Архитектура и алгоритмы для индексации всей музыки ВКонтакте

Алексей Акулович
vk.com

HL 2016 HighLoad++

Профессиональная конференция разработчиков высоконагруженных систем

Сколько музыки в ВК?

Сколько музыки в ВК?

Сколько музыки в ВК?

Сколько музыки в ВК?

Сколько музыки в ВК?

Сколько музыки в ВК?

Сколько музыки в ВК?

400

миллионов треков

4 ПБ файлов

+150к (1.5 ТБ) в день

Есть что послушать

Что с этим можно делать?

Найти в поиске по названию из ID3 тегов

Что с этим можно делать?

Найти в поиске по названию из ID3 тегов

Дубли

Что с этим можно делать?

Найти в поиске по названию из ID3 тегов

Дубли

Несоответствие названий и содержимого

Что с этим можно делать?

Найти в поиске по названию из ID3 тегов

Дубли

Несоответствие названий и содержимого

И всё

Что хочется уметь делать

Фильтровать дубли в поиске

Что хочется уметь делать

Фильтровать дубли в поиске

Предлагать варианты лучшего качества

Что хочется уметь делать

Фильтровать дубли в поиске

Предлагать варианты лучшего качества

Обложки, тексты, исполнители, …

Что хочется уметь делать

Фильтровать дубли в поиске

Предлагать варианты лучшего качества

Обложки, тексты, исполнители, …

Легализация

Что хочется уметь делать

Фильтровать дубли в поиске

Предлагать варианты лучшего качества

Обложки, тексты, исполнители, …

Легализация

А ещё быть устойчивым к

Перекодированию (смена битрейта и т.п.)

А ещё быть устойчивым к

Перекодированию (смена битрейта и т.п.)

Добавлению тишины/шума в начало/конец

А ещё быть устойчивым к

Перекодированию (смена битрейта и т.п.)

Добавлению тишины/шума в начало/конец

Убиранию небольшой части из начала/конца

Возьмём готовое решение!

Что вышло?

Решение лишь создавало отпечатки

Что вышло?

Решение лишь создавало отпечатки

Поиск по ним мы сочинили свой

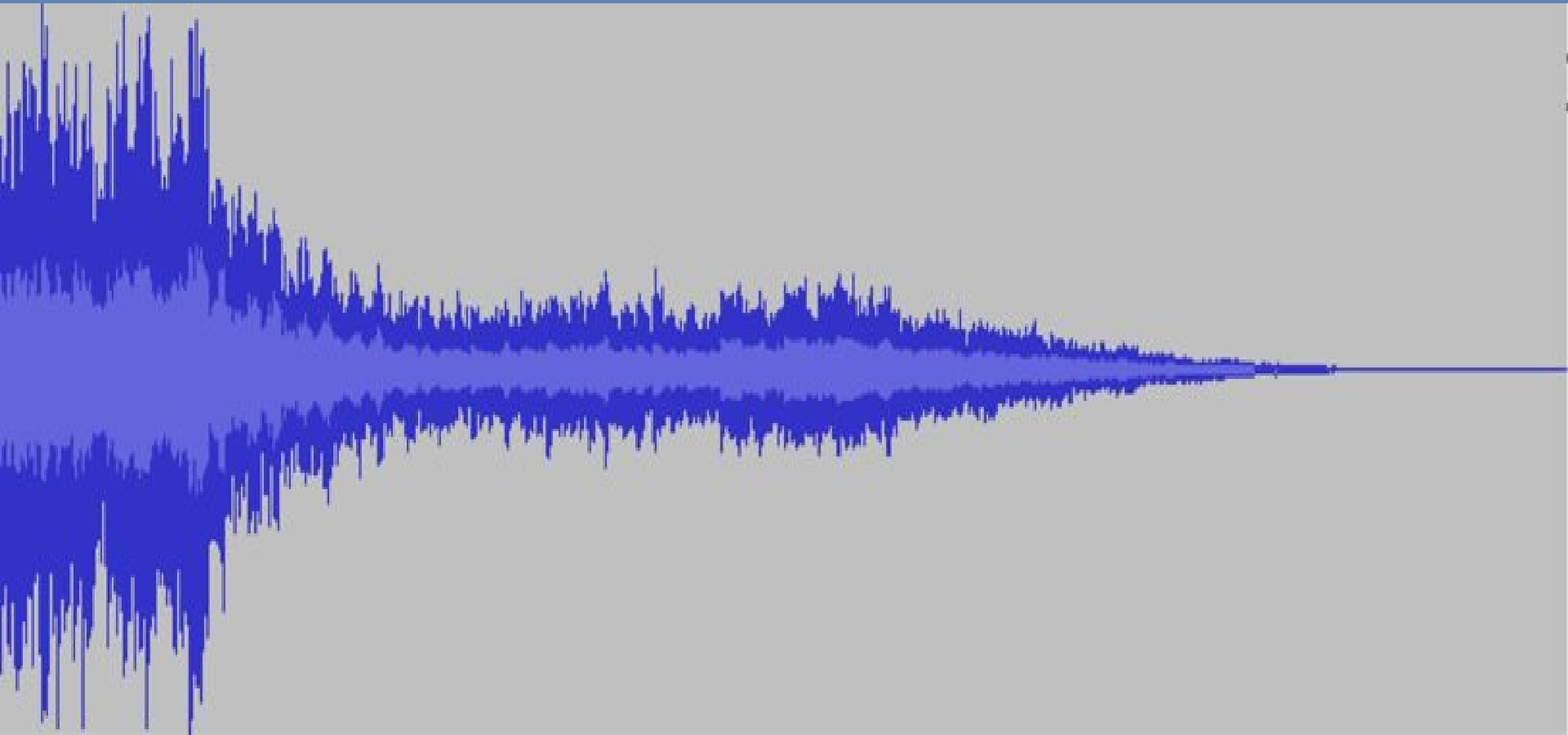
Что вышло?

