Adam Kilgarriff Keynote

The Right Rhymes

SMART LEXICOGRAPHY IN FULL EFFECT ?

Matt Kohl

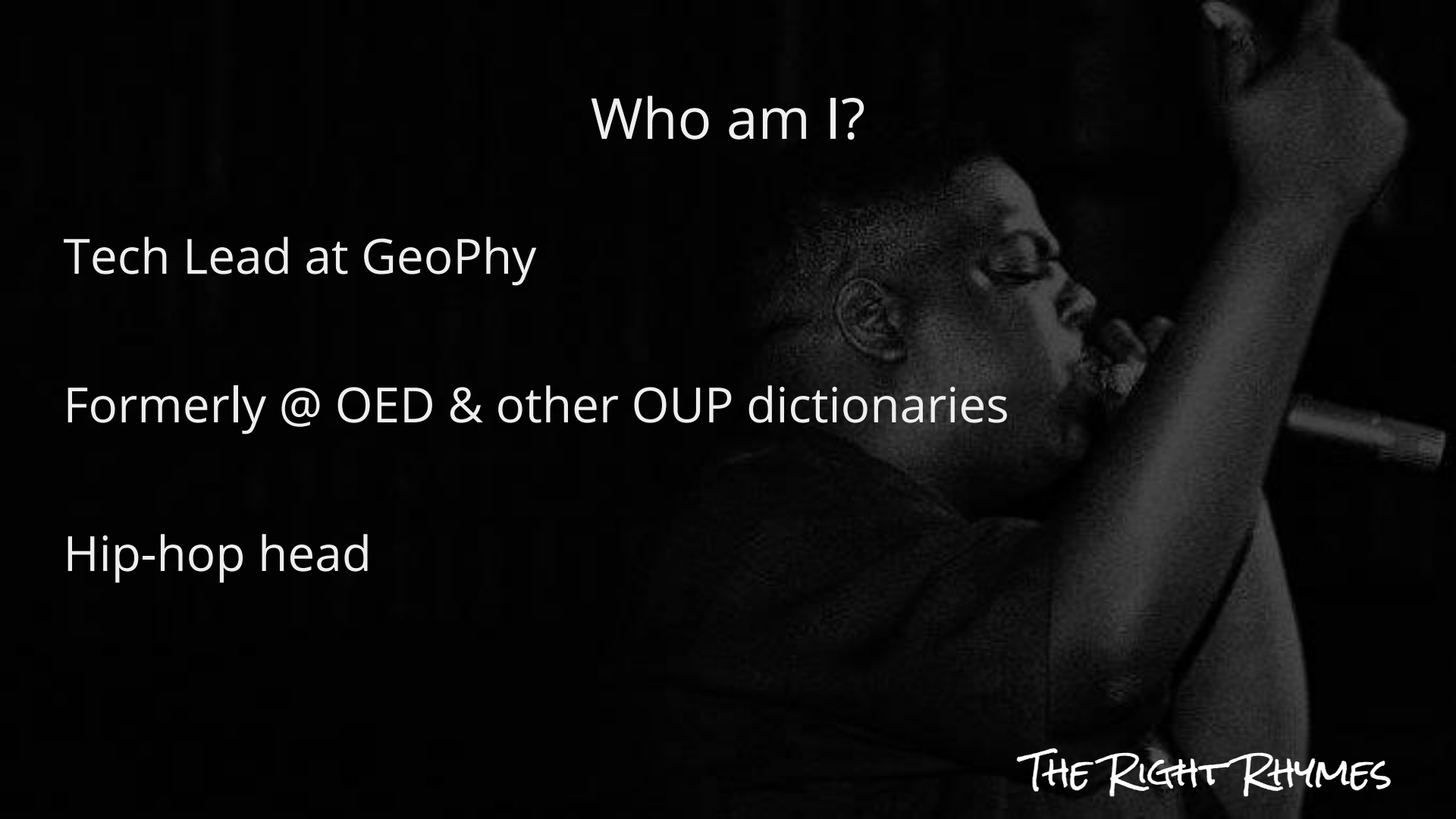
THE RIGHT RHYMES

Who am I?

Tech Lead at GeoPhy

Formerly @ OED & other OUP dictionaries

Hip-hop head



Who are you?
Lexicographers?
Technologists?
Hip-hop music lovers?
THE RIGHT RHYMES

