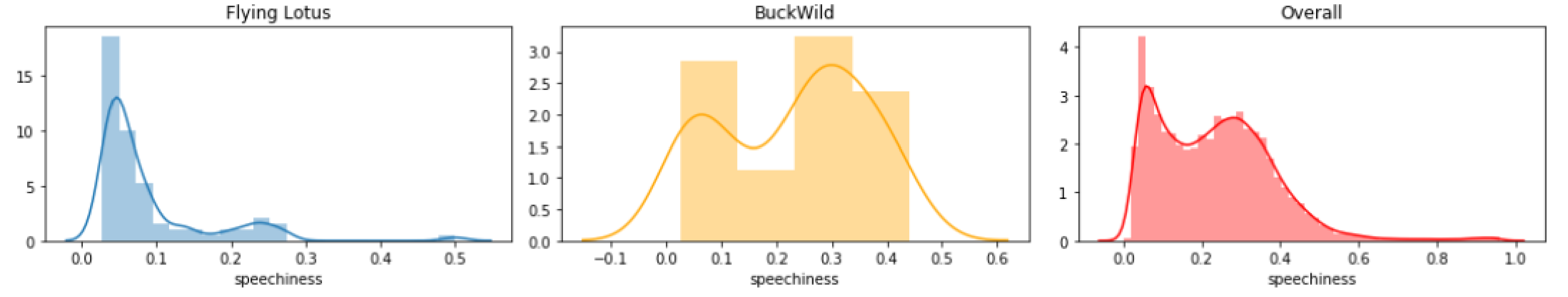
…What's the deal with Flying Lotus?

Hypothesis: vocals are hindering model accuracy

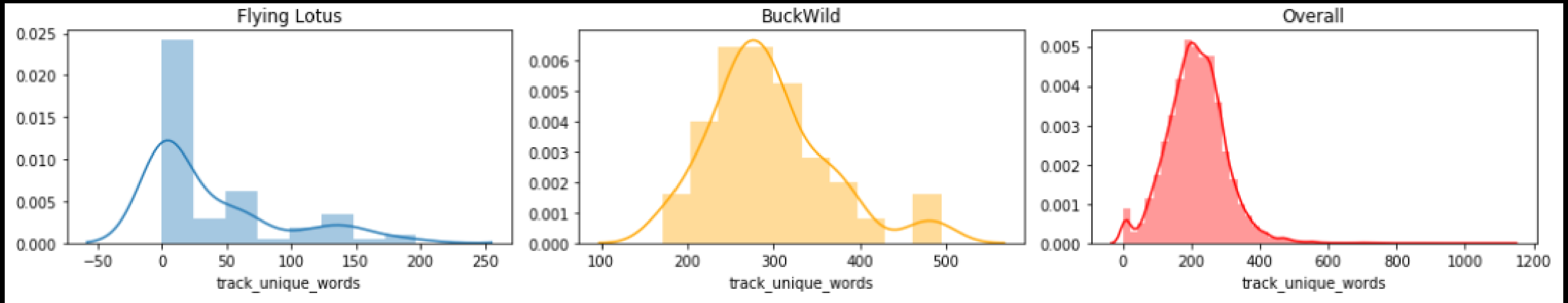
Engineer: Audio

Question: Are audio features an effective means of predicting the track producer?

Speechiness

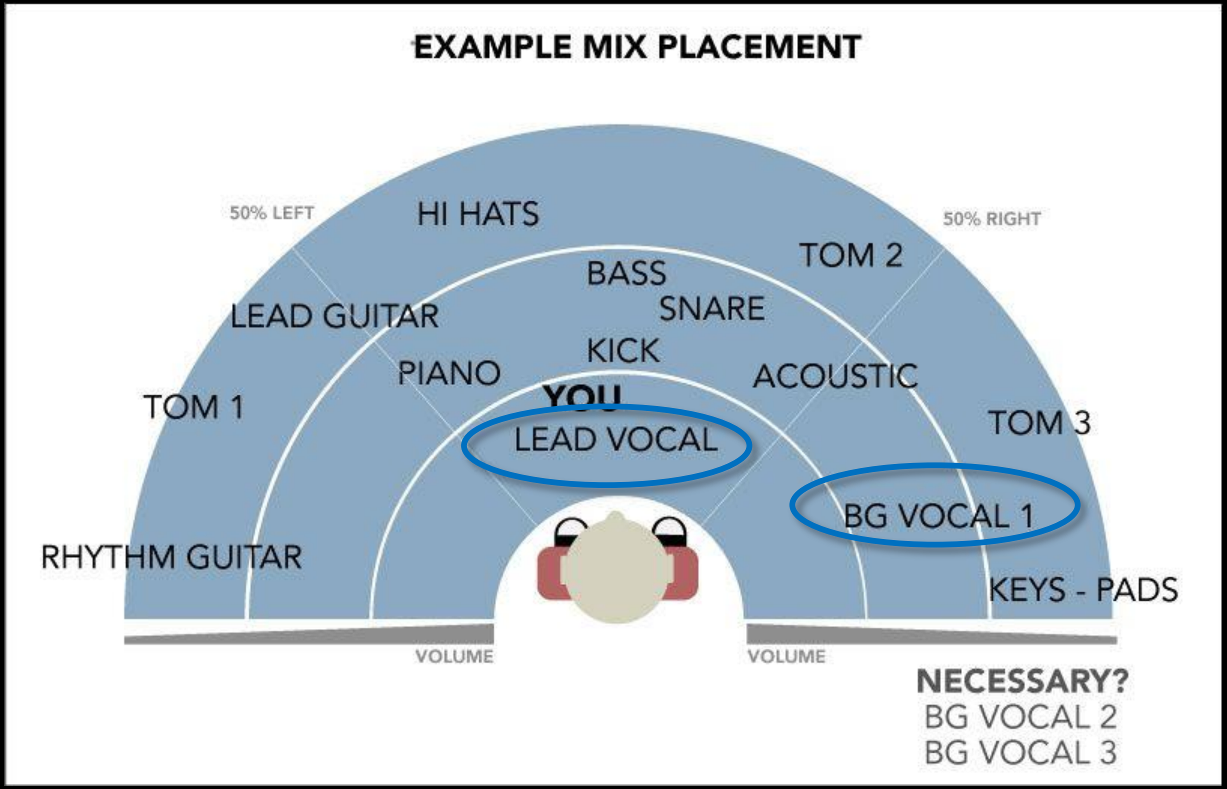


Unique Word Count



Engineer: Audio

Question: Are audio features an effective means of predicting the track producer?



In stereo, vocals recorded at "center" of music mix, can be isolated

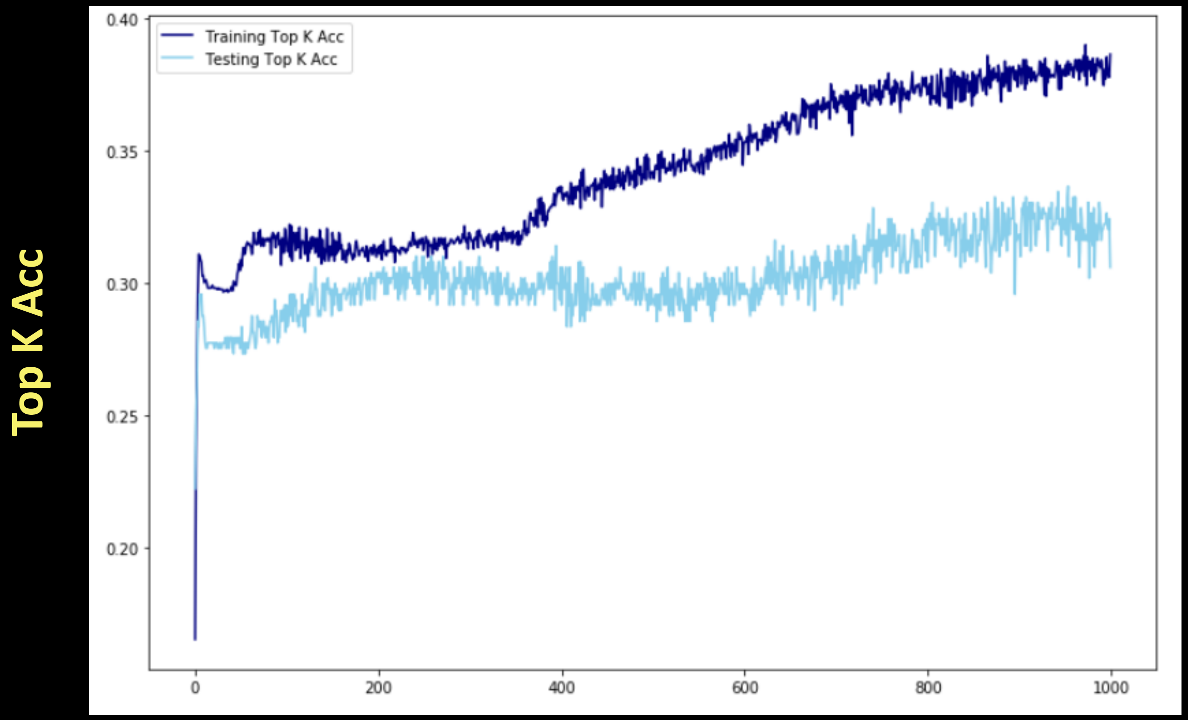
Original

Reduced Vocals

Engineer: Audio

Question: Are audio features an effective means of predicting the track producer?

NN Train v Test



Issues

- Loss of Other Instruments
- Fidelity Loss
- MIR Knowledge

Engineer: Audio

Question: Are audio features an effective means of predicting the track producer?

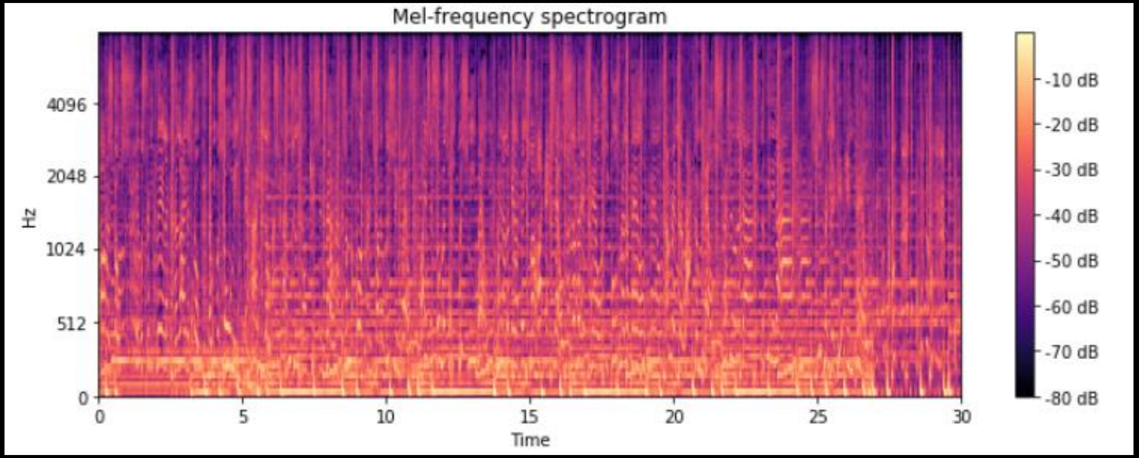
Better Processing

+

Audio Imaging

+

Convolutional Neural Network



Learning

- Our current set of features are not sufficient to predict a track's producer in order to impute
- Consider Convolution Neural Network

GENIUS

Engineer: Lyrics

Skill

Vocabulary

Sentiment

Topics

Engineer: Lyrics

GENIUS

Rhyme Density

A/B/A/B Scheme

Here it is the groove slightly transformed
Just a bit of a break from the norm
Just a little somethin' to break the monotony
Of all that hardcore dance that has gotten to be
A little bit out of control it's cool to dance
And think of the summers of the past
Adjust the base and let the alpine blast
Pop in my CD and let me run a rhyme
And put your car on cruise and lay back 'cause this is summertime

Will Smith - Summertime

Internal Rhyming

I'll knock you bitches into next week with a haymaker,
And straight-razor your face when you land seven days later,
I said pray your soul to keep when you go to sleep,
But you sold the lease on your own beliefs like Roman priests,
You probably told the beast Hip-Hop needs it's own police,
To patrol the streets and shows whenever something dope's released...

