Metadata problems in the music industry (and how to match music metadata)

Sertan Şentürk – Lead Data Scientist, R&D
Kobalt Music Group, London

Berlin MIR Meetup
Monday, September 30, 2019

OUTLINE

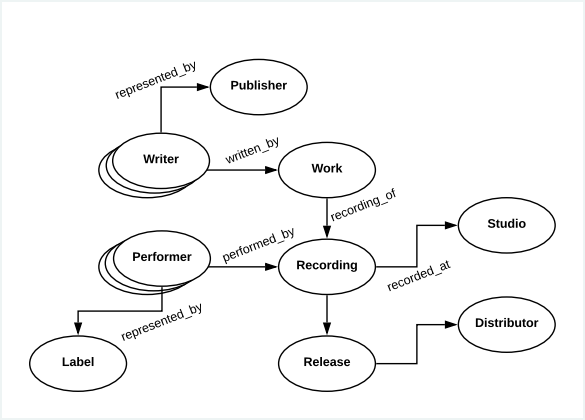
- Introduction
- Behind the Scenes
- Metadata Problems
- Metadata Matching
- Summary

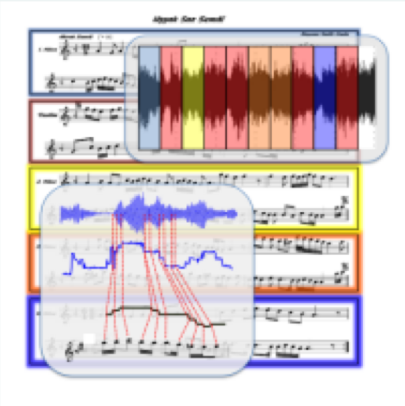
Disclaimers

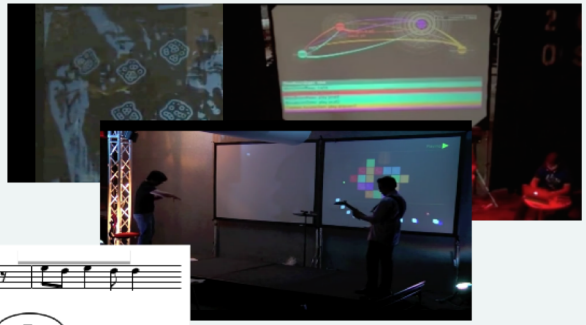
- I will try to cover a lot of bases
  - But we will only touch the surface
- I will not have much time to be specific for each case
  - Don’t hesitate at the moment to ask for more
- I wish to make more people aware
  - So I can get lazier ^_^

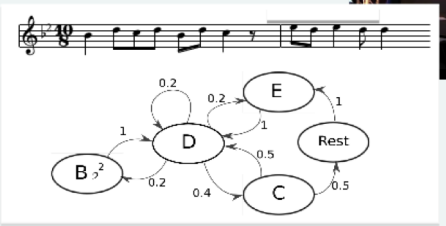
ABOUT ME

- Since 2018: Data Scientist @ Kobalt Music Group
- Previously (& briefly) @ SoundCloud
- 2017: PhD @ Music Technology Group, Universitat Pompeu Fabra
- 2011: MSc @ Georgia Tech Center for Music Technology