Agenda

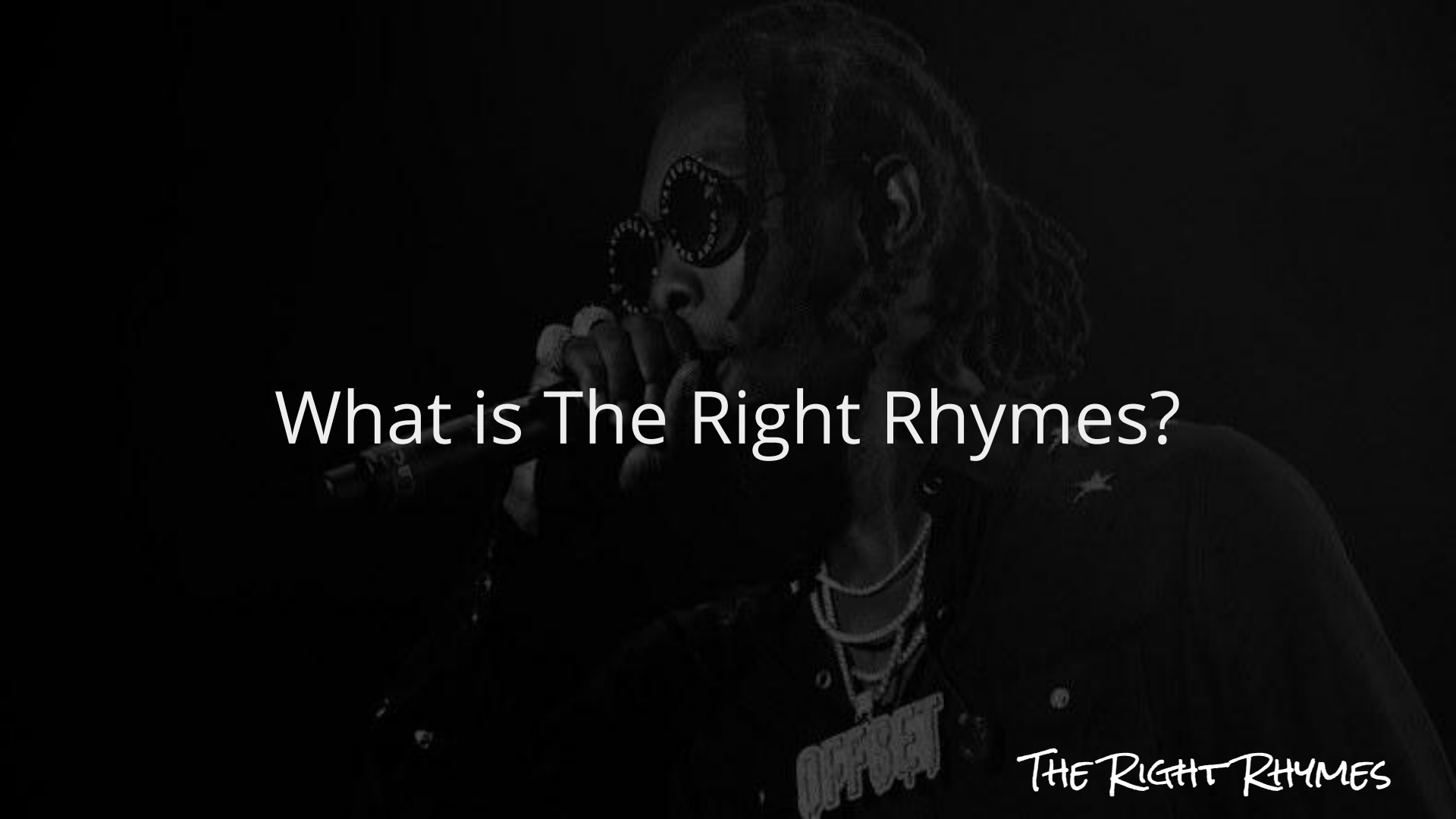
What is the Right Rhymes?