Diabolic – Stand By

Engineer: Lyrics

Rhyme Density

Phonetic Representation

but	you	sold	the	lease
b ʌ	j u:	s **əʊ** l d	ð ə	l **i:** s
on	your	own	beliefs	like
ɒ n	ɔ: ə	**əʊ** n	b ɪ l **i:** f s	l aɪ k
Roman	priests			
r **əʊ** m ə n	p r **i:** s t s			

…
I'll knock you bitches into next week with a haymaker,
And straight-razor your face when you land seven days later,
I said pray your soul to keep when you go to sleep,
But you sold the lease on your own beliefs like Roman priests,
You probably told the beast Hip-Hop needs it's own police,
To patrol the streets and shows whenever something dope's released,
…

$$\textbf{Avg Rhyme Density} = \frac{\text{\# of Potential Rhymes Per Line}}{\text{\# of Lines in Song}}$$

Engineer: Lyrics

Vocabulary

Complexity
(# of 3+ syllable words in vocab)

Size
(# of unique words in first 5K words)

Sentiment (VADER)

Positive

Negative

Neutral

NOTE: VADER is Lexicon based. Can get better results using supervised approach

Engineer: Lyrics

GENIUS

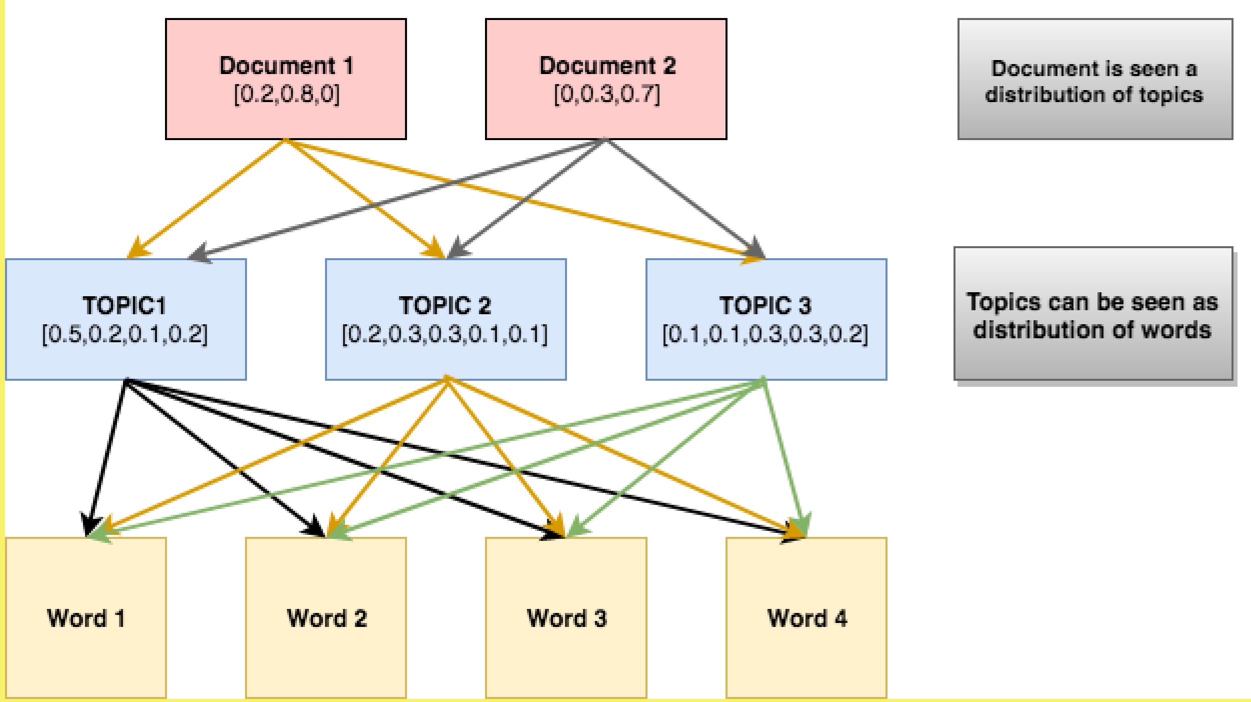
Question: What is the most effective way to extract **topics** from text?

LDA

Latent

Dirichlet

Allocation



Engineer: Lyrics

Question: What is the most effective way to extract **topics** from text?

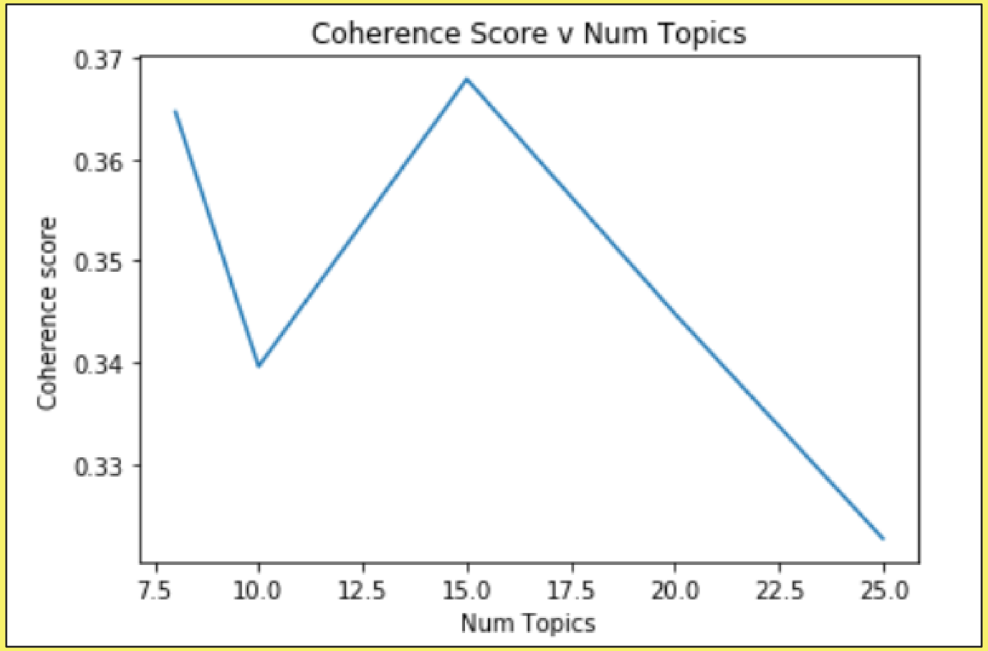
- Remove Stopwords
- **Profanity**
- Lemmatize
- Bag of Words
- Build Your Dict
- Select # Topics
- Score Model

Engineer: Lyrics

GENIUS

Question: What is the most effective way to extract **topics** from text?

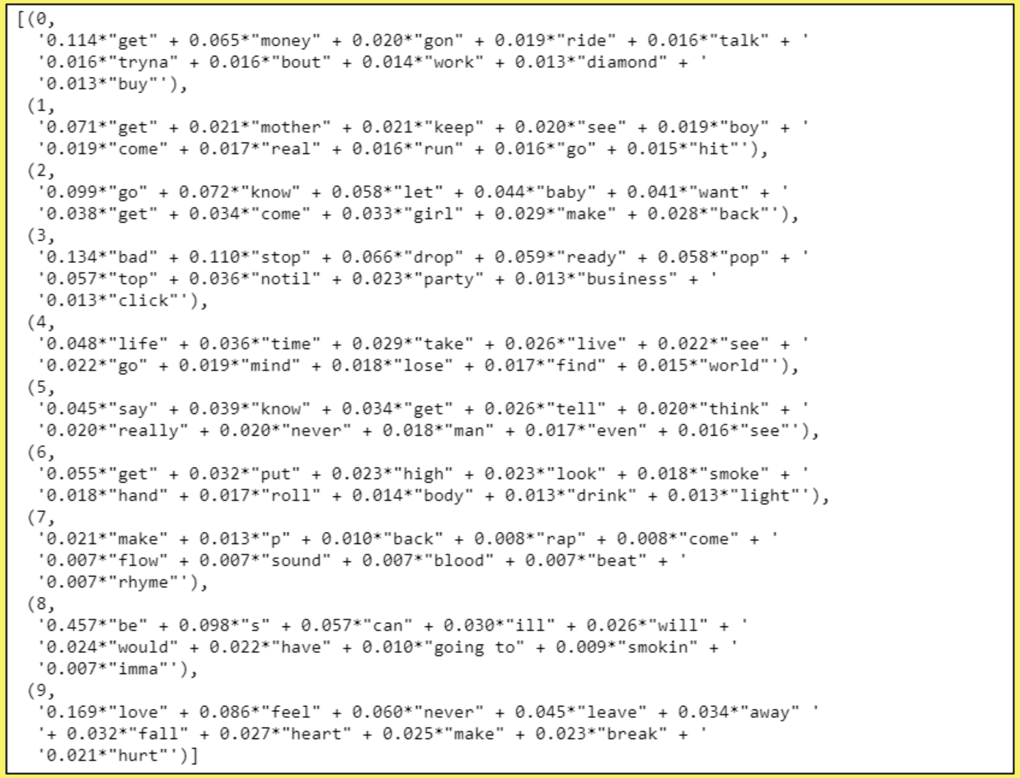
- **Coherence Score:** Degree of semantic similarity between top words used to describe a topic
- Must be balanced with interpretability…



Range: 0 (Bad) to **1** (Perfect)

Engineer: Lyrics

Question: What is the most effective way to extract **topics** from text?

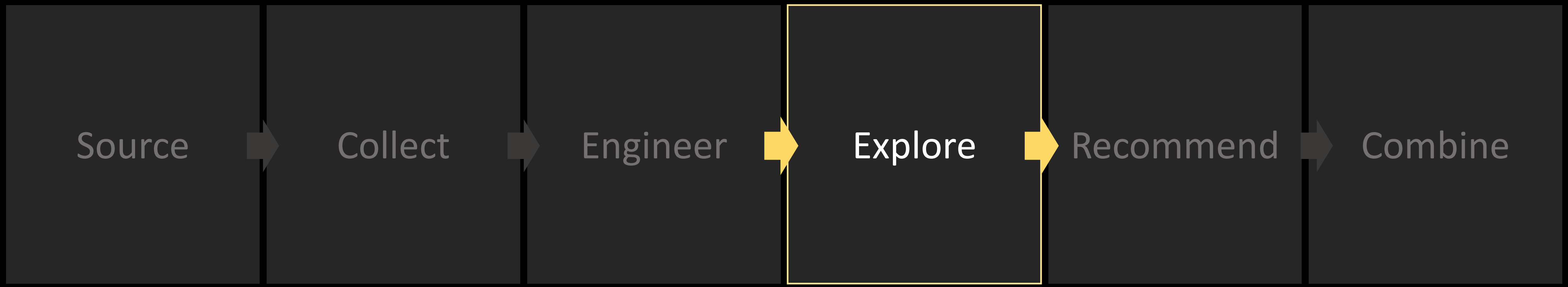


```
[(0,
  '0.114*"get" + 0.065*"money" + 0.020*"gon" + 0.019*"ride" + 0.016*"talk" + '
  '0.016*"tryna" + 0.016*"bout" + 0.014*"work" + 0.013*"diamond" + '
  '0.013*"buy"'),
 (1,
  '0.071*"get" + 0.021*"mother" + 0.021*"keep" + 0.020*"see" + 0.019*"boy" + '
  '0.019*"come" + 0.017*"real" + 0.016*"run" + 0.016*"go" + 0.015*"hit"'),
 (2,
  '0.099*"go" + 0.072*"know" + 0.058*"let" + 0.044*"baby" + 0.041*"want" + '
  '0.038*"get" + 0.034*"come" + 0.033*"girl" + 0.029*"make" + 0.028*"back"'),
 (3,
  '0.134*"bad" + 0.110*"stop" + 0.066*"drop" + 0.059*"ready" + 0.058*"pop" + '
  '0.057*"top" + 0.036*"notil" + 0.023*"party" + 0.013*"business" + '
  '0.013*"click"'),
 (4,
  '0.048*"life" + 0.036*"time" + 0.029*"take" + 0.026*"live" + 0.022*"see" + '
  '0.022*"go" + 0.019*"mind" + 0.018*"lose" + 0.017*"find" + 0.015*"world"'),
 (5,
  '0.045*"say" + 0.039*"know" + 0.034*"get" + 0.026*"tell" + 0.020*"think" + '
  '0.020*"really" + 0.020*"never" + 0.018*"man" + 0.017*"even" + 0.016*"see"'),
 (6,
  '0.055*"get" + 0.032*"put" + 0.023*"high" + 0.023*"look" + 0.018*"smoke" + '
  '0.018*"hand" + 0.017*"roll" + 0.014*"body" + 0.013*"drink" + 0.013*"light"'),
 (7,
  '0.021*"make" + 0.013*"p" + 0.010*"back" + 0.008*"rap" + 0.008*"come" + '
  '0.007*"flow" + 0.007*"sound" + 0.007*"blood" + 0.007*"beat" + '
  '0.007*"rhyme"'),
 (8,
  '0.457*"be" + 0.098*"s" + 0.057*"can" + 0.030*"ill" + 0.026*"will" + '
  '0.024*"would" + 0.022*"have" + 0.010*"going to" + 0.009*"smokin" + '
  '0.007*"imma"'),
 (9,
  '0.169*"love" + 0.086*"feel" + 0.060*"never" + 0.045*"leave" + 0.034*"away" '
  '+ 0.032*"fall" + 0.027*"heart" + 0.025*"make" + 0.023*"break" + '
  '0.021*"hurt"')]
```