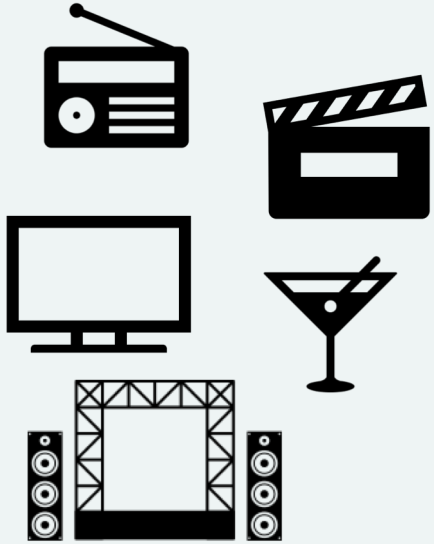


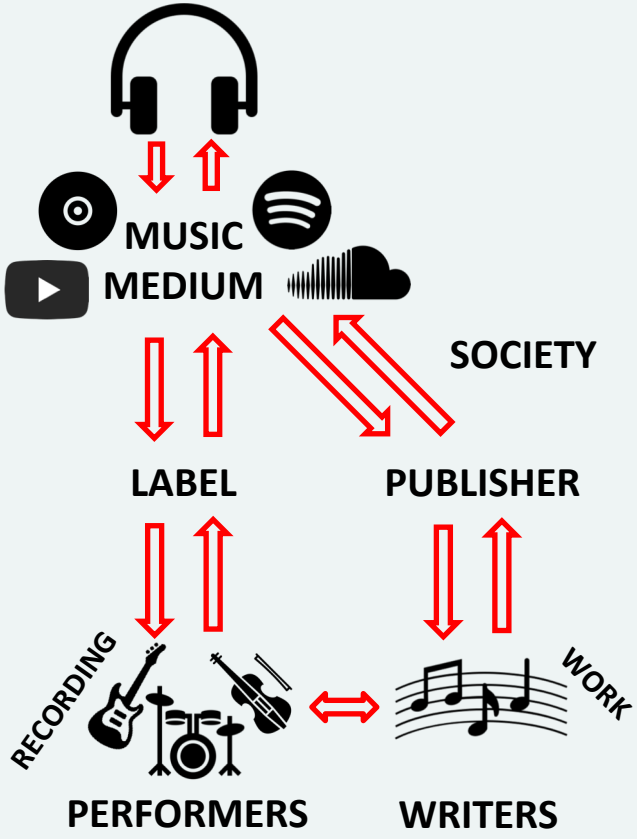




Behind the scenes of the music industry

MUSIC INDUSTRY IN BRIEF (!)