Peeking under the hood

Writing the dictionary

My latest stab at a lyrics corpus

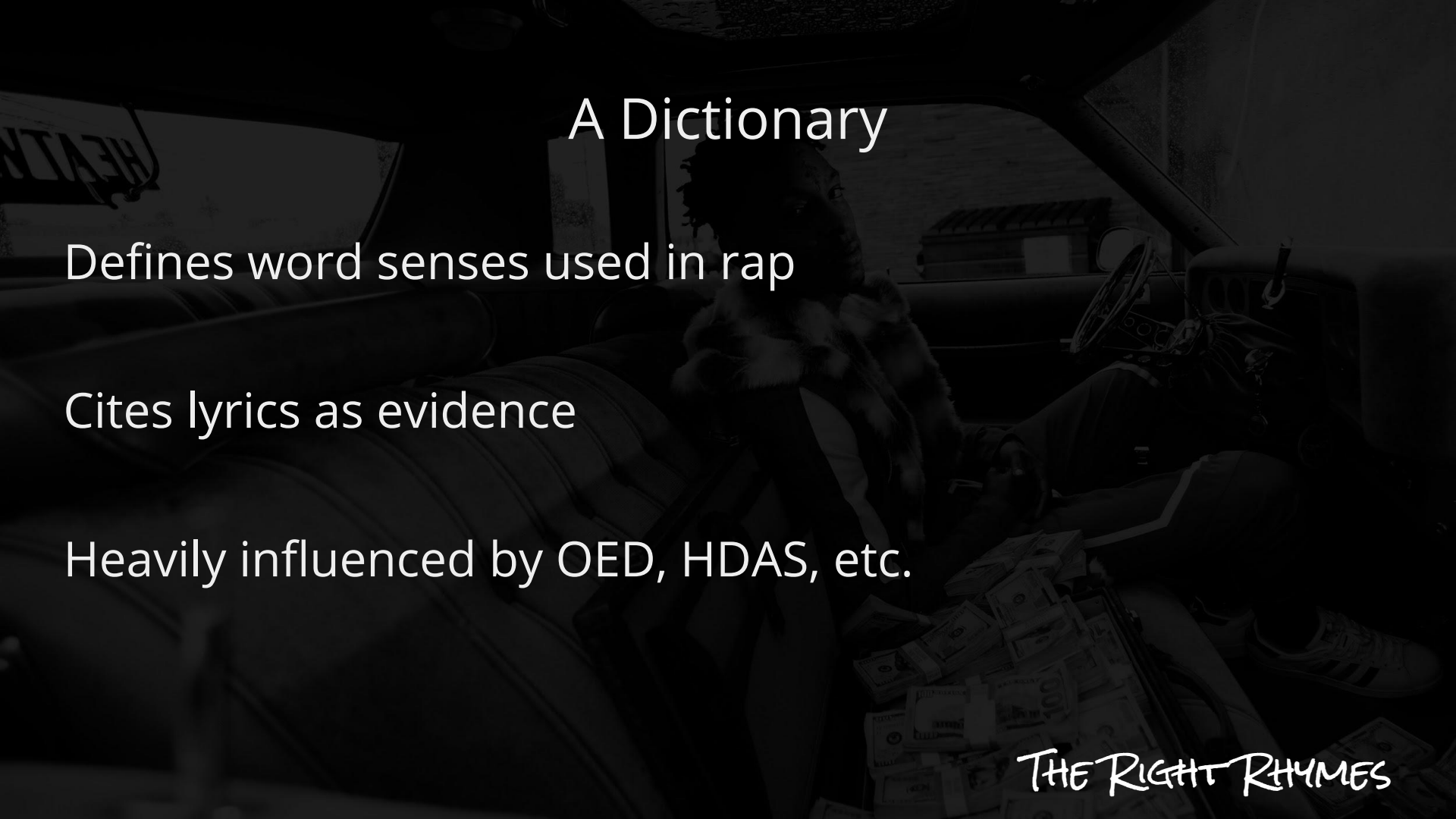
THE RIGHT RHYMES



What is The Right Rhymes?