Topics
The Hustle
Family
Lust
Partying
Reflection
Storytelling
Drugs
Skills
Aspirations
Love

Learning

- LDA is a powerful tool for topic extraction but will require subjectivity (num topics, topic interpretation)
- Can be used to programmatically generate tags based on content



Source

Collect

Engineer

Explore

Recommend

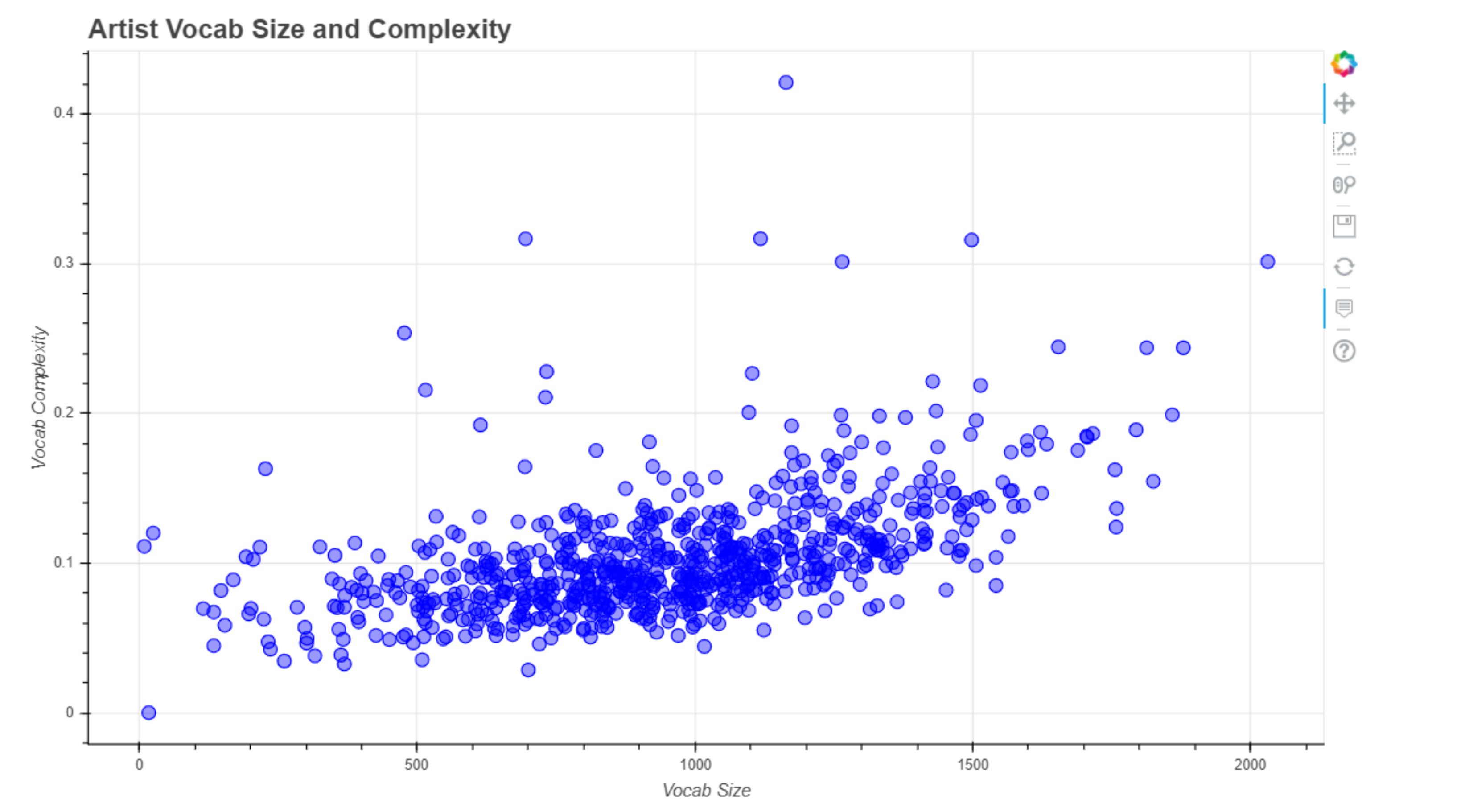
Combine

- **Lyrical Deep Dive**

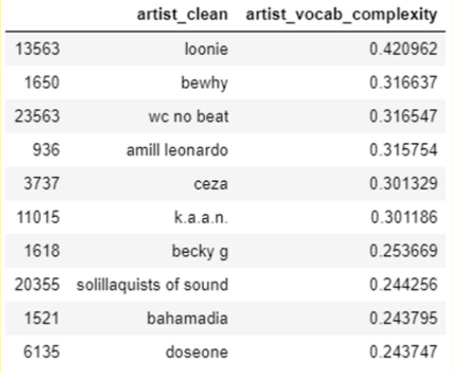
Lyrical Insights: Vocab

GENIUS

Rappers by Vocabulary

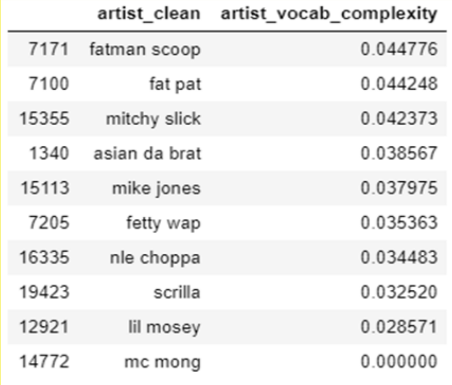


High Vocab Complexity



	artist_clean	artist_vocab_complexity
13563	loonie	0.420962
1650	bewhy	0.316637
23563	wc no beat	0.316547
936	amill leonardo	0.315754
3737	ceza	0.301329
11015	k.a.a.n.	0.301186
1618	becky g	0.253669
20355	solillaquists of sound	0.244256
1521	bahamadia	0.243795
6135	doseone	0.243747

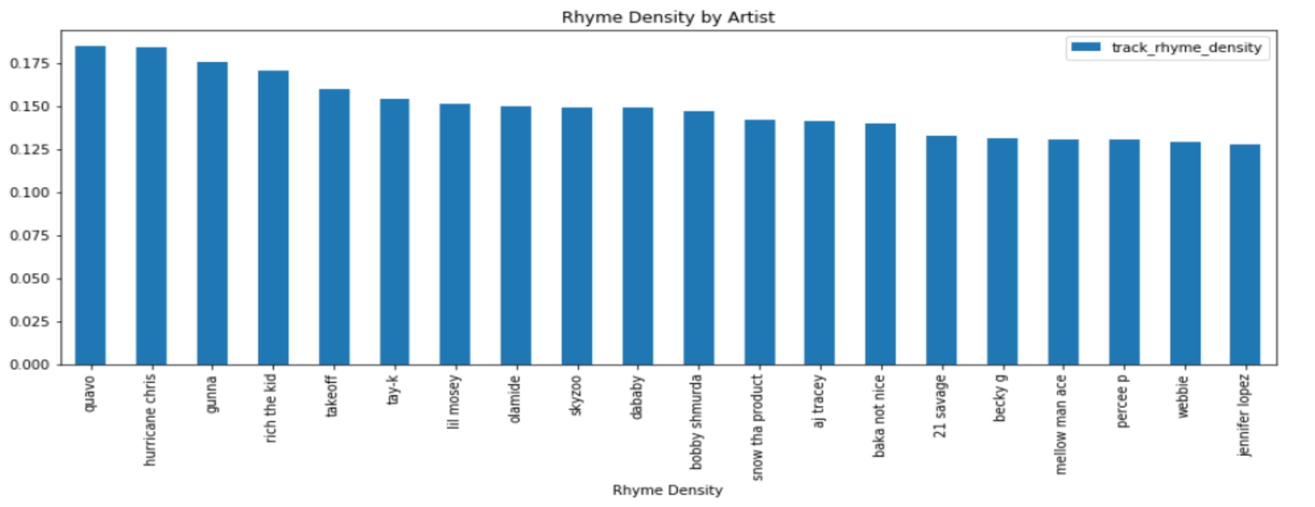
Low Vocab Complexity



	artist_clean	artist_vocab_complexity
7171	fatman scoop	0.044776
7100	fat pat	0.044248
15355	mitchy slick	0.042373
1340	asian da brat	0.038567
15113	mike jones	0.037975
7205	fetty wap	0.035363
16335	nle choppa	0.034483
19423	scrilla	0.032520
12921	lil mosey	0.028571
14772	mc mong	0.000000