Решение лишь создавало отпечатки

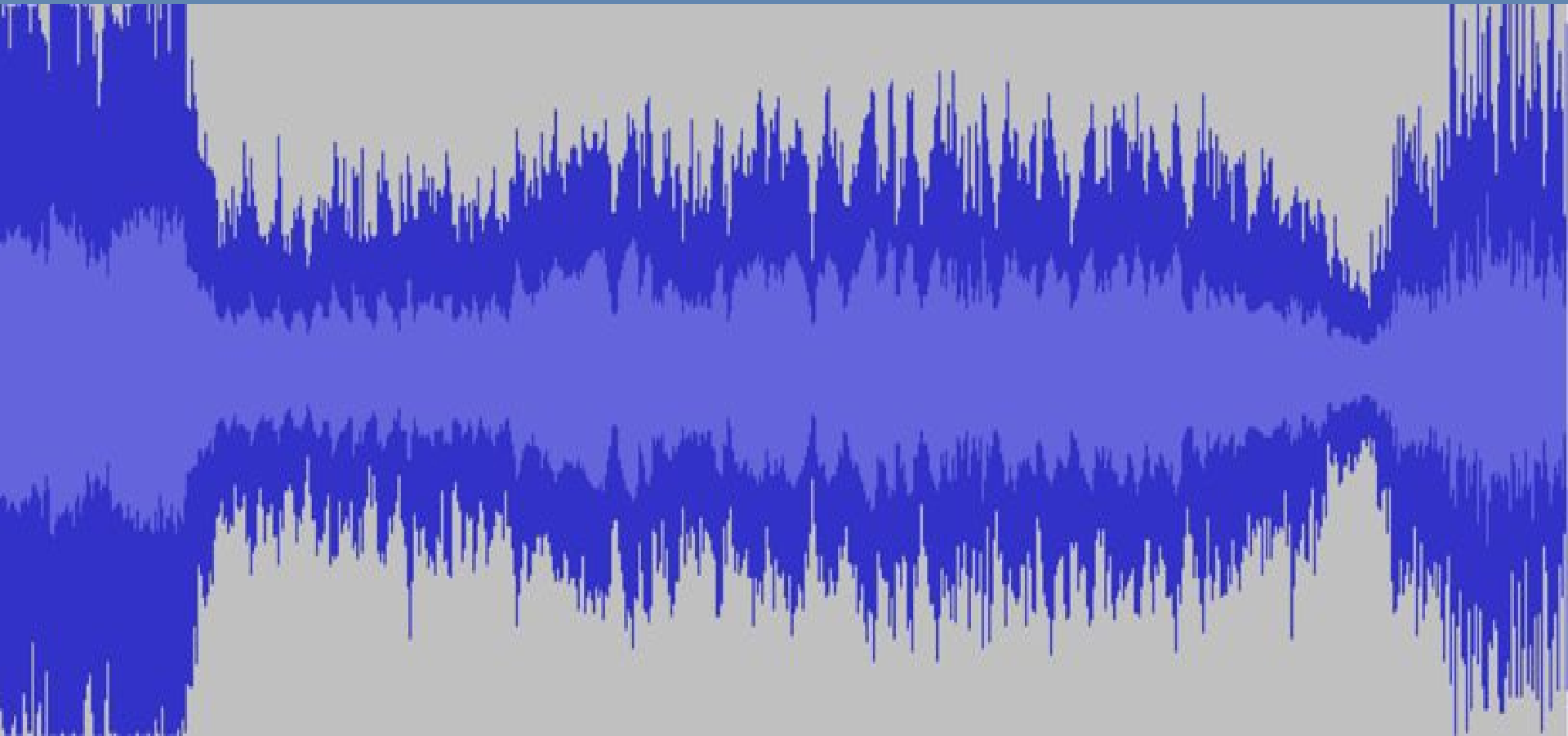
Поиск по ним мы сочинили свой

В *целом* работало на тестах

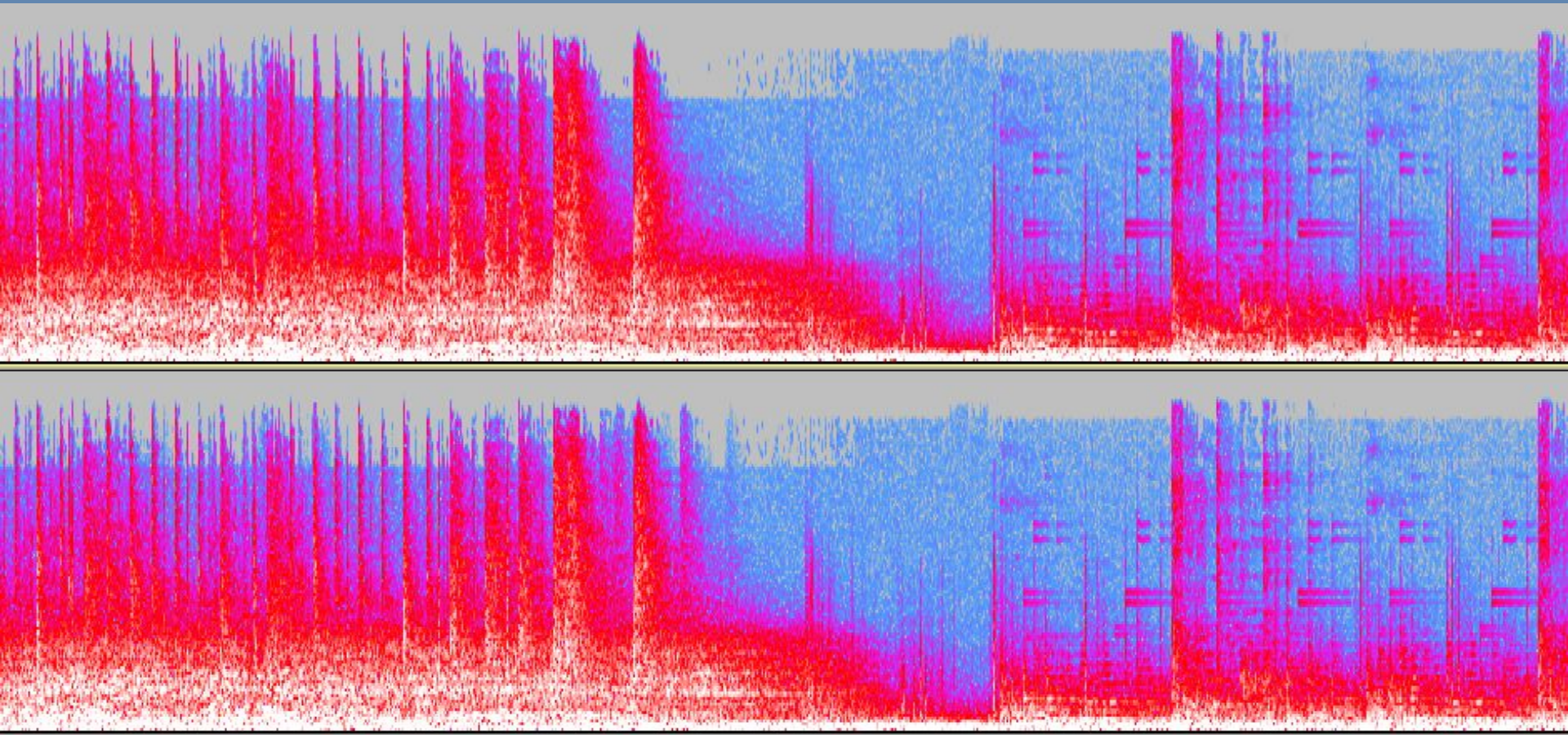
Доп. требования: live, радио, подкасты



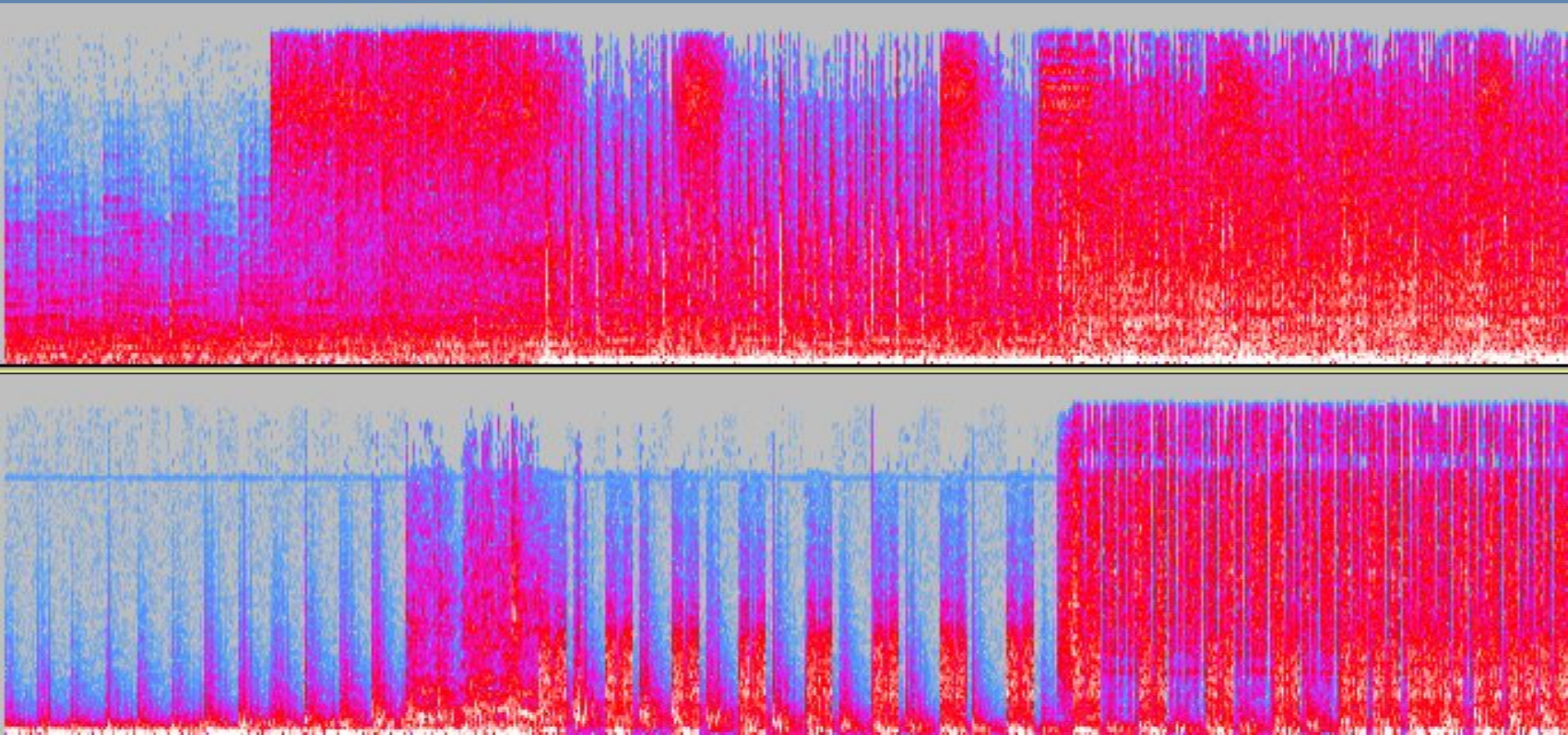
Доп. требования: live, радио, подкасты



Доп. требования: эквалайзер



Доп. требования: склейки, каверы



Доп. требования: зашумленный фрагмент

Доп. требования: не вышло

Что делать?

Что делаем?

Из файла достать аудио фреймы

Что делаем?

Из файла достать аудио фреймы

Совершить некую магию

Что делаем?

Из файла достать аудио фреймы

Совершить некую магию

Получить некие данные

Что делаем?