THE RIGHT RHYMES

Some motivation
THE RIGHT RHYMES



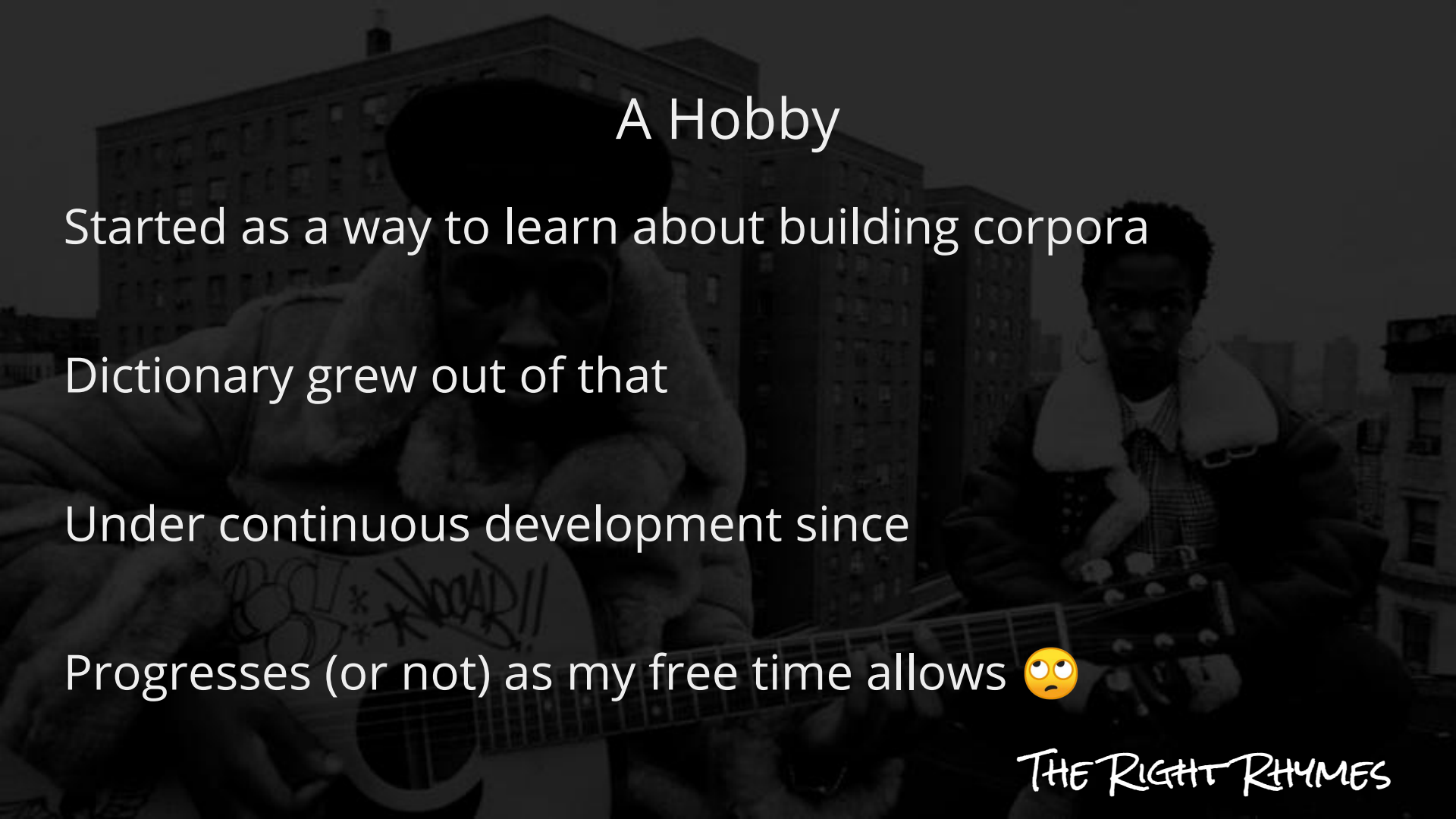
A Dictionary

Defines word senses used in rap

Cites lyrics as evidence

Heavily influenced by OED, HDAS, etc.

THE RIGHT RHYMES



A Hobby

Started as a way to learn about building corpora

Dictionary grew out of that

Under continuous development since

Progresses (or not) as my free time allows 🙄

THE RIGHT RHYMES

A Passion Project

THE RIGHT RHYMES

A Web Application
Speaking of … demo?
THE RIGHT RHYMES

What’s under the hood?
THE RIGHT RHYMES

An old-school web application

Written in the Django Web Framework

Persistence in PostgreSQL database

Deployed on a DigitalOcean droplet (1 vCPU, 2GB RAM)