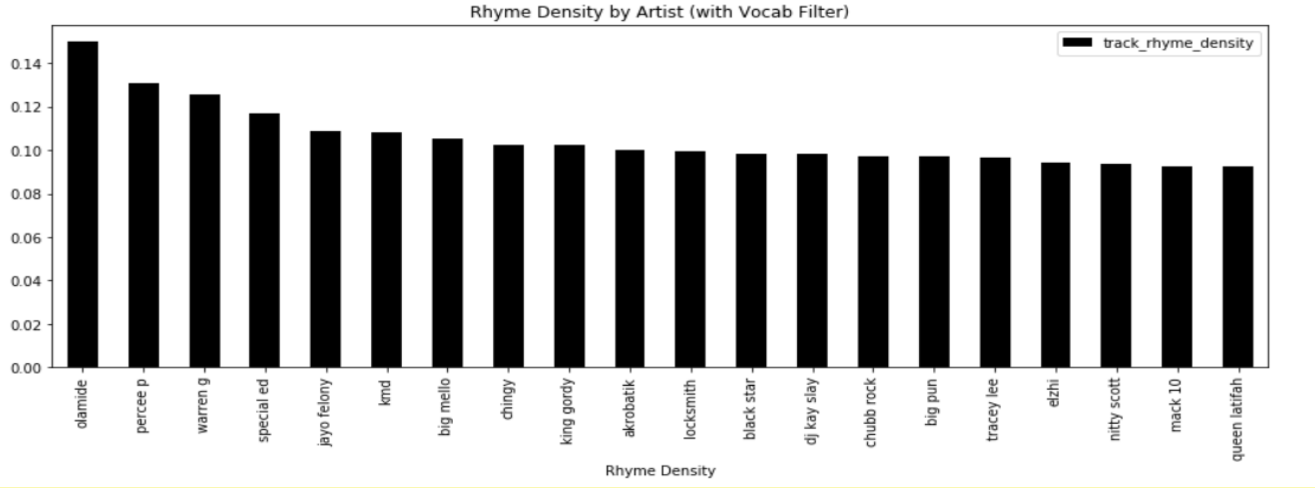
Lyrical Insights: Flow

Rhyme Density



- Quavo (Migos)
- Hurricane Chris
- Gunna
- Rich the Kid

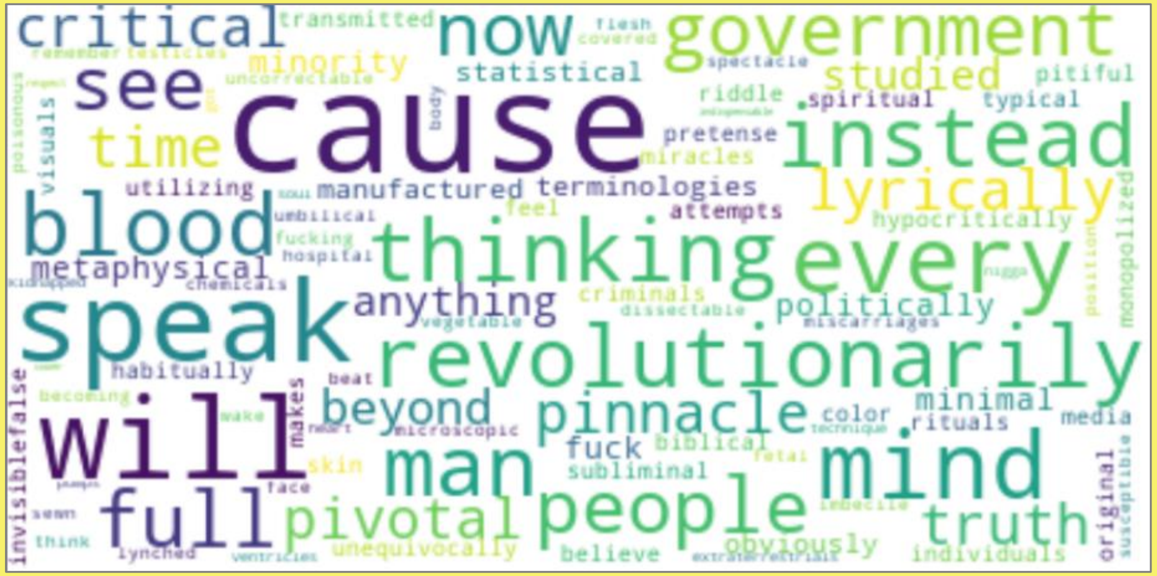
Rhyme Density & Vocab Size



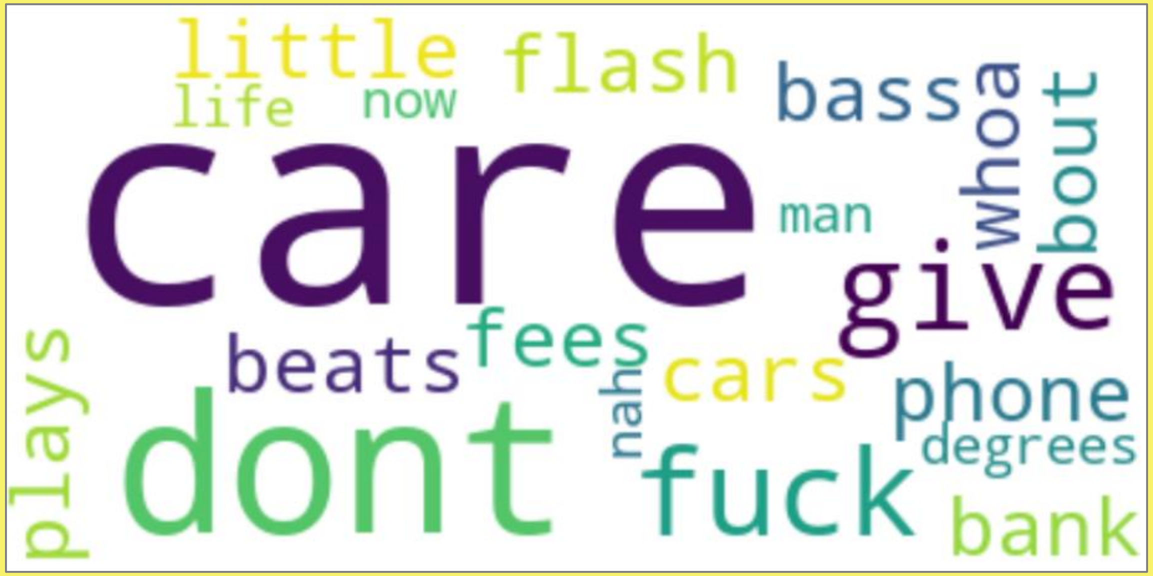
- Olamide, Percee P -- Bilingual!
- **Warren G, Big Pun**
- **Akrobatik**

Lyrical Insights: Complexity

Most Complex Track	Least Complex Track
Immortal Technique: Speak Your Mind	**JPEGMAFIA:** *My Thoughts On Neogaf Dying*



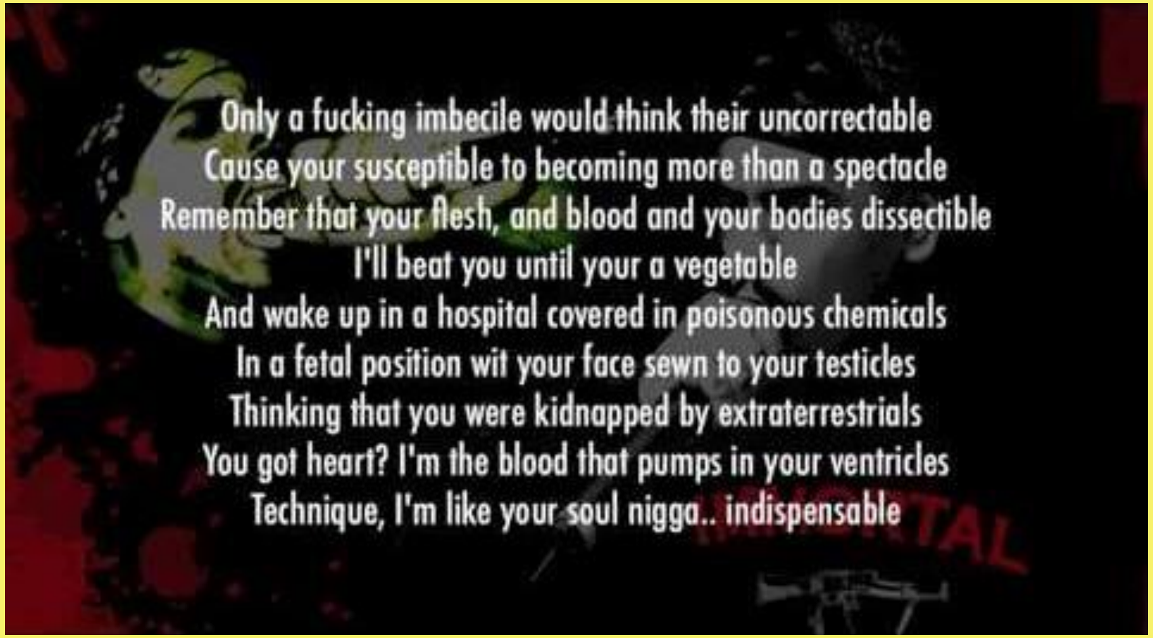
*Based on track vocab complexity



*Based on track vocab complexity

Lyrical Insights: Complexity

Most Complex Track	Least Complex Track
Immortal Technique: Speak Your Mind	**JPEGMAFIA:** *My Thoughts On Neogaf Dying*



Only a fucking imbecile would think their uncorrectable
Cause your susceptible to becoming more than a spectacle
Remember that your flesh, and blood and your bodies dissectible
I'll beat you until your a vegetable
And wake up in a hospital covered in poisonous chemicals
In a fetal position wit your face sewn to your testicles
Thinking that you were kidnapped by extraterrestrials
You got heart? I'm the blood that pumps in your ventricles
Technique, I'm like your soul nigga.. indispensable

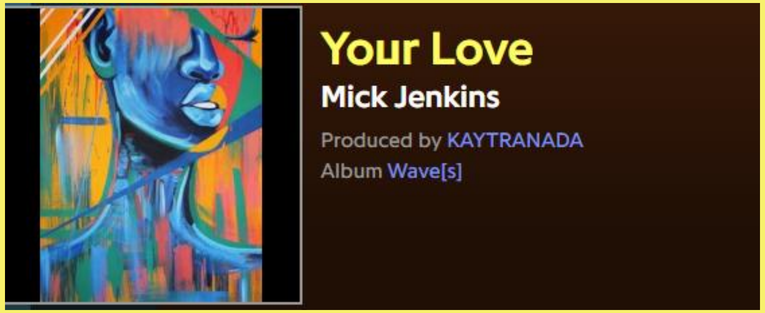
Lyrical Insights: Sentiment

GENIUS

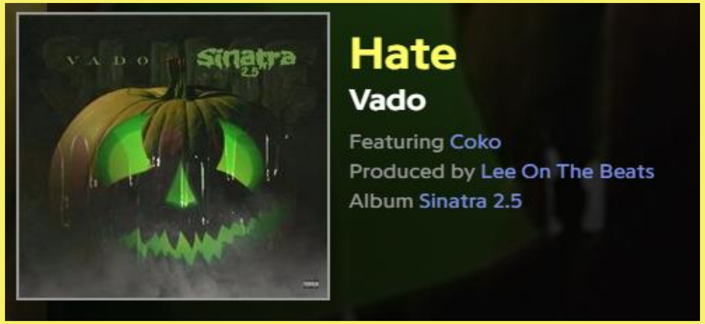
☺ High Positive Sentiment Score

☹ High Negative Sentiment Score

CORRECT

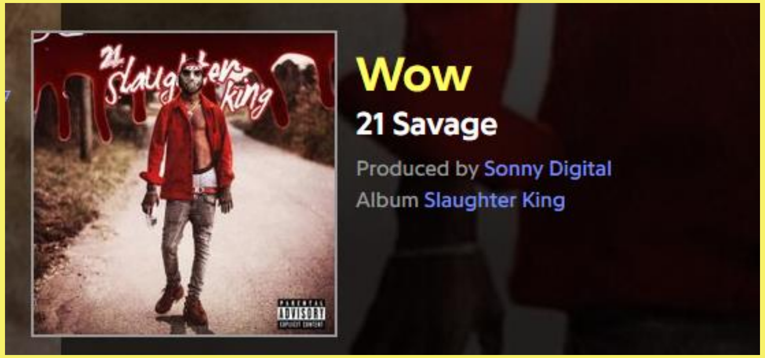


Your Love
Mick Jenkins
Produced by KAYTRANADA
Album Wave[s]

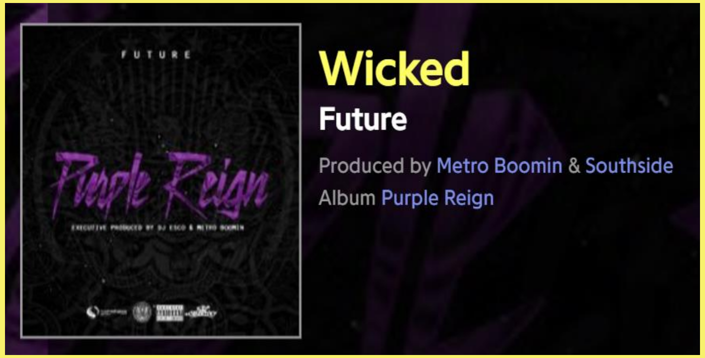


Hate
Vado
Featuring Coko
Produced by Lee On The Beats
Album Sinatra 2.5

INCORRECT



Wow
21 Savage
Produced by Sonny Digital
Album Slaughter King

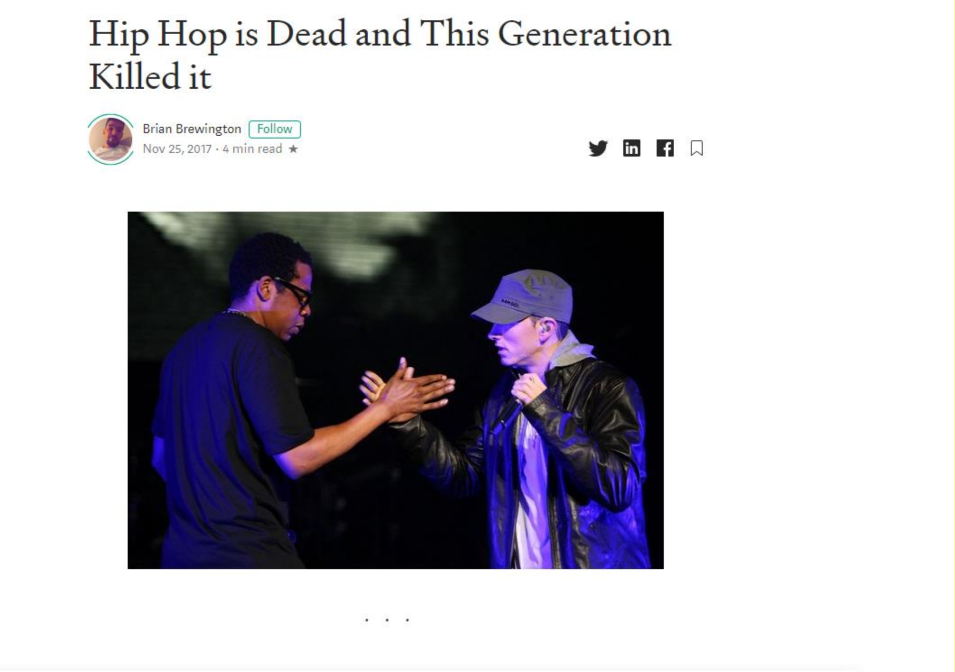


Wicked
Future
Produced by Metro Boomin & Southside
Album Purple Reign

Lyrical Insights: Macro Changes

GENIUS

Question: How has the genre changed? Is hip hop dead as an art form?