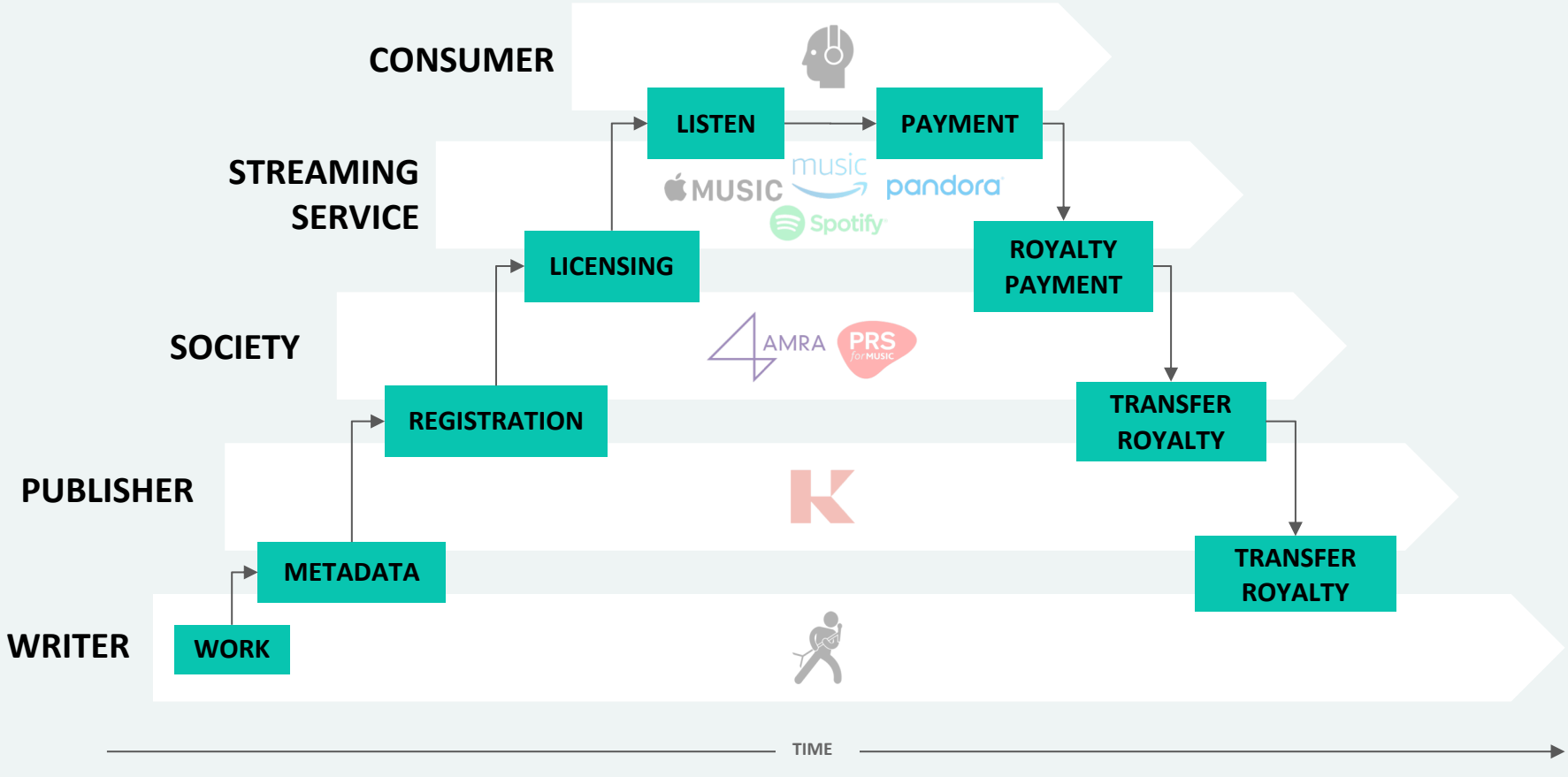




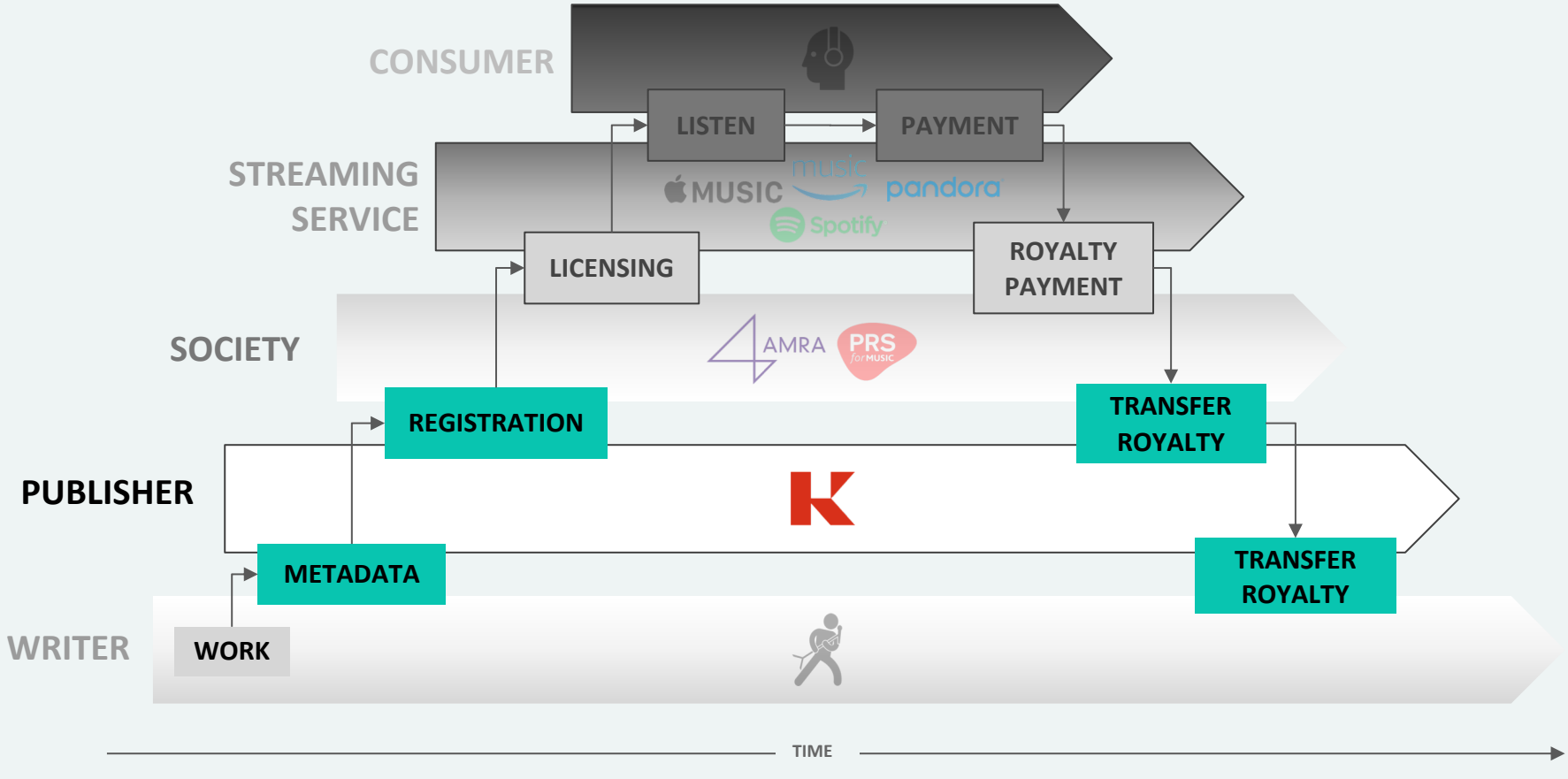
COMPLEX ECOSYSTEM

CONSUMER		100 000 000's...
RETAILER		1000's...
SOCIETY	AMRA	100's...
PUBLISHER / RECORD LABEL	Kobalt AWAL A KOBALT COMPANY	1000's...
CREATOR		1 000 000's...