Из файла достать аудио фреймы

Совершить некую магию

Получить некие данные

Как-то сравнивать файлы по этим данным

Декодирование MP3

MP3 на вход → аудио данные на выходе

Библиотека libmad[1] + свой враппер на Go[2]

1 — http://www.underbit.com/products/mad/
2 — https://github.com/AterCattus/fennec-tiny (файл mad.go)

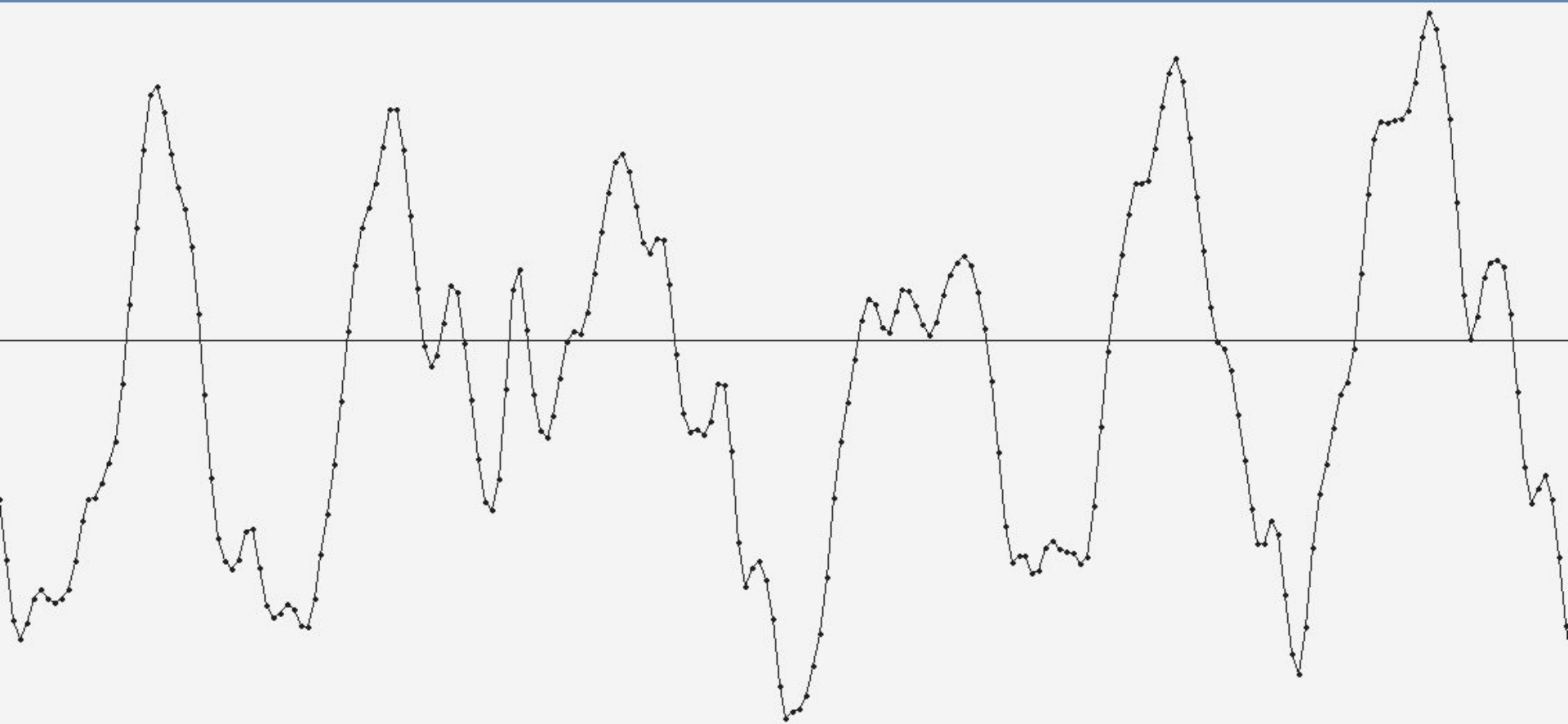
Декодирование MP3

MP3 на вход → аудио данные на выходе

Библиотека libmad[1] … пер на Go[2]

1 — http://www.underbit.com/products/mad/
2 — https://github.com/AterCattus/fennec-tiny (файл mad.go)

Декодирование MP3: амплитуда



Что делаем?