Hip Hop is Dead and This Generation Killed it

Brian Brewington Follow

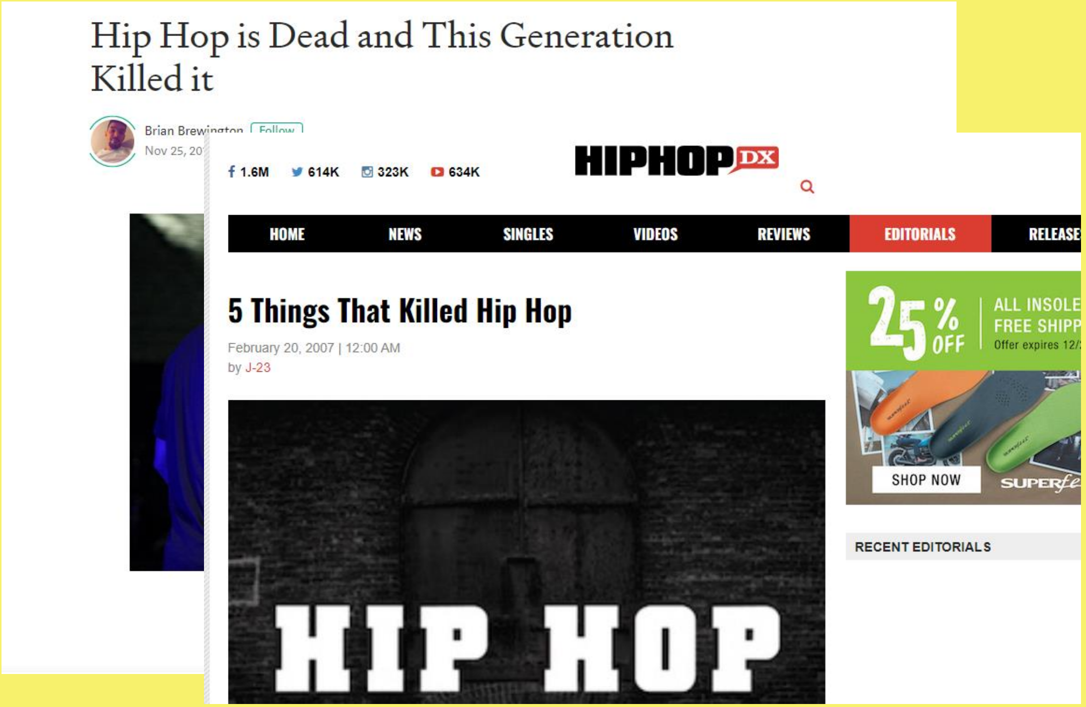
Nov 25, 2017 · 4 min read ★

Hip Hop
1979 - 2017

Lyrical Insights: Macro Changes

GENIUS

Question: How has the genre changed? Is hip hop dead as an art form?



Hip Hop is Dead and This Generation Killed it

Brian Brewington Follow
Nov 25, 20

HIPHOPDX

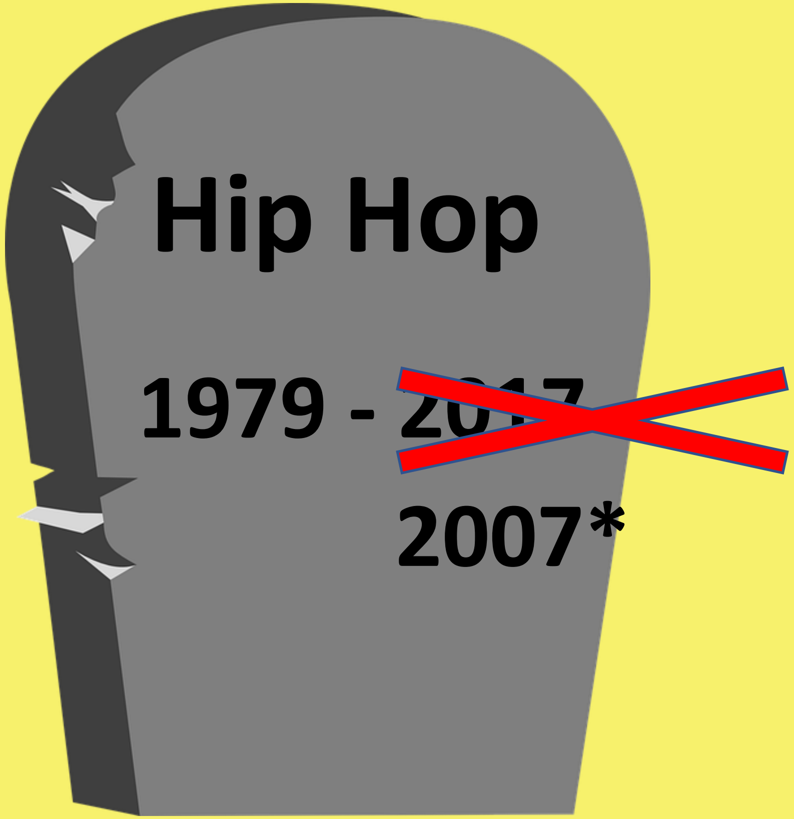
1.6M 614K 323K 634K

HOME NEWS SINGLES VIDEOS REVIEWS EDITORIALS RELEASE

5 Things That Killed Hip Hop

February 20, 2007 | 12:00 AM
by J-23

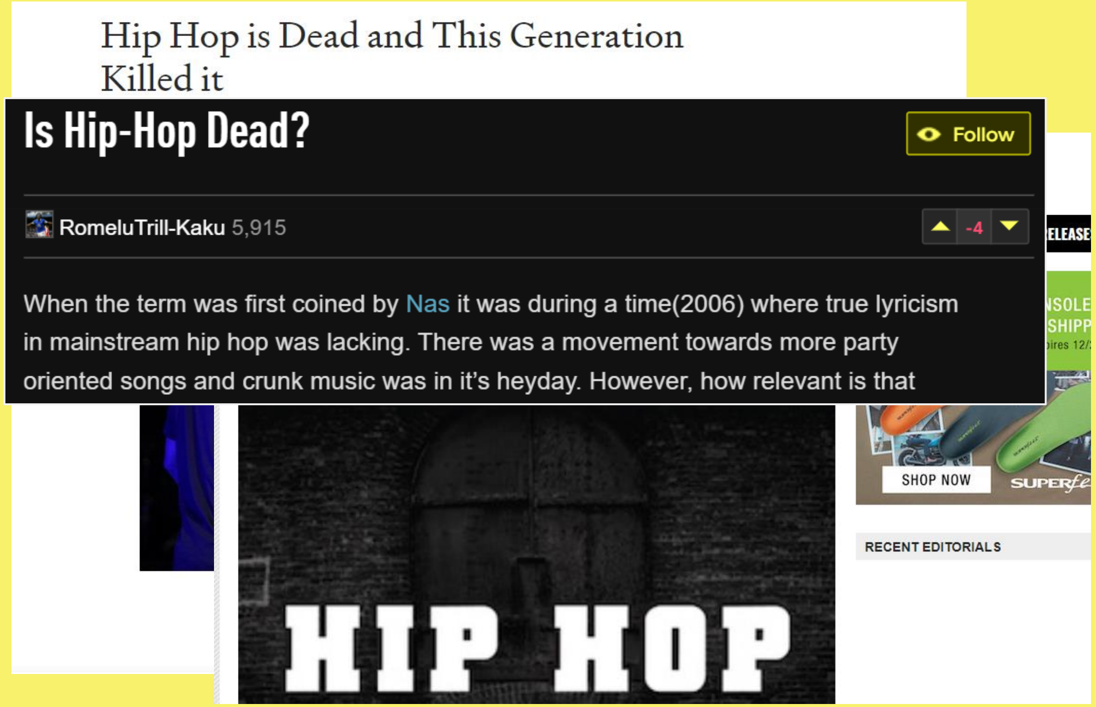
RECENT EDITORIALS



Lyrical Insights: Macro Changes

GENIUS

Question: How has the genre changed? Is hip hop dead as an art form?



Hip Hop is Dead and This Generation Killed it

Is Hip-Hop Dead?

RomeluTrill-Kaku 5,915

When the term was first coined by Nas it was during a time(2006) where true lyricism in mainstream hip hop was lacking. There was a movement towards more party oriented songs and crunk music was in it's heyday. However, how relevant is that

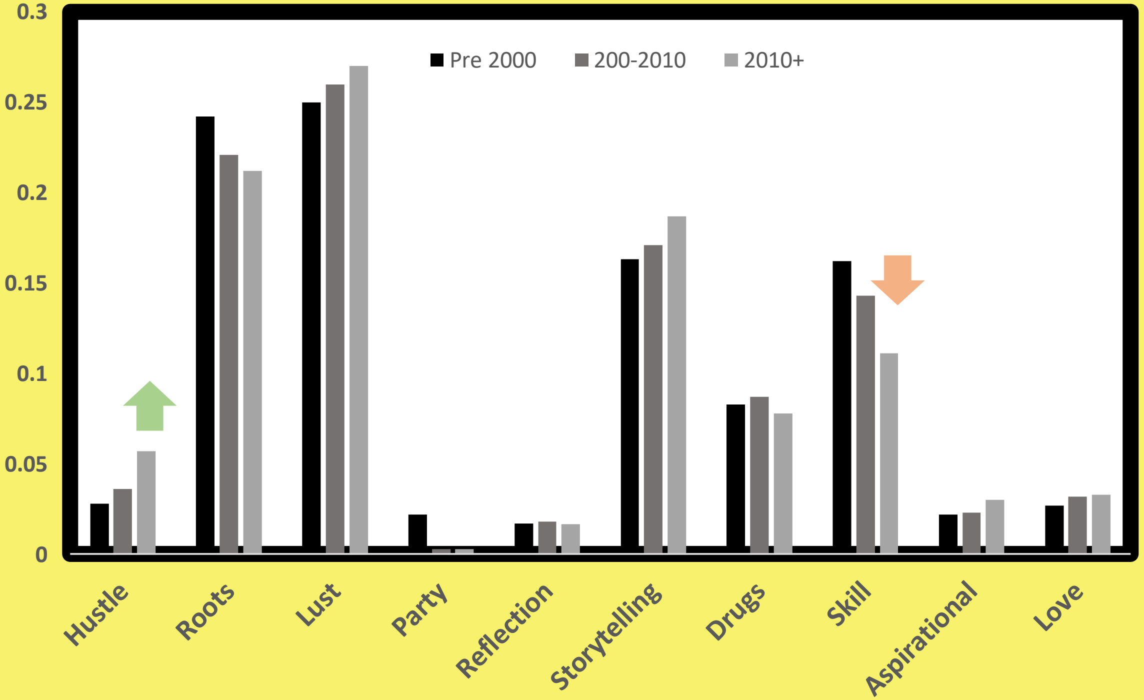
Hip Hop
1979 - ~~2017~~
~~2007*~~
2006*

Lyrical Insights: Macro Changes

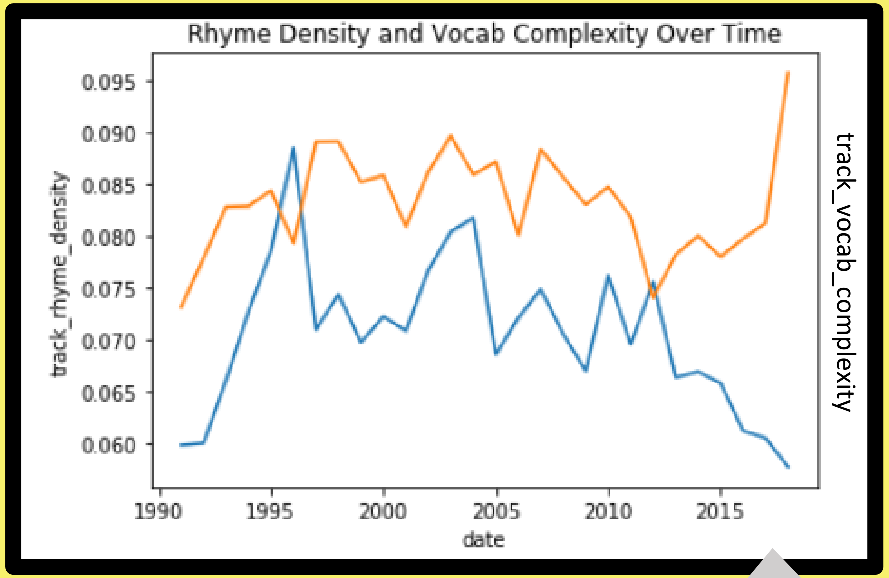
GENIUS

Question: How has the genre changed? Is hip hop dead as an art form?