THE RIGHT RHYMES

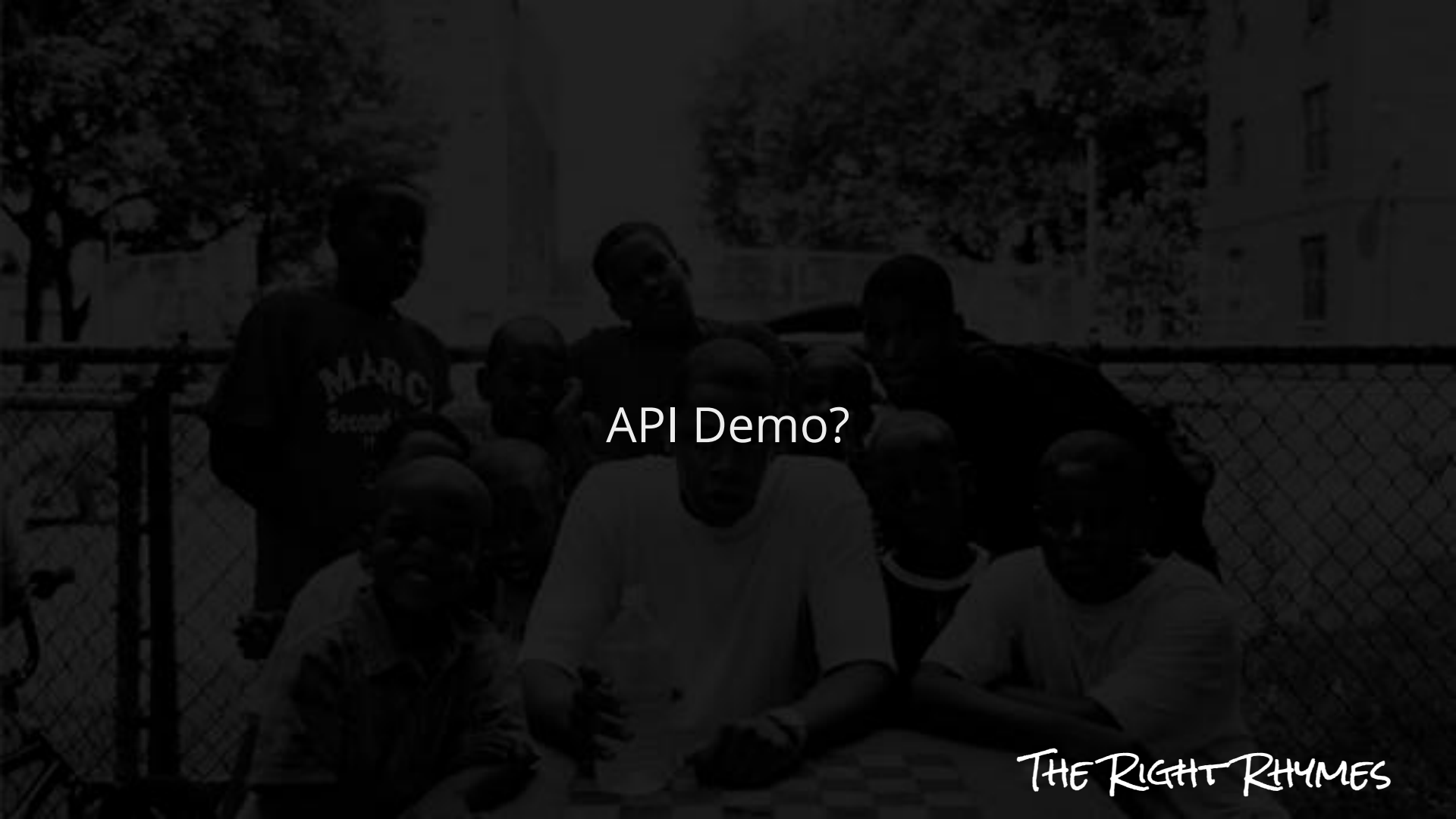
A RESTish API

Declared with Django Rest Framework

Endpoints service the front-end

Not really designed for public consumption...