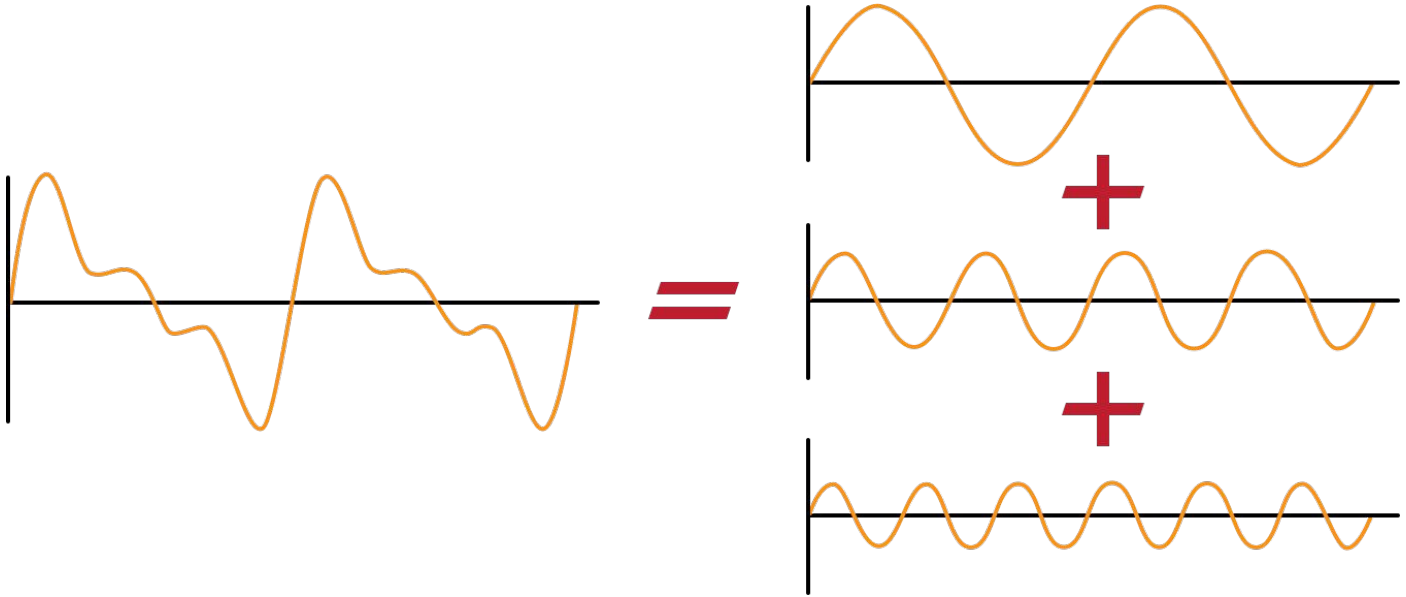
~~Из файла достать аудио фреймы~~

Совершить некую магию

Получить некие данные

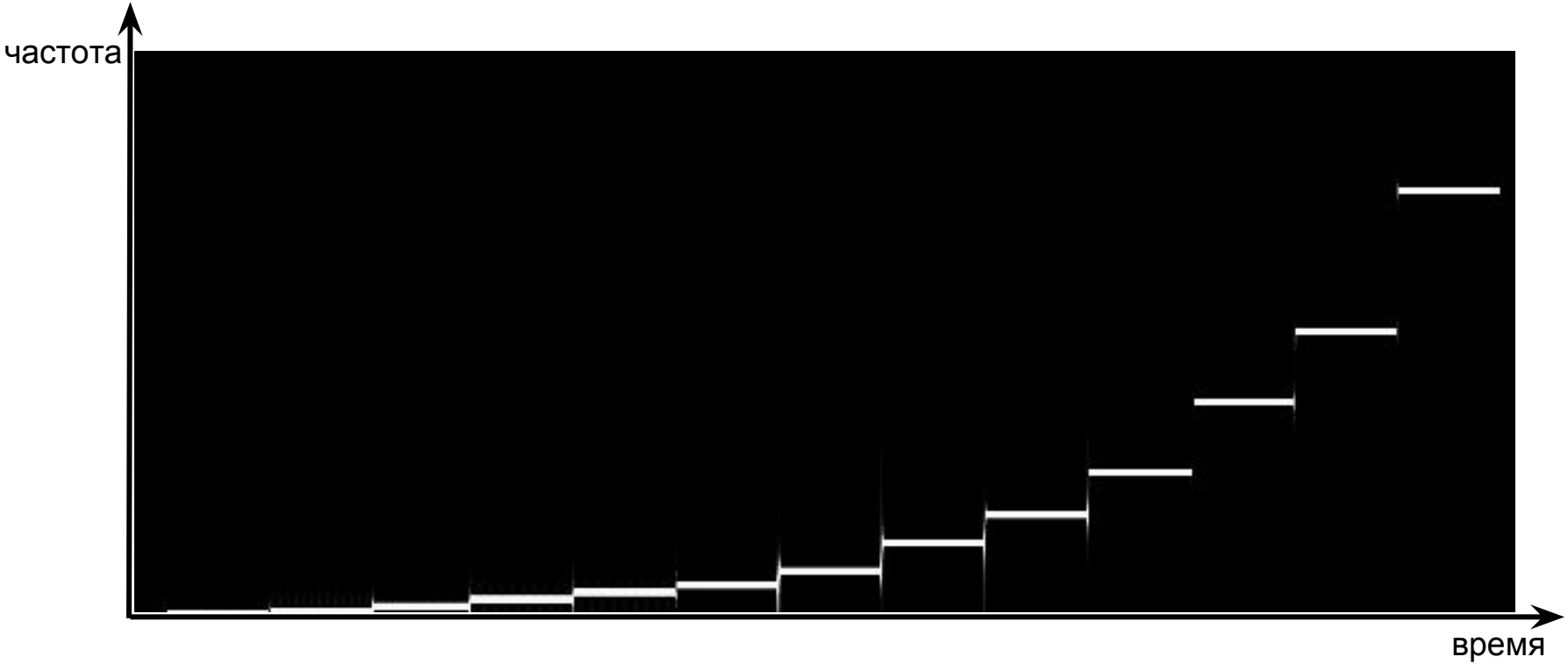
Как-то сравнивать файлы по этим данным

Звуковая волна

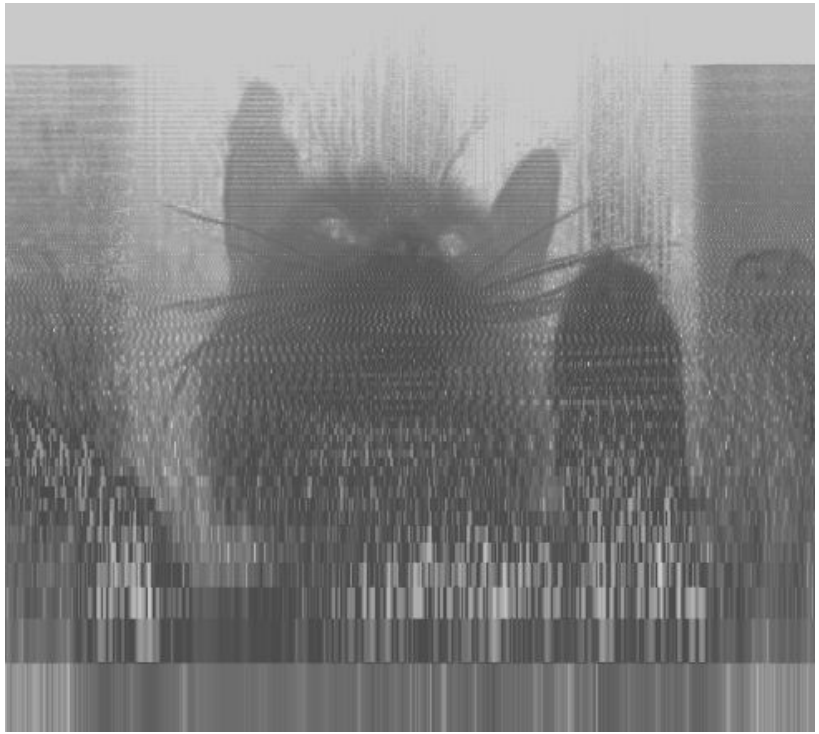


Преобразование Фурье

Спектрограмма



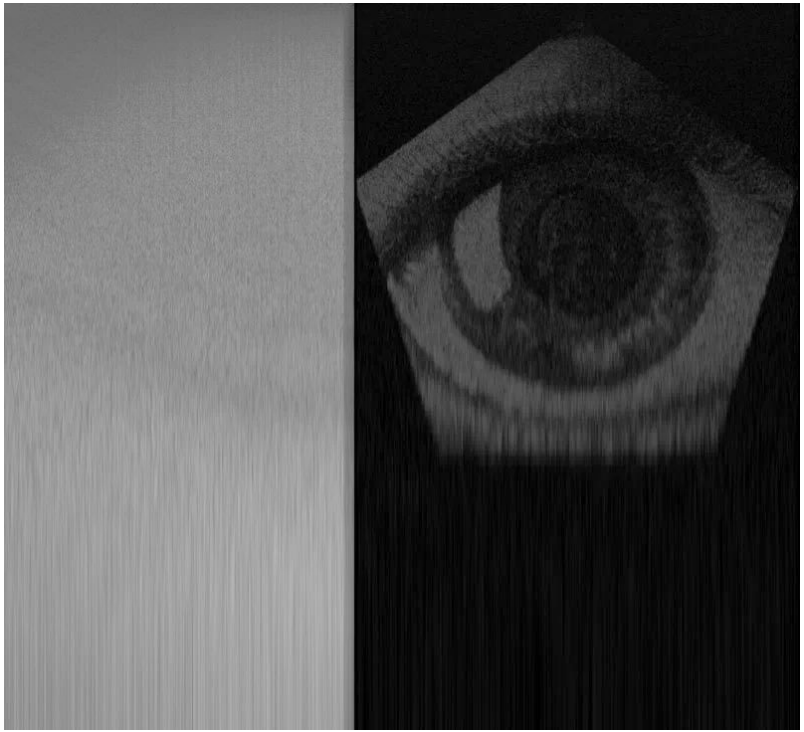
Спектрограмма



Songs About My Cats

Venetian Snares

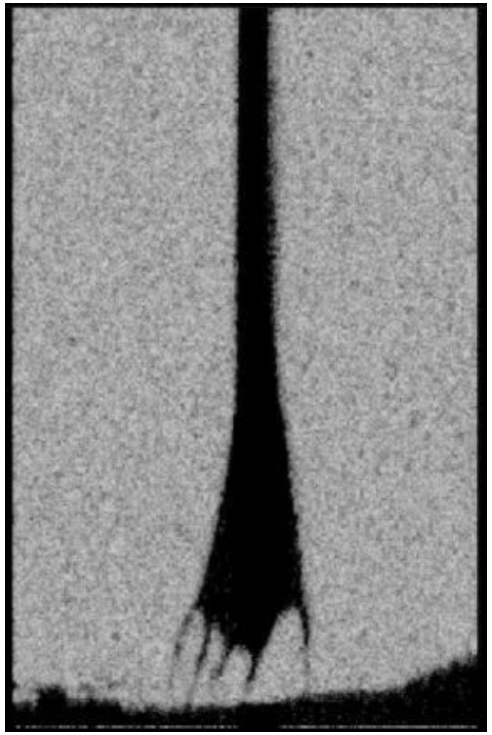
Спектрограмма



Continuum

Disasterpeace (FEZ Soundtrack)

Спектрограмма



My Violent Heart

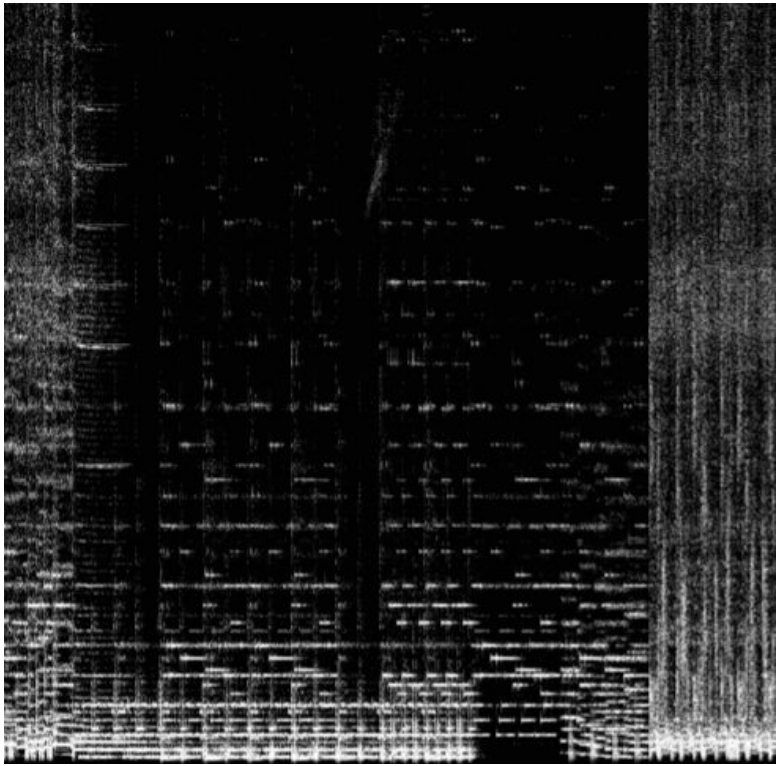
Nine Inch Nails

Спектрограмма

Transitions

DJ Sonix

Спектрограмма



Обычная песня

Займет 10 ПБ

А было 4 ПБ

Что делаем?