Topics



Flow and Vocab

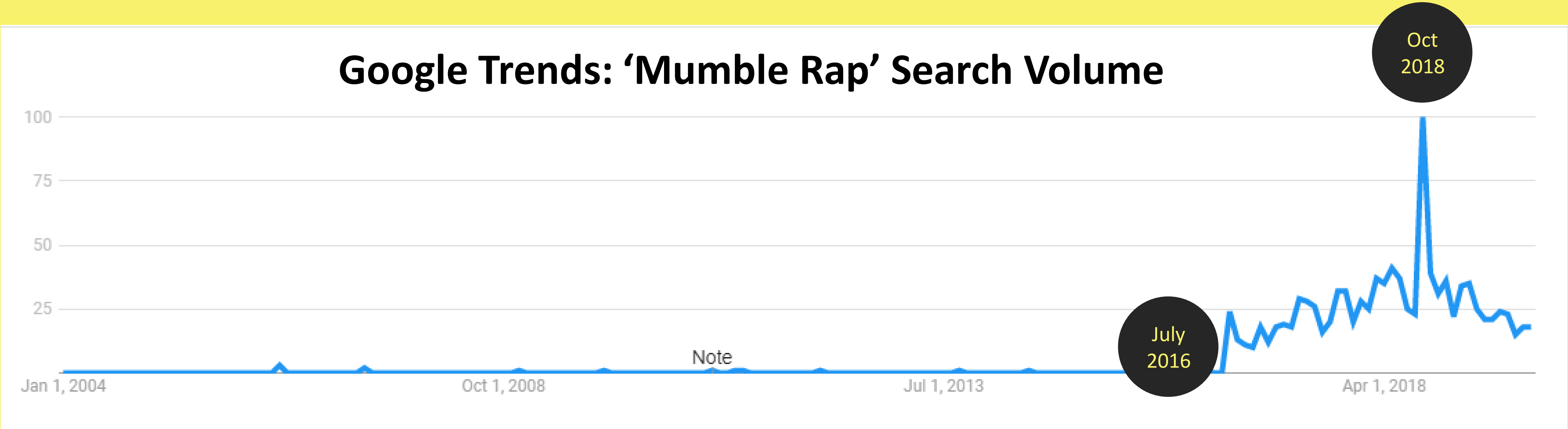


2015: Migos first major mixtape

Lyrical Insights: Macro Changes

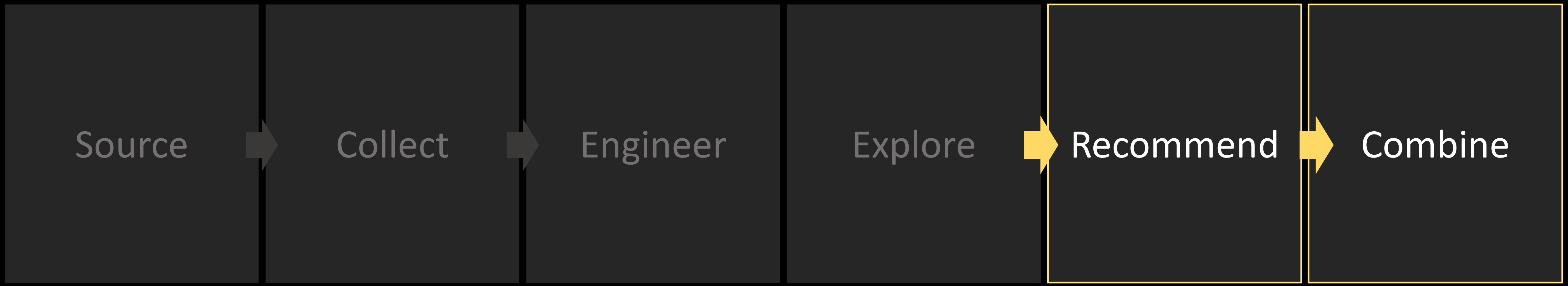
GENIUS

Question: How has the genre changed? Is hip hop dead as an art form?



Learning

- While it's not dead, it's become much more about flow than about the actual lyrical content



Source

Collect

Engineer

Explore

Recommend

Combine

- Types of Engines
- Watchouts
- Measurement
- Observations

- Hybrid Approach

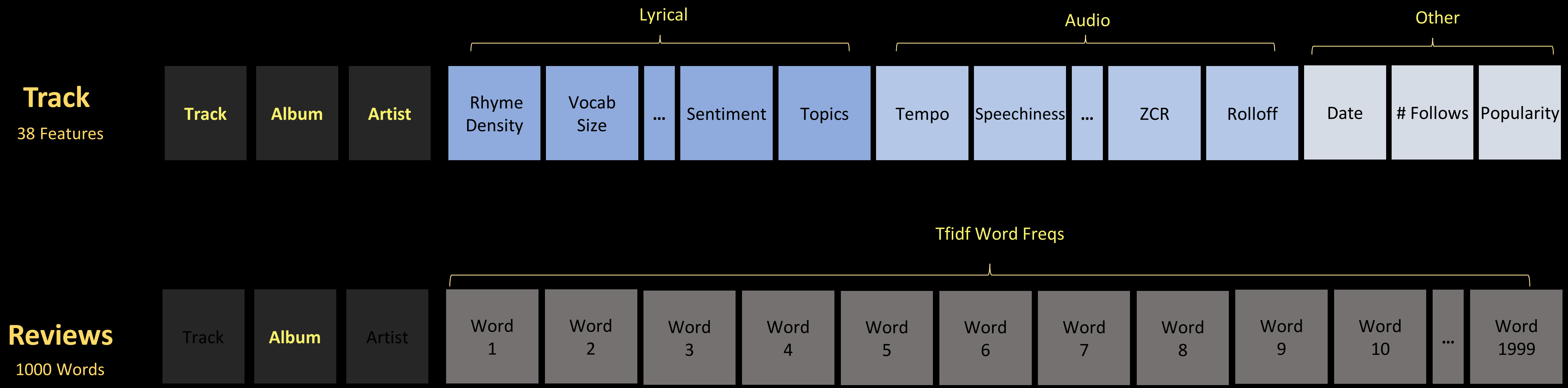
Recommend: Data Summary

Track

38 Features

			Lyrical					Audio					Other		
Track	**Album**	**Artist**	Rhyme Density	Vocab Size	...	Sentiment	Topics	Tempo	Speechiness	...	ZCR	Rolloff	Date	# Follows	Popularity

Recommend: Data Summary



Track

38 Features

Lyrical

Audio

Other

Track	**Album**	**Artist**	Rhyme Density	Vocab Size	…	Sentiment	Topics	Tempo	Speechiness	…	ZCR	Rolloff	Date	# Follows	Popularity

Reviews

1000 Words

Tfidf Word Freqs

Track	**Album**	Artist	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	…	Word 1999

Recommend: Data Summary

				Lyrical					Audio					Other		
Track 38 Features	**Track**	**Album**	**Artist**	Rhyme Density	Vocab Size	...	Sentiment	Topics	Tempo	Speechiness	...	ZCR	Rolloff	Date	# Follows	Popularity

				Tfidf Word Freqs											
Reviews 1000 Words	Track	**Album**	Artist	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	...	Word 1000

				Binary Follow Flags											
Followers 151K	Track	Album	**Artist**	N00b_ saibot	XXX COD420	Gooch Crucible	Squiggly Bob	$mokeyyy	kartashov	TruSwag	YBN Mystique	Astroboy	efray	...	Squiggly Bob

Recommend: Types of Recommenders

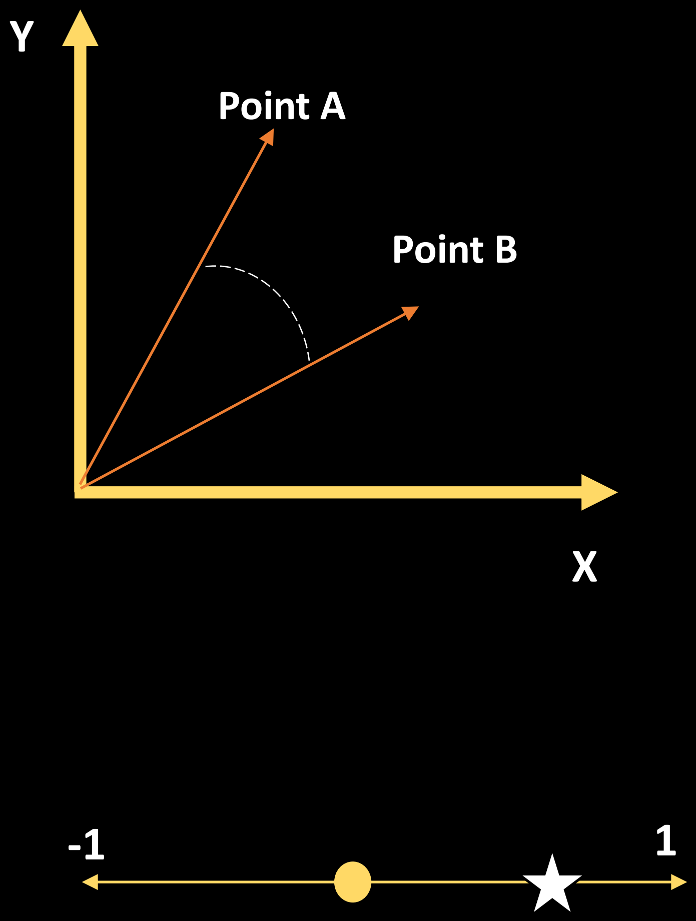
Bucket	Requirements	Advantages	Disadvantages
Knowledge-Based	• Significant knowledge of the domain in question	• Minimal reliance on data or specific user preferences. • Approach for high complexity/barrier items	• Need extensive domain knowledge • Scalability a factor
Heirarchical	• Clean product/content management system	• Simple to implement (I bought a printer, so I probably need ink)	• Limited personalization opportunity
★ **Content Based Filtering**	• Clean and consistent metadata at the item level	• No community required	• Overengineering possible, leading to **unexpected** results • Best when aligned to user preferences
★ **Collaborative Filtering**	• Large historical data set • Set of users and preferences	• No domain knowledge required • Simple implementations can be effective	• Sparsity of Data • Cold-Start User Problem

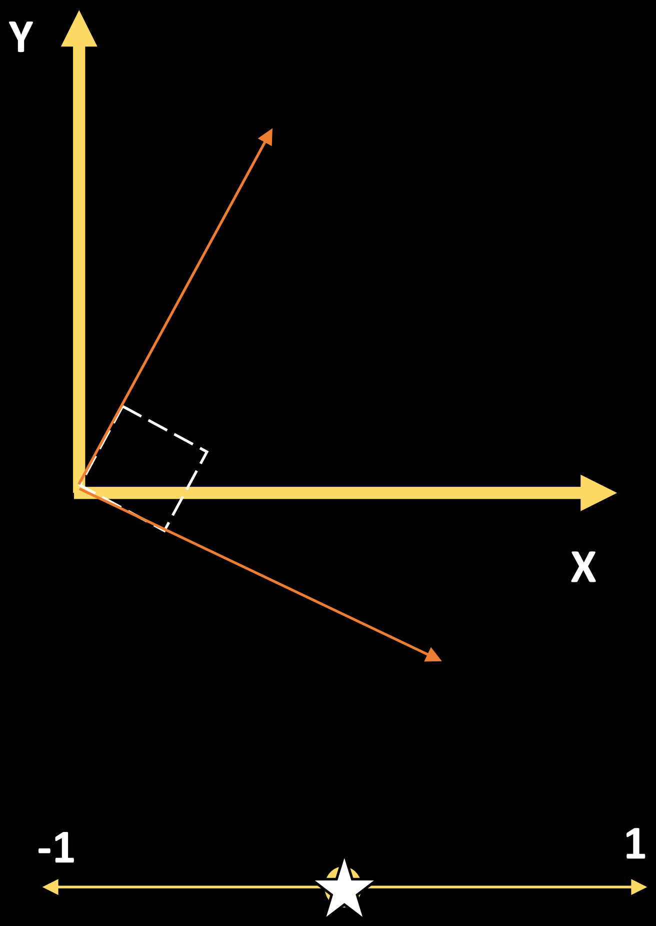
Recommend: Watchouts

What are the pitfalls of recommender systems?