THE RIGHT RHYMES


API Demo?
THE RIGHT RHYMES

Lessons learned

Django is great

JavaScript is powerful, but a whole different skill-set

Design is hard, interview your users

I do not scale 🙄

Let’s talk data
THE RIGHT RHYMES

Dictionary Data Overview

One big XML document 😳

Validated with a custom XML Schema

Versioned in Git

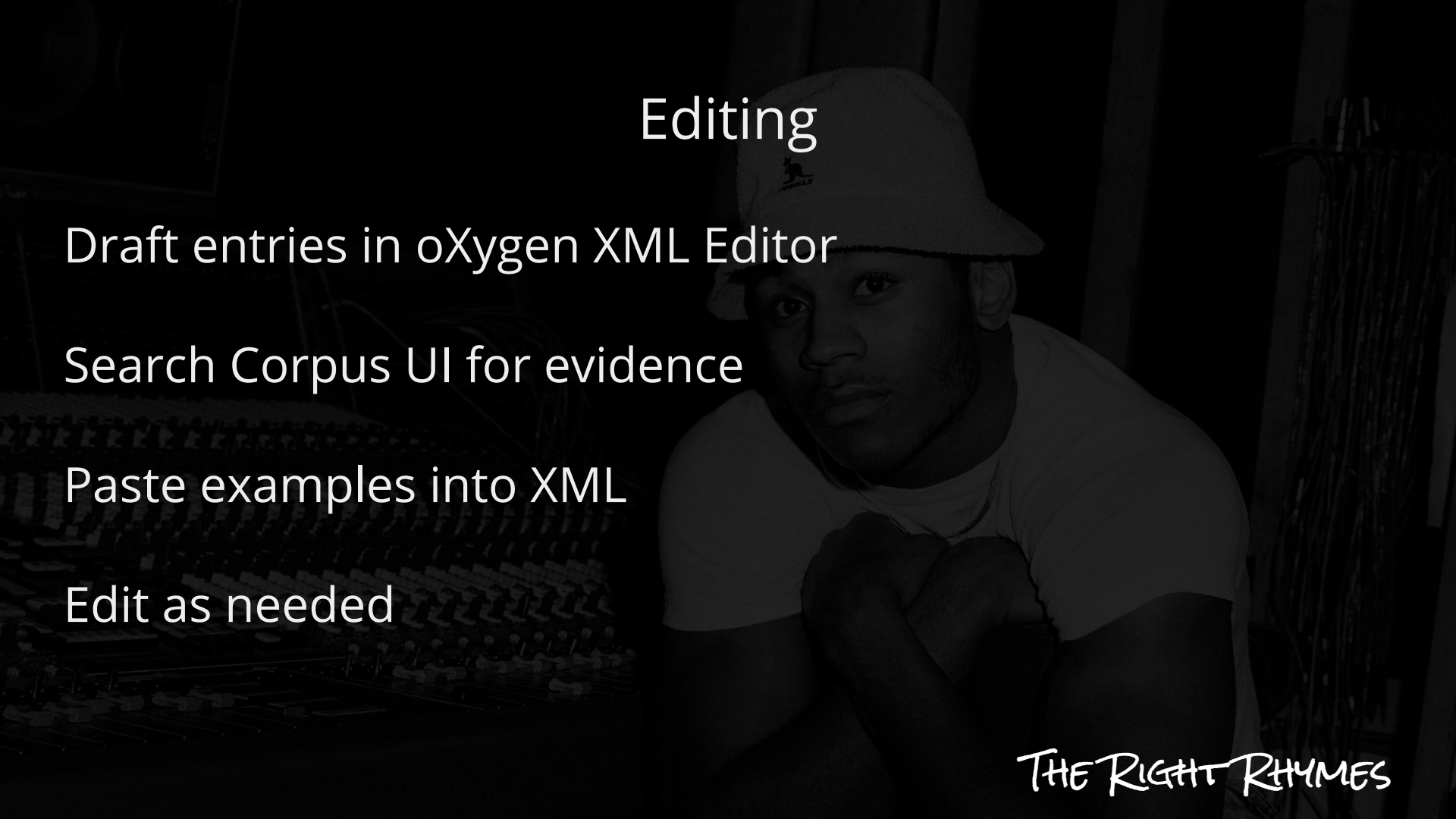
Editing

Draft entries in oXygen XML Editor

Search Corpus UI for evidence

Paste examples into XML

Edit as needed



Publishing

Denormalize XML with XSLT pipeline

Push to GitLab

Pull from GitLab & load into Django

Post-process for optimization

Shall we look at some XML?
THE RIGHT RHYMES

Benefits

Control over the data model

I am comfortable with XML

Low-cost tooling

THE RIGHT RHYMES

Pitfalls

Open-source software 💔 XML

Yet another technology stack

oXygen has no web interface

Zero integration with my other tools

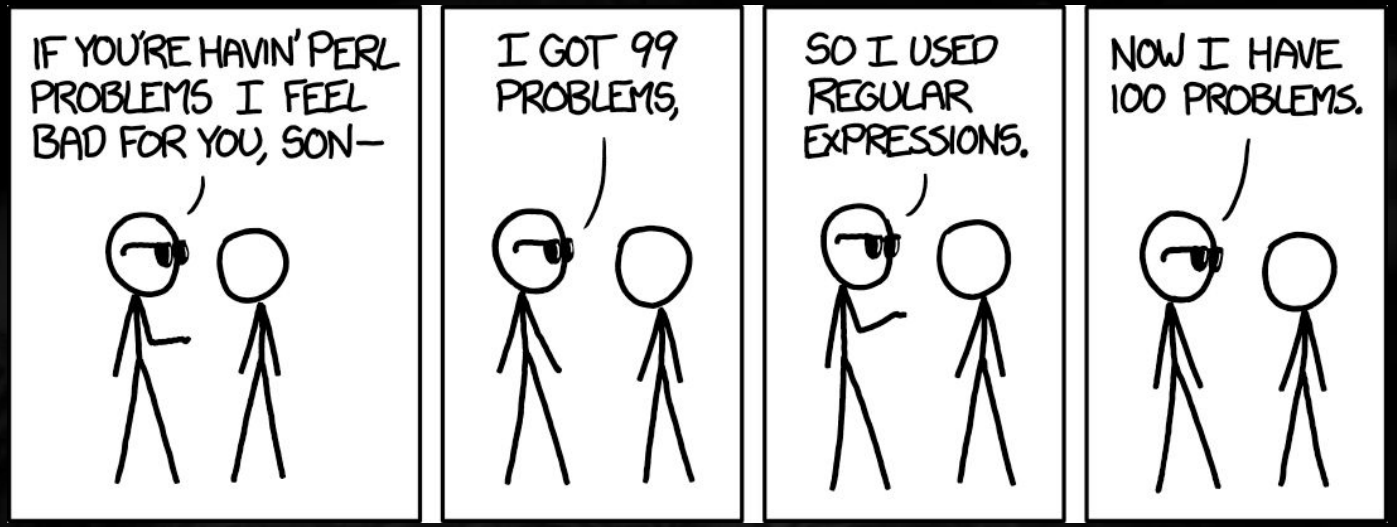
My latest stab at a lyrics corpus
THE RIGHT RHYMES

But before that, some history

THE RIGHT RHYMES

2012

My first attempt ...