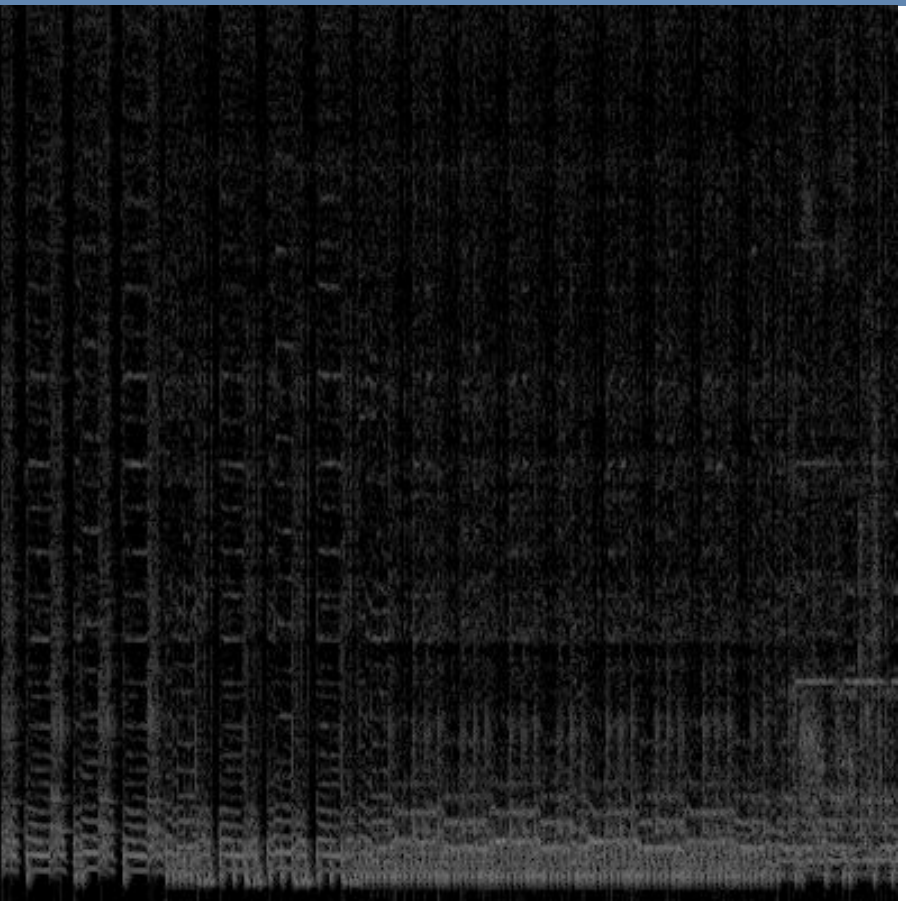
~~Из файла достать аудио фреймы~~

~~Совершить некую магию~~

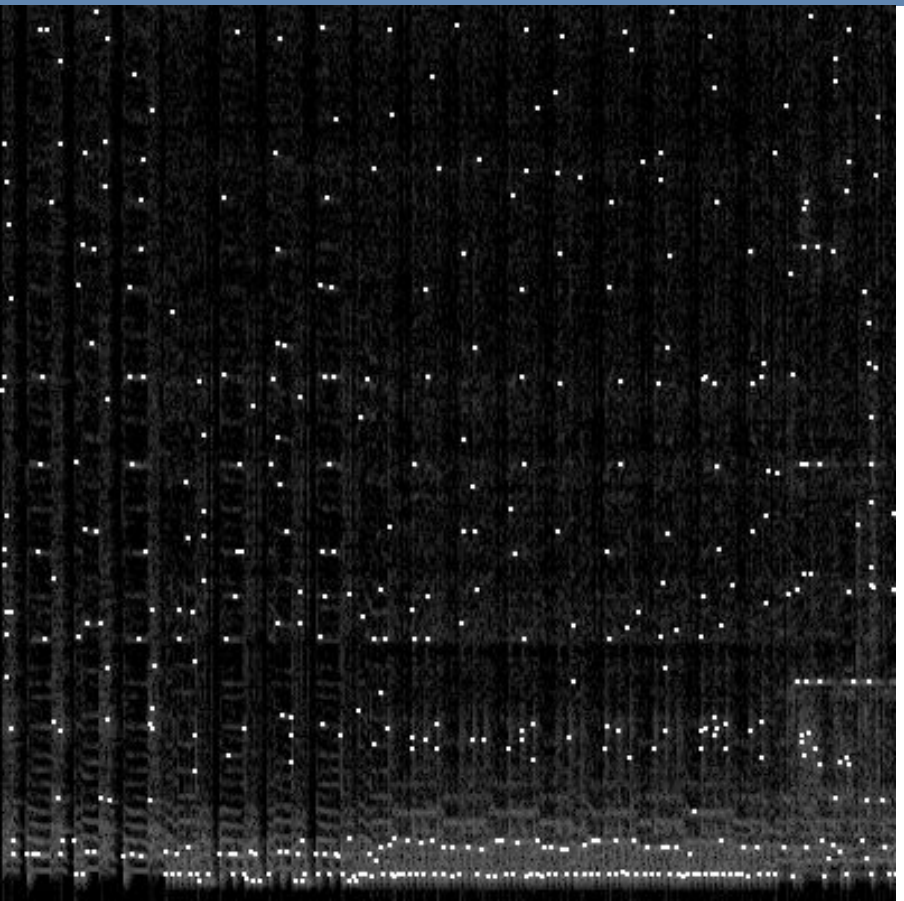
Получить некие данные

Как-то сравнивать файлы по этим данным

Уменьшение объема данных

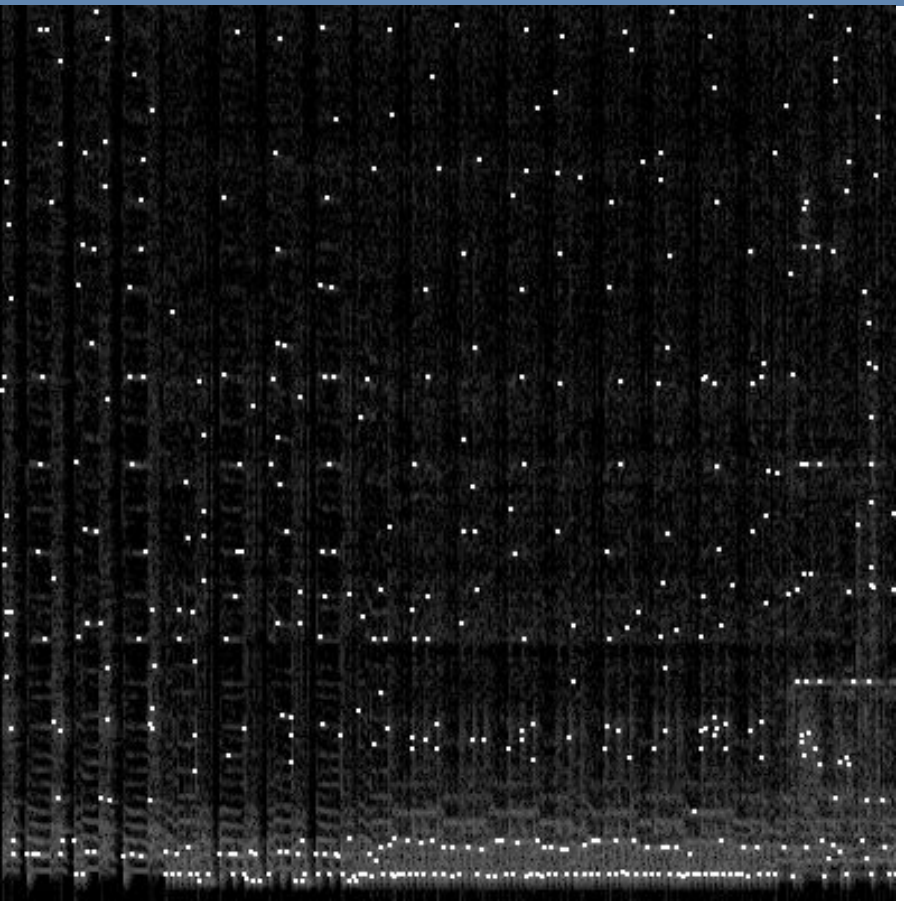


Уменьшение объема данных



В **200** раз
меньше данных

Уменьшение объема данных



Пики это (время, частота)

Время → uint32
Частота → uint32
(время, частота) → uint64

Отпечаток: массив uint64

Что делаем?