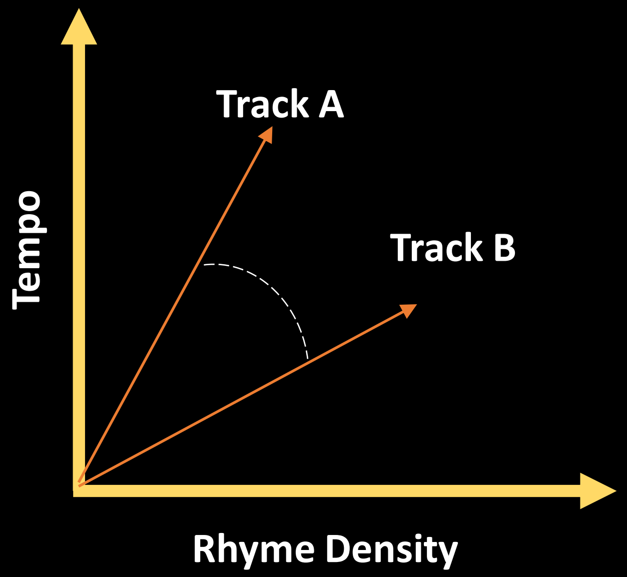
Watchouts	Description	Example
Synonymy	• Products or content that looks **very similar** but is in fact different.	• Eminem – *Without Me* • Emimen – *Without Me* (Remastered) • Eminem – *Without Me* (Radio Edit)
Novelty vs Correctness	• Combining novelty and correctness produces serendipitous moments for the user	• Danny Brown – *Bruiser Brigade* • Casey Veggies – *PNCINTLOFWGKTA*
Scalability	• How effectively an engine can be scaled as underlying data grows • ***Sparsity*** can help here	• Collaborative filter + massive social networks • Page Views
New Additions	• The introduction of a new user or product into your algorithm – how will they be treated?	• New user with no profile • Users with low rating volume • New item with minimal metadata

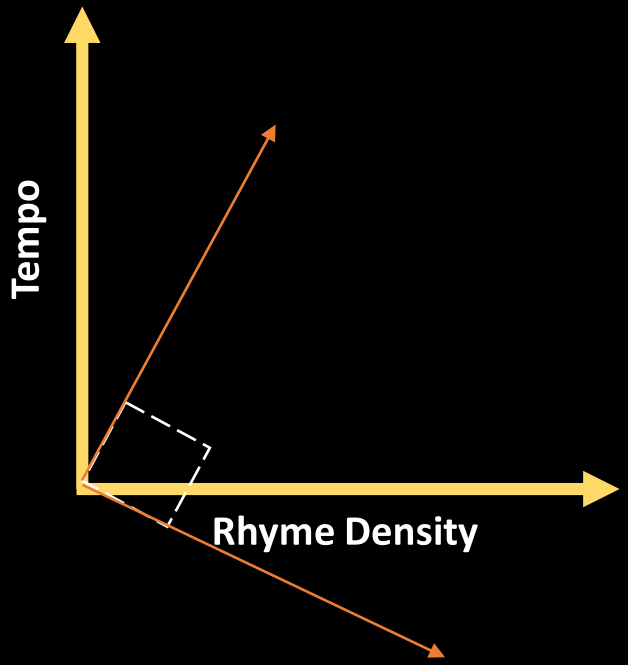
Recommend: Measurement (Cosine Similarity)





Recommend: Measurement (Cosine Similarity)





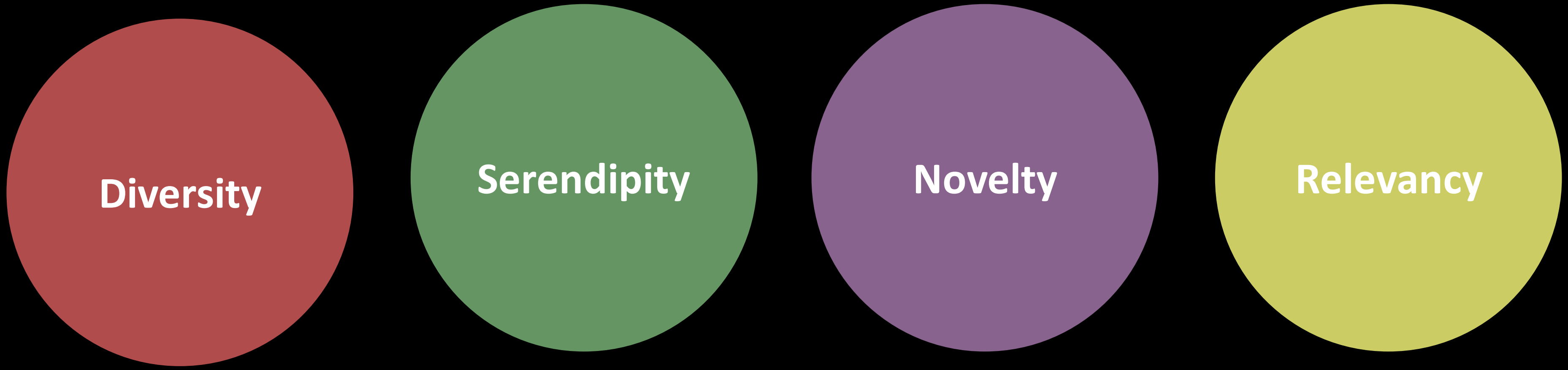
Close Relationship

RECOMMEND

No Relationship

~~RECOMMEND~~

Recommend: Measurement (Macro KPIs)



Diversity

Serendipity

Novelty

Relevancy

Recommend: Collaborative Filtering (Followers)

Current Artist	Users Also Listen To…	Similarity
 Kendrick Lamar	Schoolboy Q	0.38
	A$AP Rocky	0.37
	Nas	0.34
	J. Cole	0.33
	Chance the Rapper	0.31
	Big Sean	0.30
	Ab-Soul	0.30
	…	

Current Artist	Users Also Listen To…	Similarity
 Chance The Rapper	Childish Gambino	0.36
	Kendrick Lamar	0.33
	Schoolboy Q	0.31
	Ab-Soul	0.29
	Mac Miller	0.28
	A$SAP Rocky	0.27
	Vic Mensa	0.24
	…	

Diversity

Serendipity

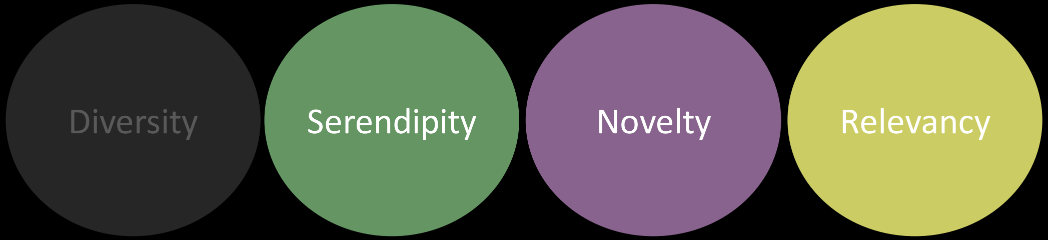
Novelty

Relevancy

Recommend: Content Based (Review Text)

Current Album	Related Reviews	Similarity
 Kendrick Lamar *To Pimp A Butterfly*	Kendrick Lamar: *DAMN*	0.89
	Kendrick Lamar: *Good Kid, MAAD City*	0.78
	Kendrick Lamar: *Section .80*	0.75
	Kendrick Lamar: *Untitled*	0.73
	Ab-Soul: *These Days*	0.66
	Logic: *The Incredible True Story*	0.54
	Schoolboy Q: *Habits and Contradictions*	0.53
	…	

Current Album	Related Reviews	Similarity
 Chance The Rapper *Acid Rap*	Chance the Rapper: *The Big Day*	0.86
	Chance the Rapper: *Coloring Book*	0.75
	Donnie Trumpet : *Surf*	0.64
	~~Jeru The Demaja: *Divine Design*~~	0.61
	Childish Gambino: *Because the Internet*	0.57
	Mackelmore: *This Unruly Mess I've Made*	0.55
	Bobby Womack: *The Bravest Man In…*	0.54
	…	



Recommend: Content Based Filtering (Track Features)

Current Track	Current Album	Current Artist
 Alright	 To Pimp a Butterfly	 Kendrick Lamar

Track	Album	Artist	Similarity
Lost Ones	*Cole World*	J Cole	0.32
~~Kiss Kiss~~	~~*Exclusive*~~	~~Chris Brown~~	~~0.31~~
Too Good	*Views From the 6*	Drake	0.31
Can't Feel My Face	*Beauty Behind Madness*	The Weeknd	0.29
Land of the Snakes	***Born Sinner***	**J Cole**	**0.28**
sing about me im dying	*To Pimp a Butterfly*	Kendrick Lamar	0.27
Cocoa Butter Kisses	*Acid Rap*	Chance the Rapper	0.27
What in XXXTarnation	***Members Only***	**XXXTentacion**	**0.26**

Diversity

Serendipity

Novelty

Relevancy

Recommend: Content Based Filtering (Track Features)

Current Track	Current Album	Current Artist
 Favorite Song	 Acid Rap	 Chance the Rapper

Track	Album	Artist	Similarity
Global	*Street Gossip*	Lil Baby	0.41
~~Codeine Dreaming~~	~~*Project Baby 2*~~	~~Kodack Black~~	~~0.40~~
Hot Shower	*The Big Day*	Chance The Rapper	0.40
I'm Straight	*Harder Than Ever*	Lil Baby	0.38
This is a Hit	***Beautiful Loser***	**Kyle**	**0.33**
~~The Rain~~	~~*Supa Dupa Fly*~~	~~Missy Elliot~~	~~0.29~~
~~Gunwalk~~	~~*I Am Not Human Being2*~~	~~Lil Wayne~~	~~0.29~~
Giv no Fucks	***Late Nights***	**Jeremih**	**0.28**