THE RIGHT RHYMES

Lessons learned

Downloading whole sites doesn’t scale

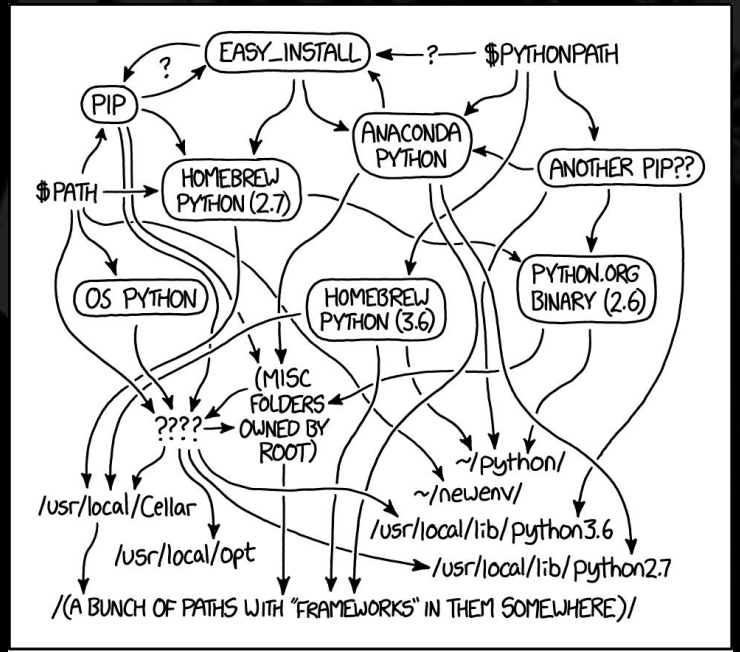
Adding structure is not trivial

Perl...

THE RIGHT RHYMES

2014

My second attempt...



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

THE RIGHT RHYMES

Lessons learned

I prefer Python to Perl; SQLite is great

Sourcing reliable metadata is hard

Data integration, also hard

Never throw anything away

2019

Query Spotify API for music metadata

Scrape Genius.com lyrics

Incorporate information from DBPedia, MusicBrainz, etc.