~~Из файла достать аудио фреймы~~

~~Совершить некую магию~~

~~Получить некие данные~~

Как-то сравнивать файлы по этим данным

Сравнение двух треков

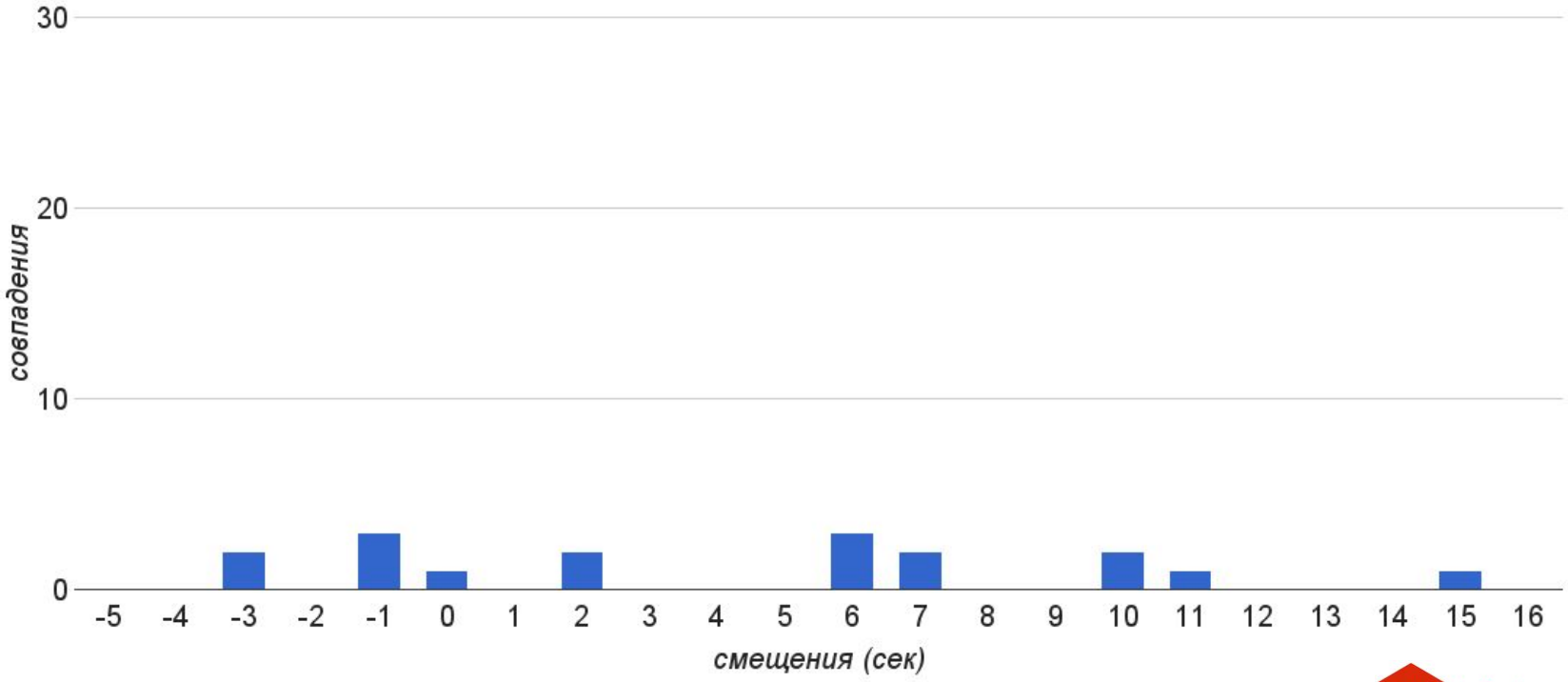
Ищем **временной сдвиг**

одного трека относительно другого

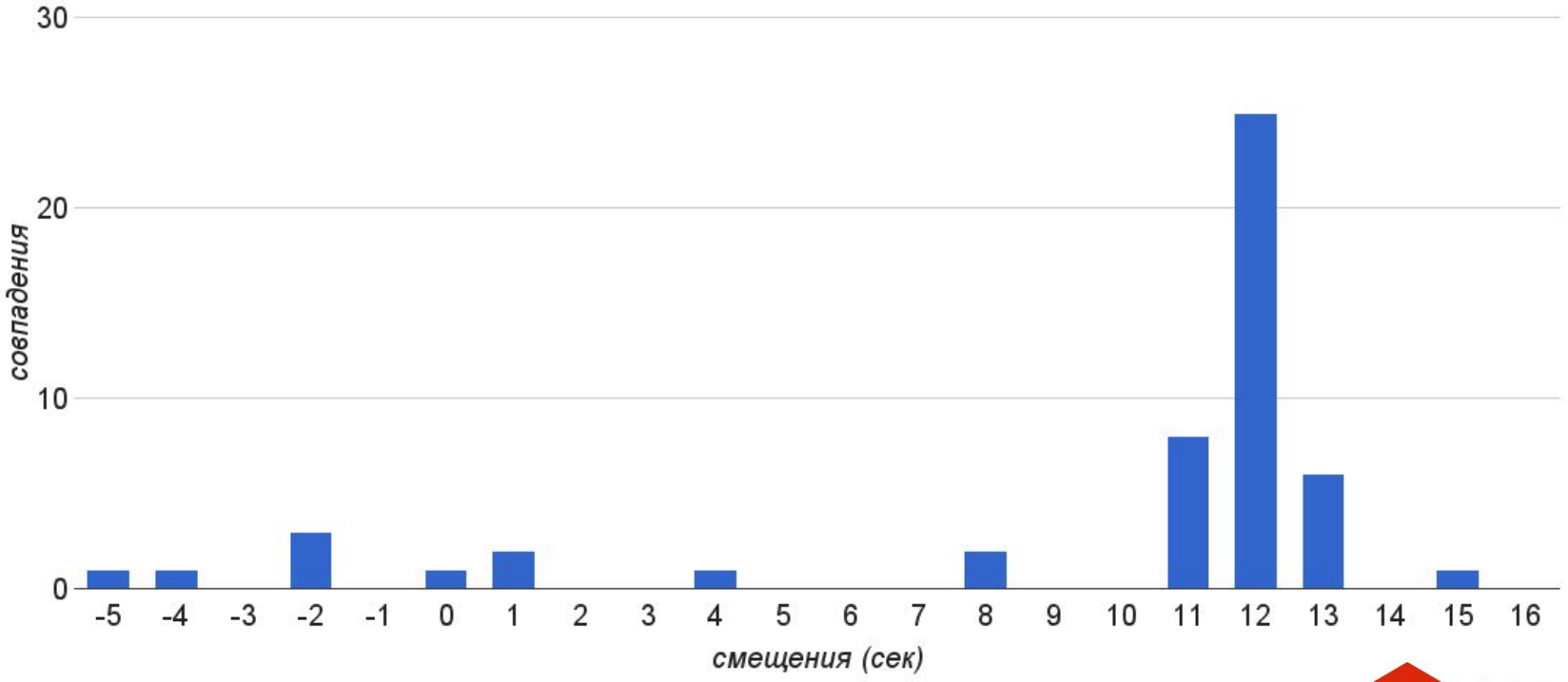
при котором получается

максимум совпадений частот

Сравнение треков: разные



Сравнение треков: общий фрагмент



Что делаем?