Diversity

Serendipity

Novelty

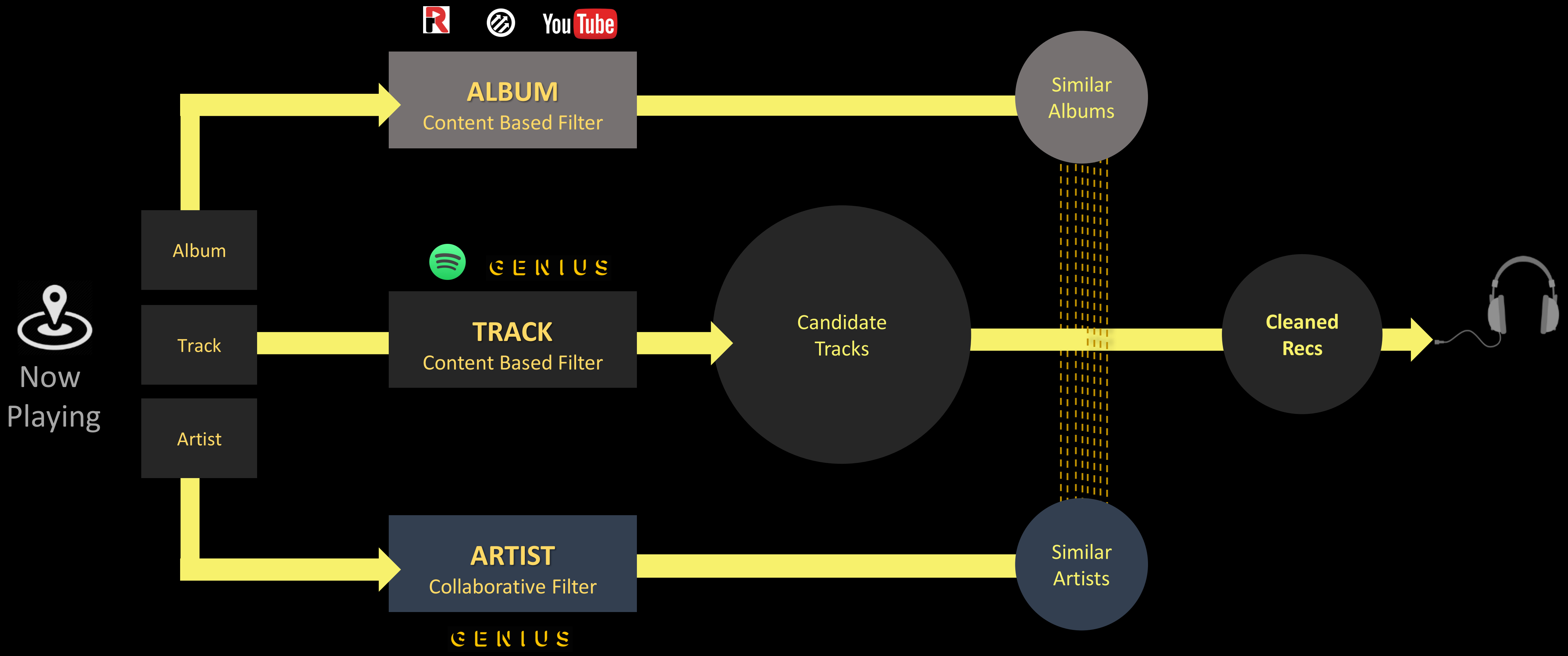
Relevancy

Engineer: Audio

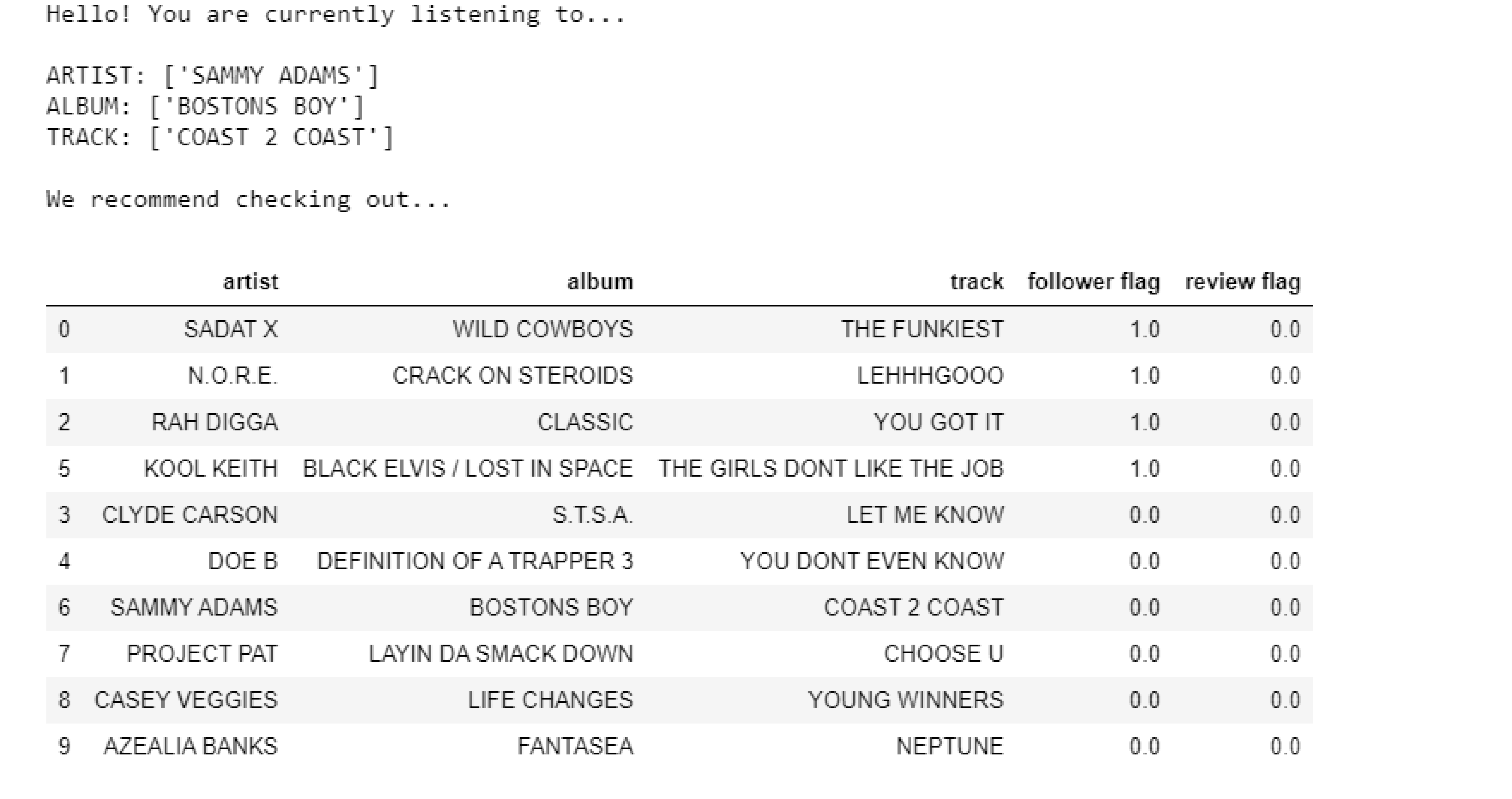
Maintain: Novelty/Serendipity

Improve: Diversity/Relevance

Recommend: Hybrid Approach



Recommend: Hybrid Approach



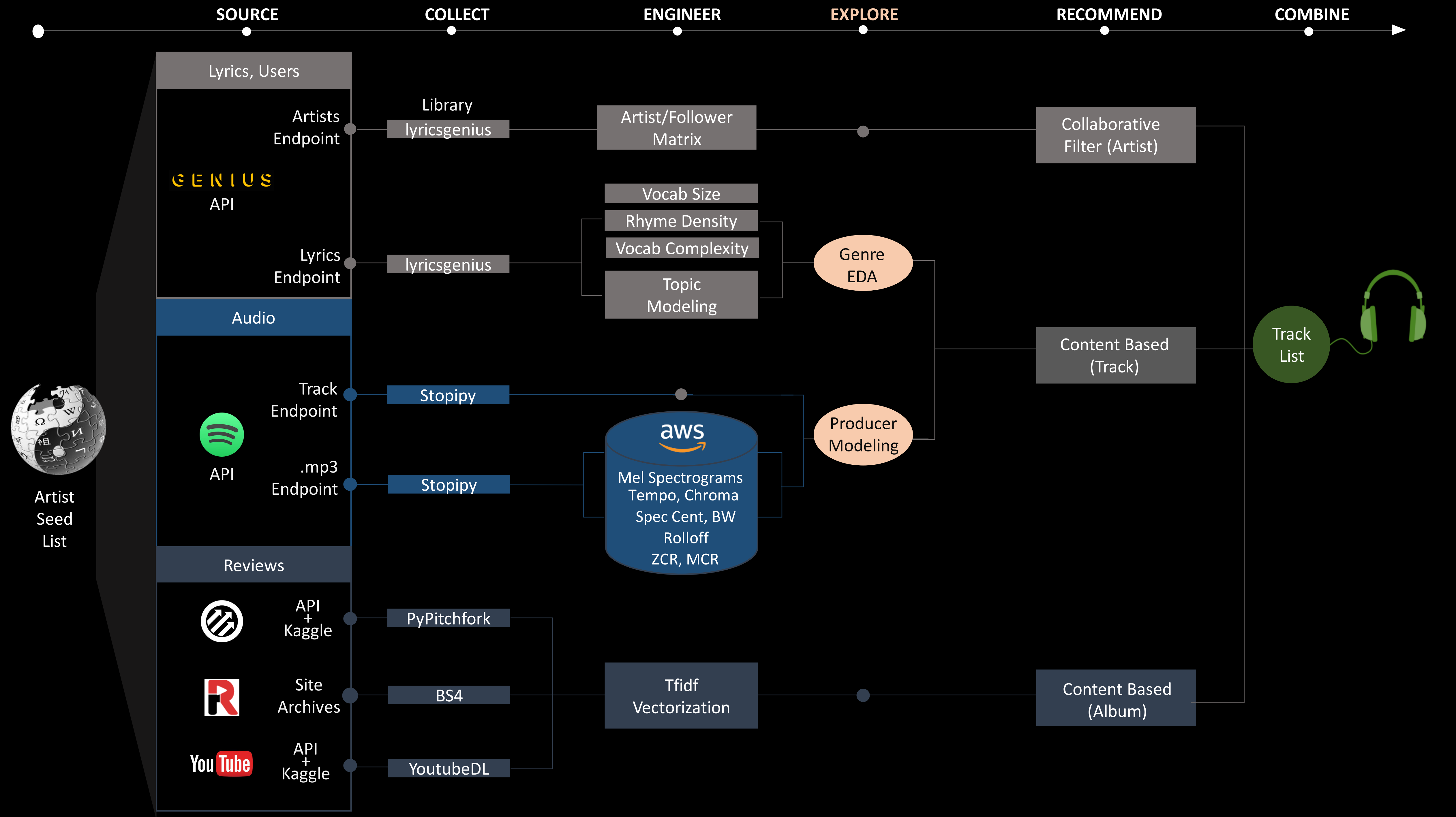
```
Hello! You are currently listening to...

ARTIST: ['SAMMY ADAMS']
ALBUM: ['BOSTONS BOY']
TRACK: ['COAST 2 COAST']

We recommend checking out...
```

	artist	album	track	follower flag	review flag
0	SADAT X	WILD COWBOYS	THE FUNKIEST	1.0	0.0
1	N.O.R.E.	CRACK ON STEROIDS	LEHHHGOOO	1.0	0.0
2	RAH DIGGA	CLASSIC	YOU GOT IT	1.0	0.0
5	KOOL KEITH	BLACK ELVIS / LOST IN SPACE	THE GIRLS DONT LIKE THE JOB	1.0	0.0
3	CLYDE CARSON	S.T.S.A.	LET ME KNOW	0.0	0.0
4	DOE B	DEFINITION OF A TRAPPER 3	YOU DONT EVEN KNOW	0.0	0.0
6	SAMMY ADAMS	BOSTONS BOY	COAST 2 COAST	0.0	0.0
7	PROJECT PAT	LAYIN DA SMACK DOWN	CHOOSE U	0.0	0.0
8	CASEY VEGGIES	LIFE CHANGES	YOUNG WINNERS	0.0	0.0
9	AZEALIA BANKS	FANTASEA	NEPTUNE	0.0	0.0

Recommend: Findings and Next Steps

Findings	Next Steps
Audio, lyrics are insufficient on their own	Front End
User data is priority #1	Dimensionality Reduction (Clustering, PCA)
Relevance + Novelty = Serendipity	User Profiles
High dimensionality can adversely impact engine	A/B Testing

Appendix



SOURCE COLLECT ENGINEER EXPLORE RECOMMEND COMBINE

Artist Seed List

Lyrics, Users

GENIUS API

Artists Endpoint

Lyrics Endpoint

Audio

API

Track Endpoint

.mp3 Endpoint

Reviews

API + Kaggle

Site Archives

API + Kaggle

Library lyricsgenius

lyricsgenius

Stopipy

Stopipy

PyPitchfork

BS4

YoutubeDL

Artist/Follower Matrix

Vocab Size

Rhyme Density

Vocab Complexity

Topic Modeling

aws

Mel Spectrograms

Tempo, Chroma

Spec Cent, BW

Rolloff

ZCR, MCR

Tfidf Vectorization

Genre EDA

Producer Modeling

Collaborative Filter (Artist)

Content Based (Track)

Content Based (Album)

Track List

Summary

	Granularity	Basic Fields	Unique Fields
	Track	• Track Title • Track ID • Artist Name • Artist ID • Album Name • Album ID • Date	• Spotify Proprietary Audio Features • Speechiness • Liveness • Acoustiness • Danciness • Etc…
GENIUS Lyrics	**Track**	• Track Title • Track ID • Artist Name • Artist ID • Album Name • Album ID	• Song Lyrics • Producers
GENIUS User Data	**Artist**	• Artist ID • Artist Name	• Follower ID • Follower Name
You Tube	**Album**	• Artist Name • Album Name	• Review Text • Review Score

Area of Focus	Questions	Learning
Scraping	*Is Selenium an effective tool for bespoke scraping needs? What are the pros and cons?	Selenium is a highly customizable tool for automated testing and more difficult scraping tasks. However, it can be slow and is susceptible to interruption (ads, site scripts).
Merging	*What are best practices for fuzzy matching?	A number of different distance-based metrics can be leveraged for fuzzy text matching. Selecting the right distance metric depends on context. To improve match rates, leverage these features to train a classificationmodel.
Hip Hop	*How has the genre changed over time? *Is hip hop "dead" as an artform?	It may be "dead" as a lyrical art form. Far less rappers are talking about their skills, focusing instead on "money" and "the hustle" in their raps. The genre has also seen a sizeable decline in rhyme density and vocab complexity.
NLP	*What is the most effective method for topic extraction? *What are the challenges?	Latent Dirichlect Allocation is the most popular approach here. This approach requires some subjectivity when it comes to settingthe number of topics and interpreting said topics. However, there are several scores (Coherence, Perplexity) that can assist.
Audio	*Can we classify production based on isolated audio features?	Production is usually a team effort, which makes this task more challenging to tackle. However, for single-producer tracks, we were able to improve our baseline accuracy from 4% to ~38%. Forour next iteration of this model, we'll look to leverage Mel Spectograms and a Convoolutional Neural Network.
Recommendations	*How do recommendations differ between collaborative filter vs content based filter?	In our implementation, content based recommendations produced more novel tracks. This is not surprising, as features like rhymedensity and positive sentiment can unearth less positive tracks.
Recommendations	*What are the major pitfalls for recommendation engines?	*Subjectivity: The idea that content cannot capture subjective info like humor, points of view (collaborative is better here) *Scalability: as features and the user base grows, the computational cost of recommendation algorithms grows significantly. *Sparsity: In particular with user-based features this is a problem. If you are leveraging user interactions with a specific page for site recommendations, for example, it may be that only 1% of users ever get to that page at all.
Recommendations	*What are some methods for hybrid recommendation engines?	TBD

Bucket	Requirements	Advantages	Disadvantages
Knowledge-Based	• Significant knowledge of the domain in question	• Minimal reliance on data or specific user preferences. • Approach for high complexity/barrier items	• Need extensive domain knowledge • Scalability a factor
Heirarchical	• Clean product/content management system	• Simple to implement (I bought a printer, I need ink)	• Less effective for more general use cases
Content Based Filtering	• Clean and consistent metadata at the item level • Optimal: user preference data	• No domain knowledge required • No community required	• Overengineering possible, leading to **unexpected** results • Best when aligned to user preferences
Collaborative Filtering	• Large historical data set • Set of users and preferences	• No domain knowledge required • Simple implementations can be effective	• Sparsity of Data • Cold-Start User Problem

Bucket	Requirements	Advantages	Disadvantages
Knowledge-Based	• Significant knowledge of the domain in question	• Minimal reliance on data or specific user preferences. • Approach for high complexity/barrier items	• Need extensive domain knowledge • Scalability a factor
Heirarchical	• Clean product/content management system	• Simple to implement (I bought a printer, I need ink)	• Less effective for more general use cases
Content Based Filtering	• Clean and consistent metadata at the item level • Optimal: user preference data	• No domain knowledge required • No community required	• Overengineering possible, leading to **unexpected** results • Best when aligned to user preferences
Collaborative Filtering	• Large historical data set • Set of users and preferences	• No domain knowledge required • Simple implementations can be effective	• Sparsity of Data • Cold-Start User Problem