THE RIGHT RHYMES

Design

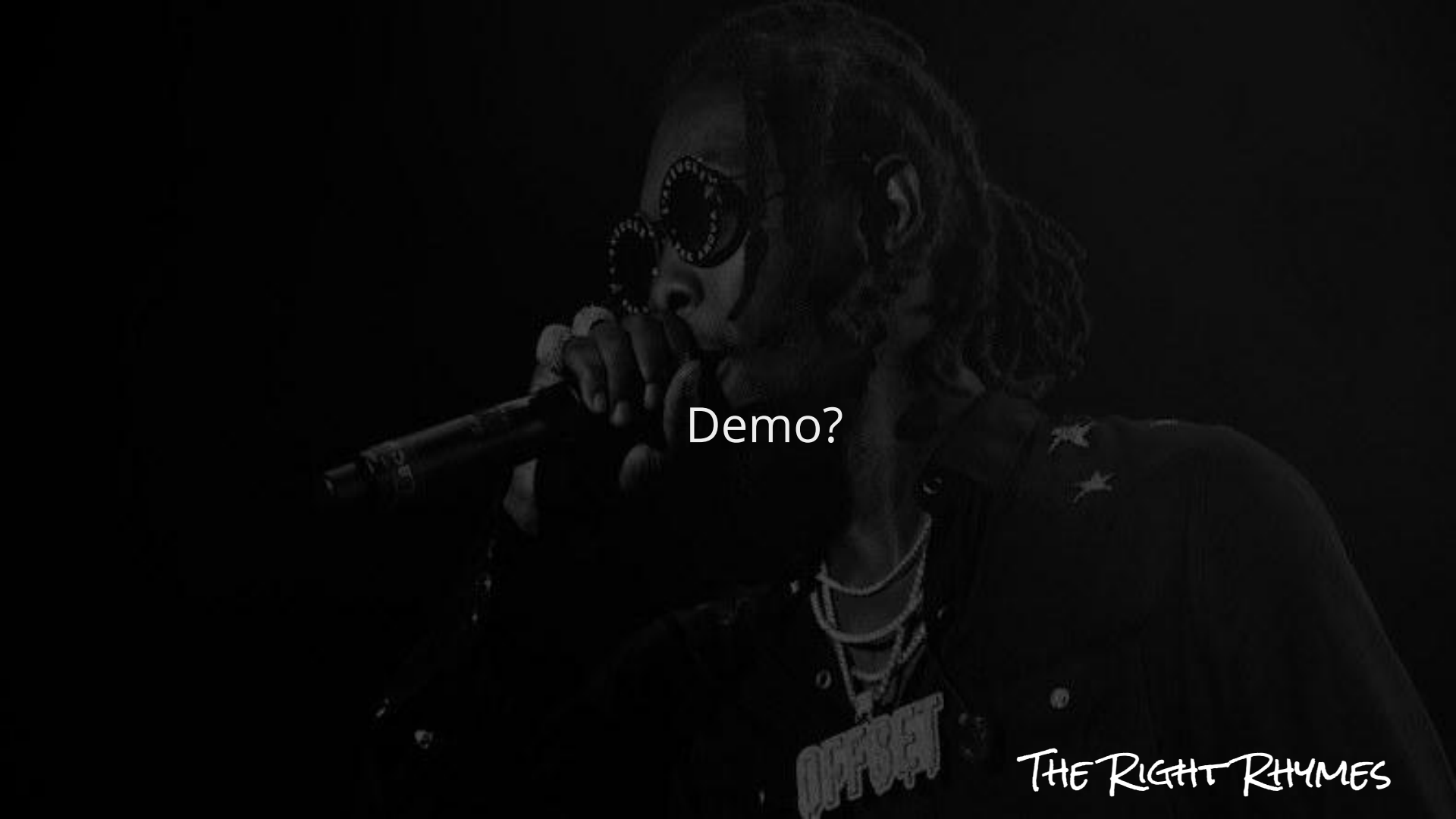
Flask application written in Python

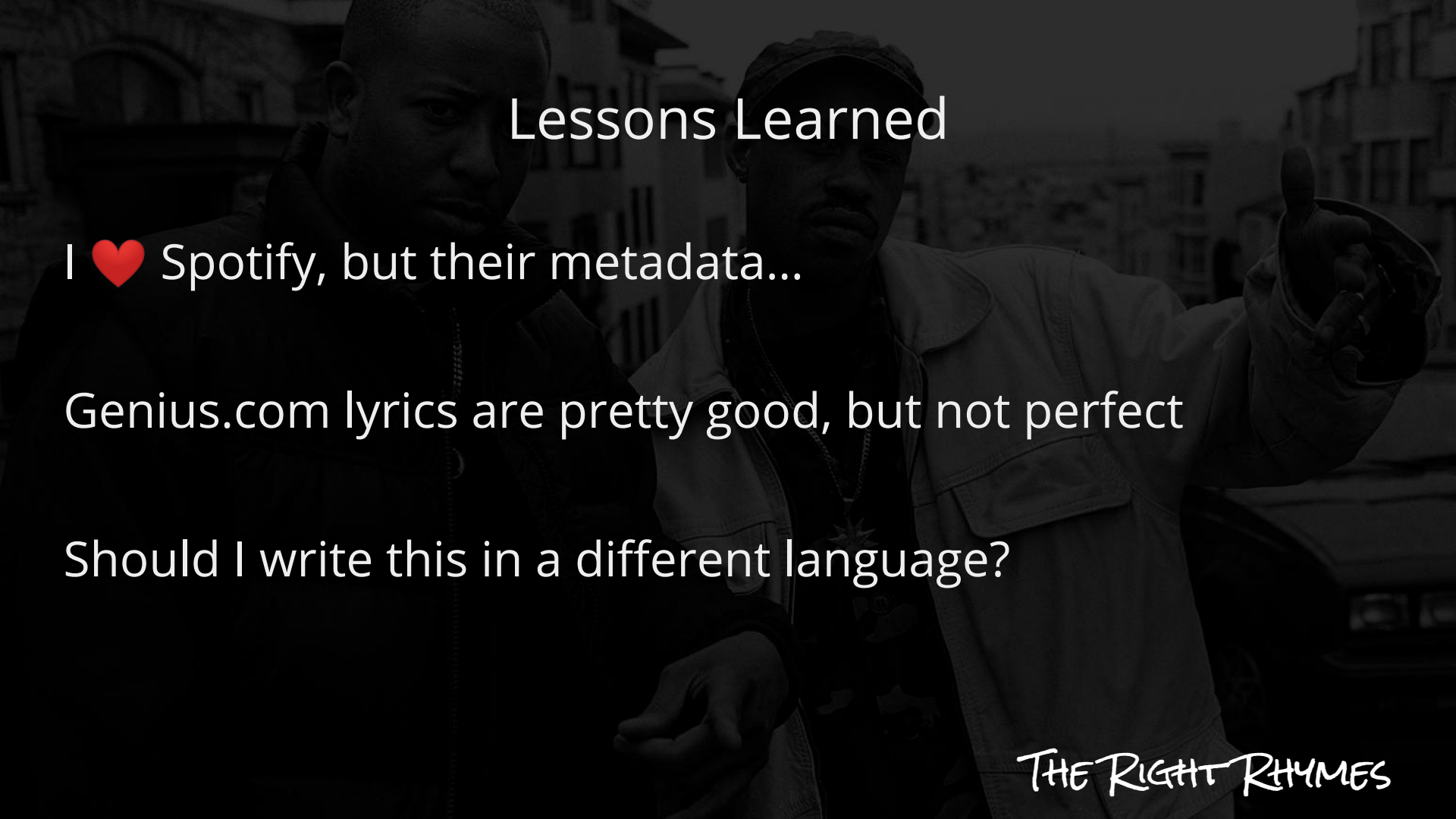
API endpoints to trigger data pipelines

Everything sinks into Postgres

Basic UI

THE RIGHT RHYMES


Demo?
OFFSET
THE RIGHT RHYMES



Lessons Learned

I ❤️ Spotify, but their metadata...

Genius.com lyrics are pretty good, but not perfect

Should I write this in a different language?

THE RIGHT RHYMES

Post-Elex Roadmap

Draft some entries 😜

Start integrating Seed 2.0 with The Right Rhymes

Feed corpus data into Sketch Engine?

Thank you so much for listening!
Any questions?
THE RIGHT RHYMES