~~Из файла достать аудио фреймы~~

~~Совершить некую магию~~

~~Получить некие данные~~

~~Как-то сравнивать файлы по этим данным~~

Можно попробовать самим

Всё ранее описанное общедоступно

https://github.com/AterCattus/fennec-tiny

Fennec SK042 Vulpes zerda Art Print by S-Schukina

Архитектура

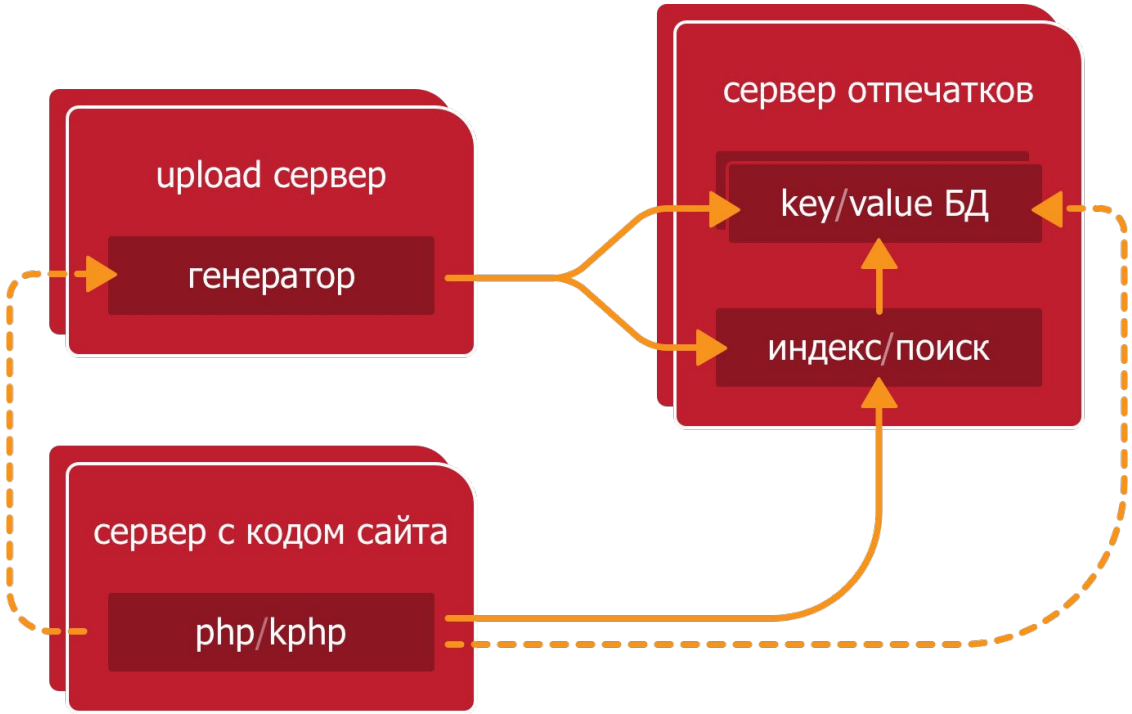
20 ТБ отпечатков

Как и где их все хранить?

Как по ним искать?

Кто подставил кролика Роджера?

Архитектура



Движок генерации отпечатков

На каждом upload сервере

Движок генерации отпечатков

На каждом upload сервере

Многопоточность

по горутине на файл

Движок генерации отпечатков

На каждом upload сервере

Многопоточность

по горутине на файл

Обработка песни:

типичная - 2-4 секунды

аудиокнига - линейно больше

Движок индексирования и поиска

Прореженные индексы

20 ТБ →100 ГБ

Да, возможны потери

Движок индексирования и поиска

Обратные индексы в памяти

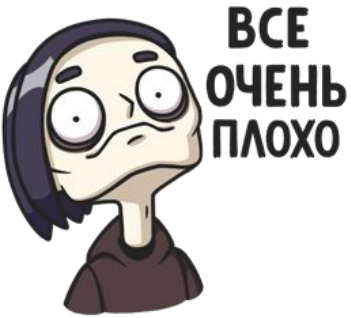
Хэши	ID треков					
0	A	C	E	M	W	X
1	B	D	N	O	S	Q
3	C	G	I	M	P	R
7	A	D	E	K	O	W
100500	F	K	P	R	V	Z

Движок индексирования и поиска

Насколько быстро движок работает?

Движок индексирования и поиска

Прогноз: 12 месяцев



Движок индексирования и поиска

Был прогноз: 12 месяцев

Повсеместное использование sync.Pool

Получили: 10 месяцев

Движок индексирования и поиска

Был прогноз: 10 месяцев

Использование container/heap

Получили: 6 месяцев

Движок индексирования и поиска

Был прогноз: 6 месяцев

Специализация container/heap

Получили: 5 месяцев

Движок индексирования и поиска

Был прогноз: 5 месяцев

Свой container/heap

Получили: 3 месяцев

Движок индексирования и поиска

Был прогноз: 3 месяцев

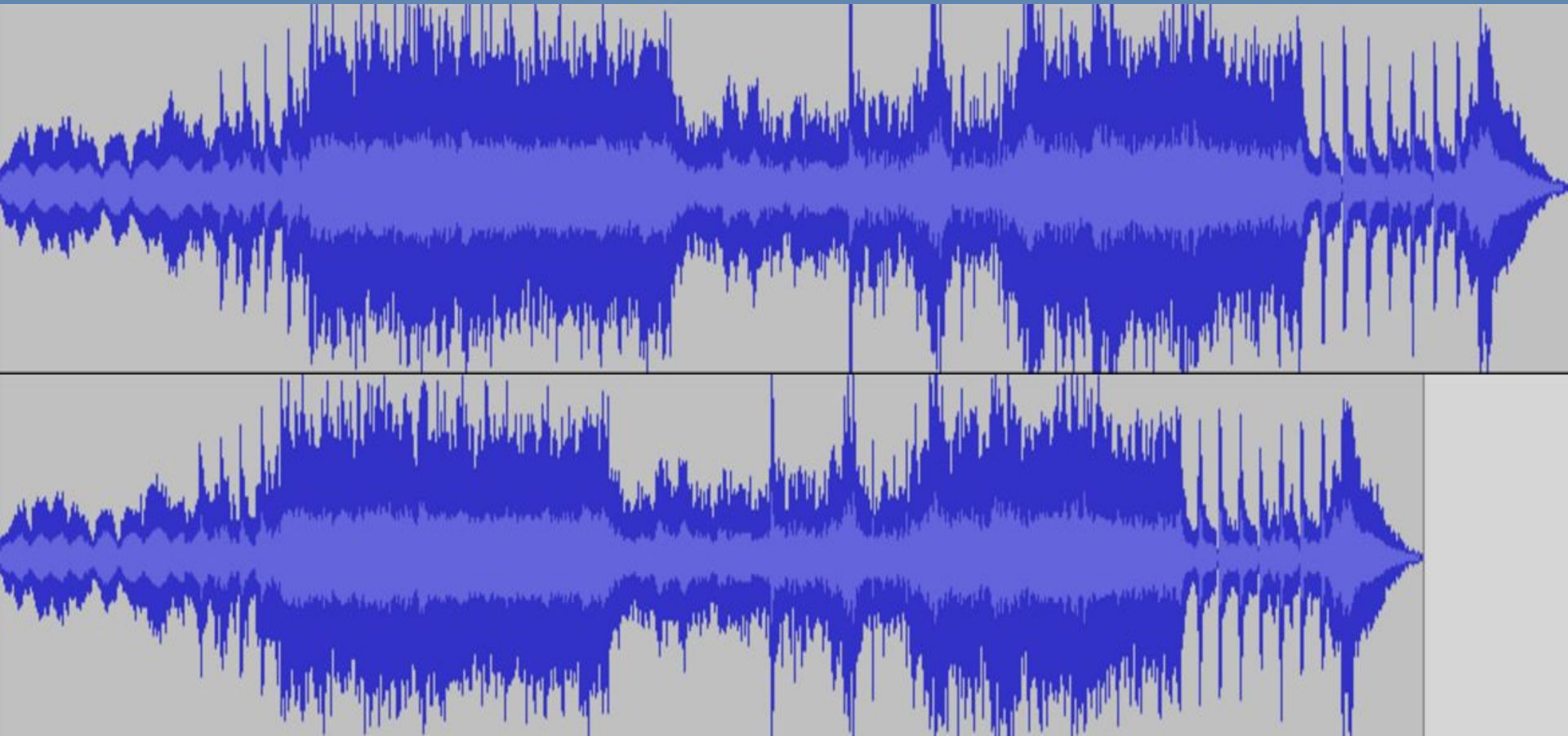
Обновления Go 1.5 → 1.6.2

Получили: 2.5 месяцев

Production тестирование

Ooops!

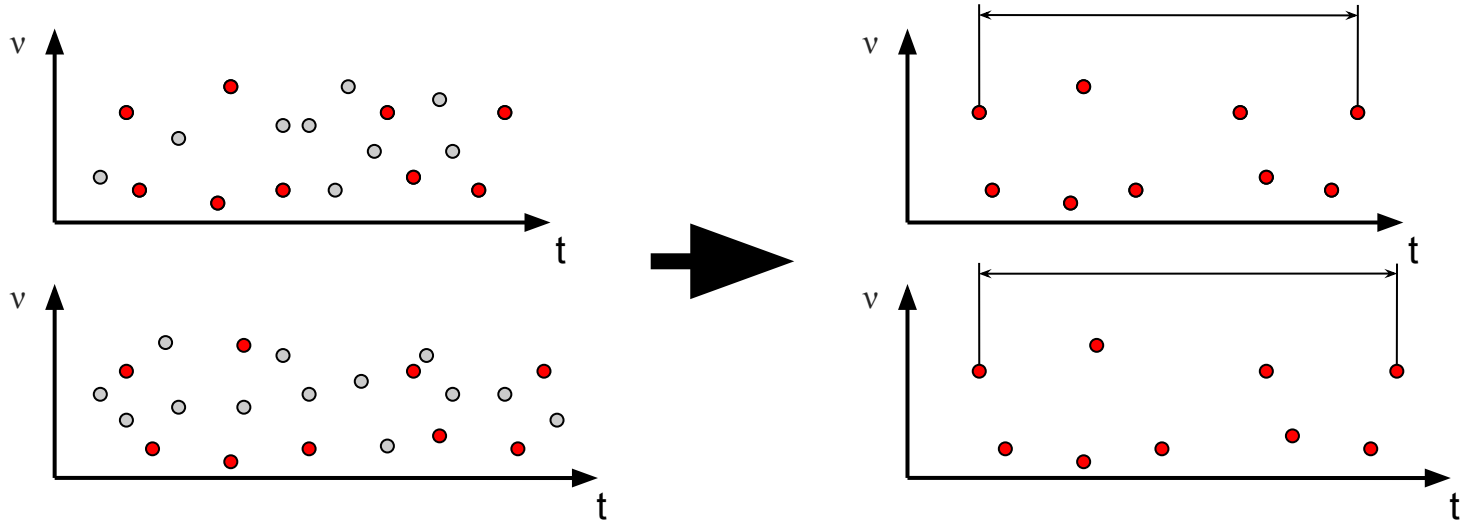
Ooops! Измененная скорость



Ooops! Измененная скорость

Наибольшая общая подпоследовательность

LCS (longest common subsequence)



Ooops! Отличия во фрагментах

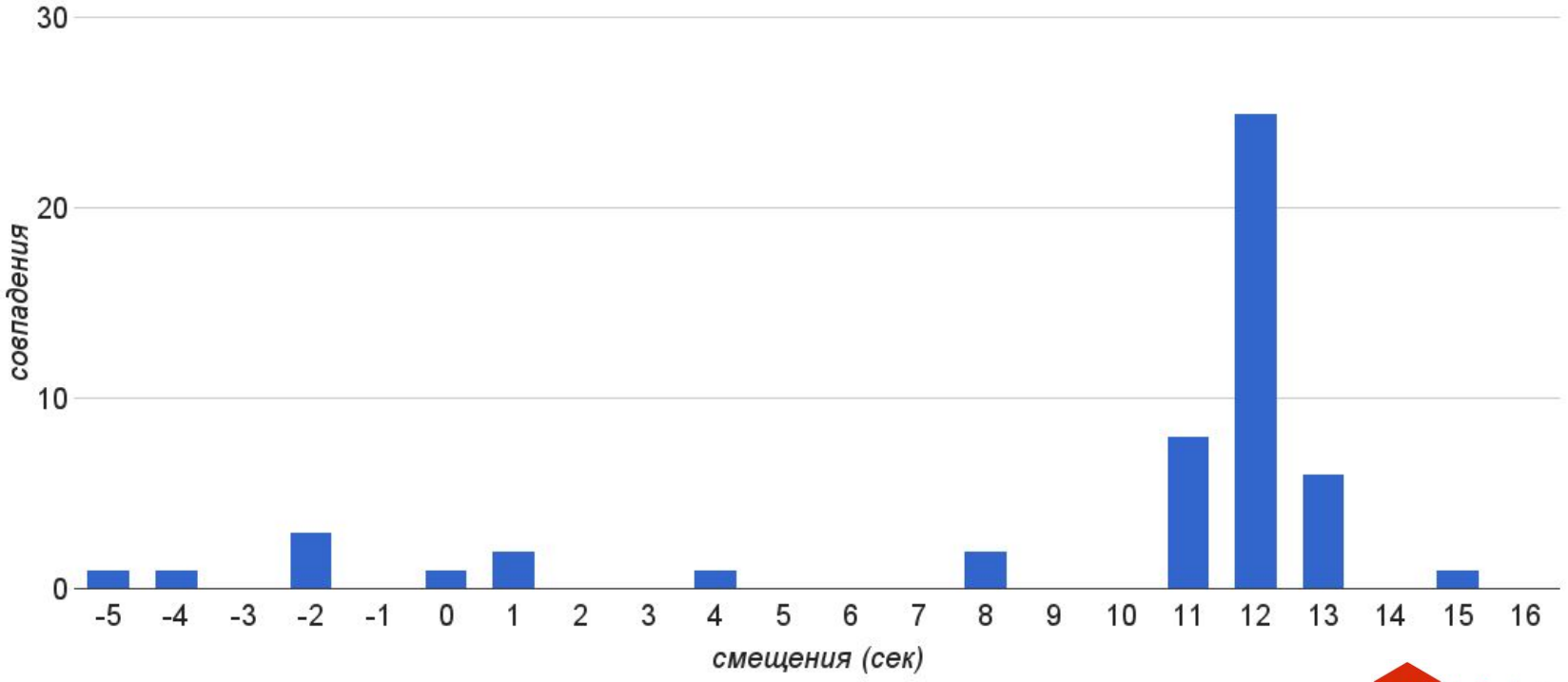
Миксы и склейки

Версии на разных языках

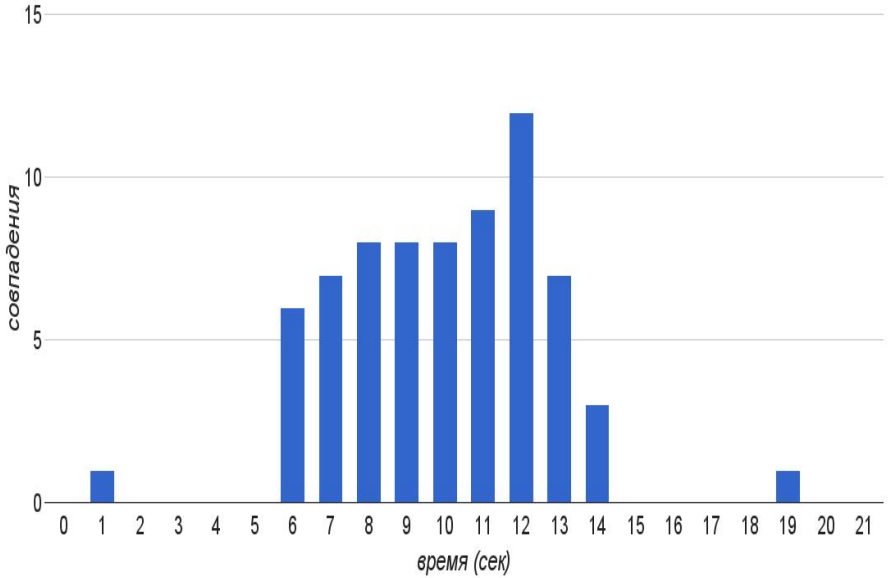
«Happy birthday»

Рэп «школьников»

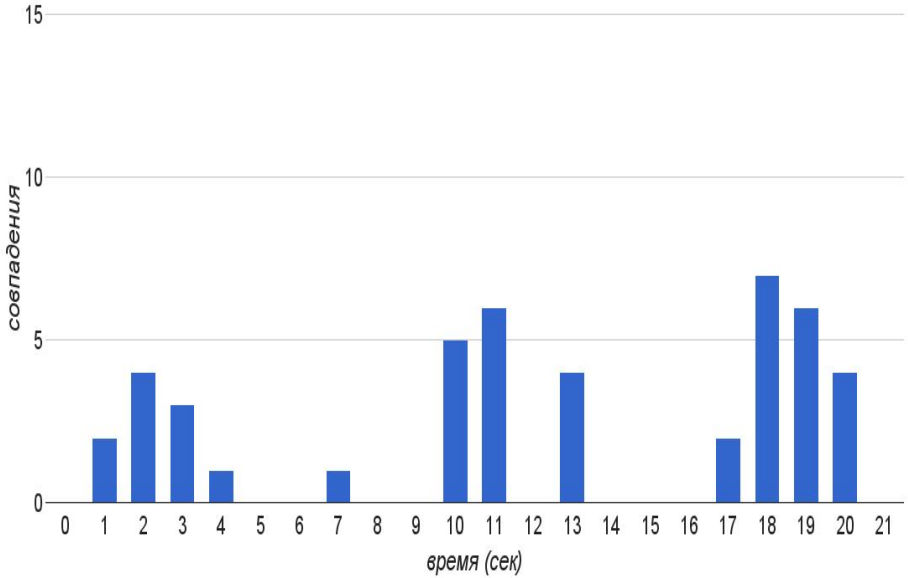
Ooops! Отличия во фрагментах



Ooops! Отличия во фрагментах



Одно совпадение



Несколько совпадений

Стресс-тест

- EQ changes Lowpass filter with 6dB octave roll-off starting at 10kHz
- Low bitrate-sample rate MP3 encoding with standard ffmpeg settings at 64kbps CBR 8kHz with an exception for 16kHz
- Phase shifting or inversion between channels
- Pitch shifting ± 2% at launch working to improve with an aim of reaching ±5% over the course of the agreement
- Speed changes 1.05x playback at launch, working to improve with an aim of reaching 1.25x over the course of the Agreement, subject to mutual agreement between the Parties
- The addition of silence, voice-over, or other non-program material at the beginning or end of the track

Не читай - презентацию скачай :)

Итоги

«Ничто не вечно, ничто не закончено
и ничто не совершенно»

Вот и всё

Презентация:
https://ater.me/conf/hl2016_audfp.pdf

Для вопросов после:
https://vk.com/ac