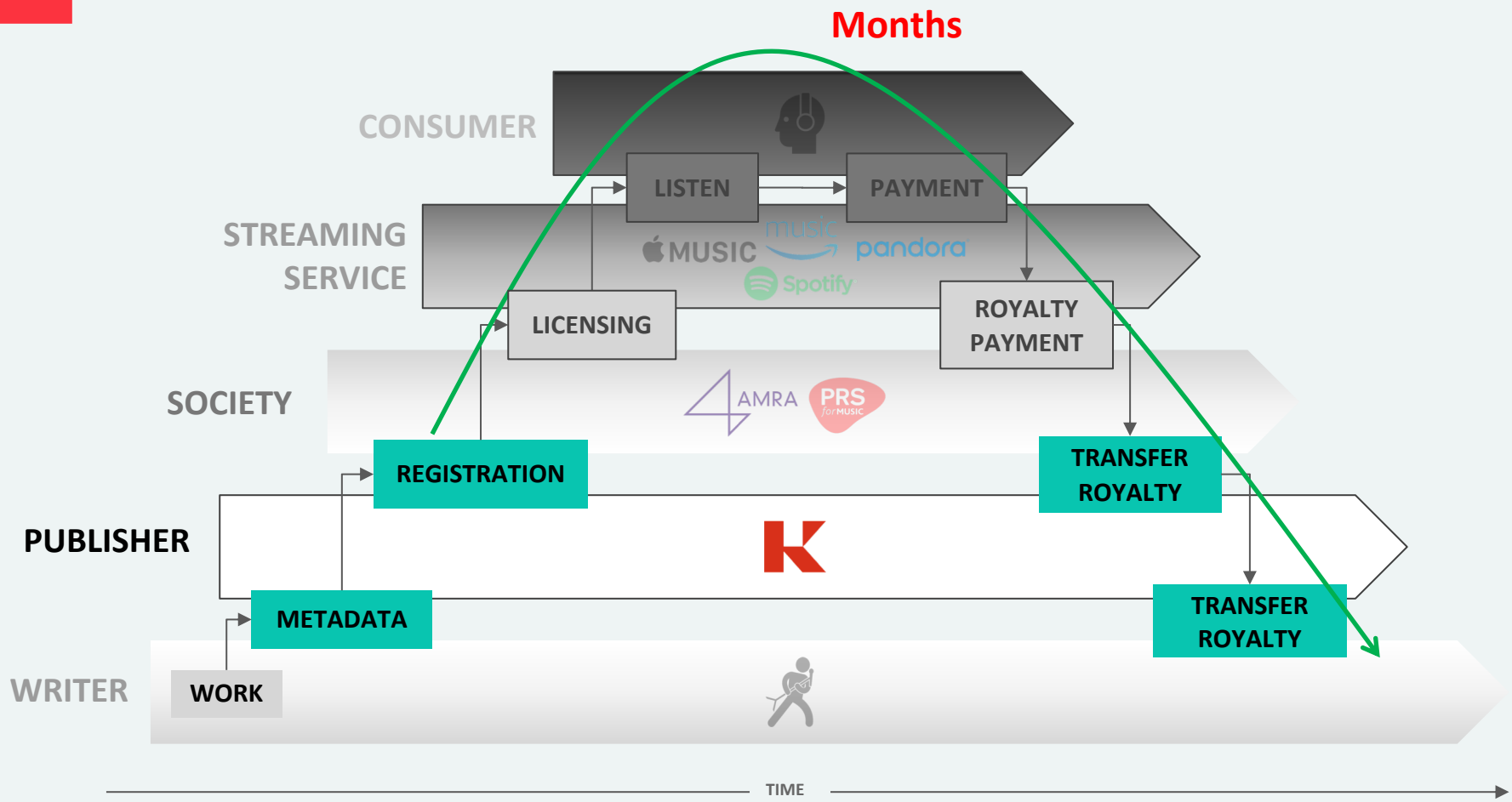
MULTI-LAYERED OPS



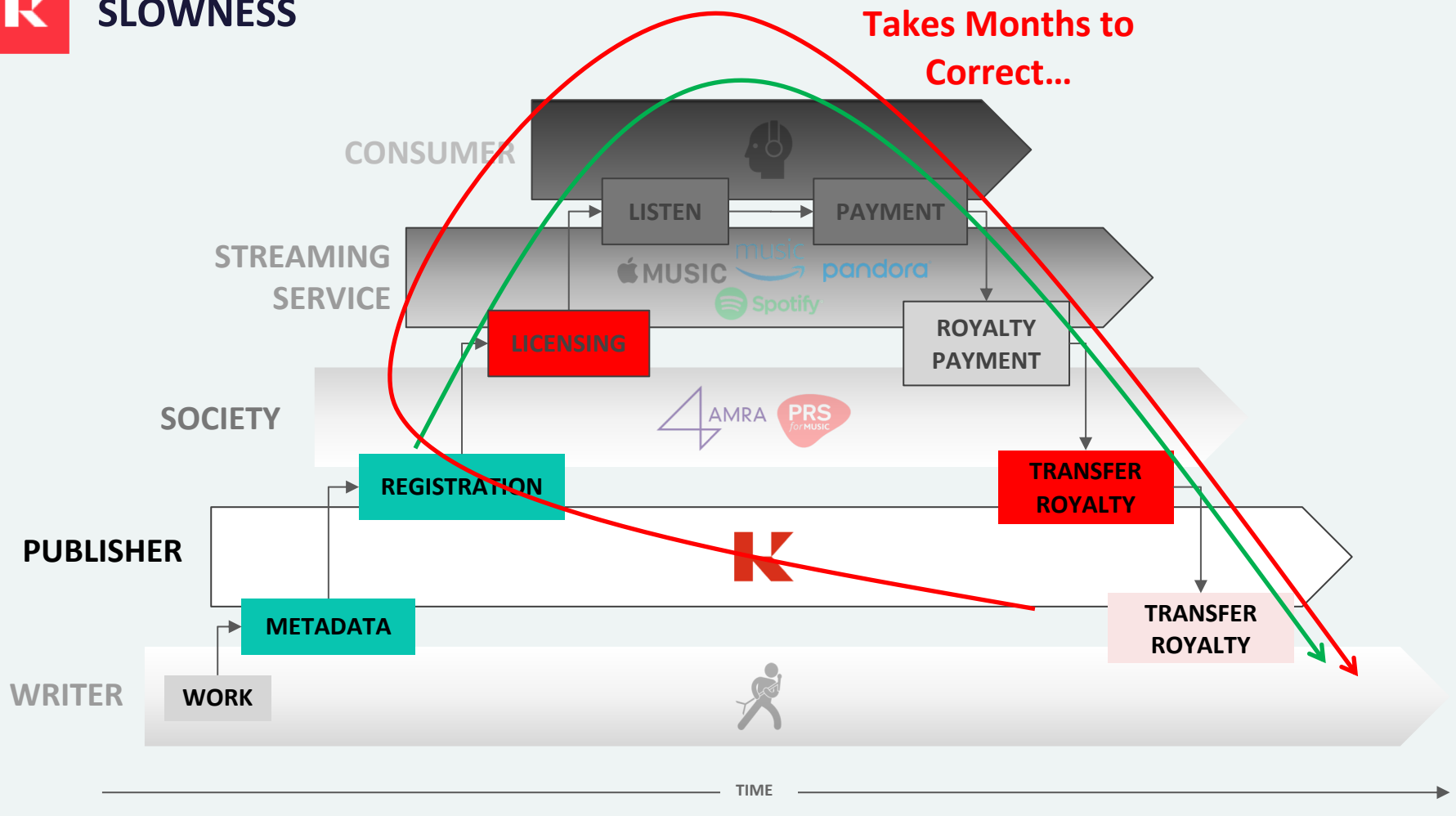
OPAQUENESS


CONSUMER
LISTEN
PAYMENT
STREAMING SERVICE
MUSIC
music
pandora
Spotify
LICENSING
ROYALTY PAYMENT
SOCIETY
AMRA
PRS for MUSIC
REGISTRATION
TRANSFER ROYALTY
PUBLISHER
METADATA
TRANSFER ROYALTY
WRITER
WORK
TIME

SLOWNESS


Months
CONSUMER
LISTEN
PAYMENT
STREAMING SERVICE
MUSIC
music
pandora
Spotify
LICENSING
ROYALTY PAYMENT
SOCIETY
AMRA
PRS for MUSIC
REGISTRATION
TRANSFER ROYALTY
PUBLISHER
METADATA
TRANSFER ROYALTY
WRITER
WORK
TIME

SLOWNESS


Takes Months to Correct…
CONSUMER
LISTEN
PAYMENT
STREAMING SERVICE
MUSIC
music
pandora
Spotify
LICENSING
ROYALTY PAYMENT
SOCIETY
AMRA
PRS for MUSIC
REGISTRATION
TRANSFER ROYALTY
PUBLISHER
METADATA
TRANSFER ROYALTY
WRITER
WORK
TIME

Metadata Problems

AMBIGUITIES

Name	Surname	Country	Genre	Role
Roger	Taylor	UK	Rock	Drums

Roger (Meddows) Taylor
Queen

Roger (Andrew) Taylor
Duran Duran

CHALLENGES

Name	Surname	Country	Genre	Role	Act
Roger Meddows	Taylor	UK	Rock	Drums	Queen
R.	Taylor	UK	Rock	Drums	Live Aid ‘85
Roger	Taylor	England	Rock‘n Roll	Percussion	Duran Duran
Roger Meddows	Taylor	UK	Rock	Guitar	The Cross
Roger	Taylor	?	None	-999	
Roger Meadows	Tylor	UK	Rock	Drums	Queen
Roger Andrew	Taylor	UK	Rock	Drums	Queen

INSUFFICIENT

INCONSISTENT

MULTI-FACETED

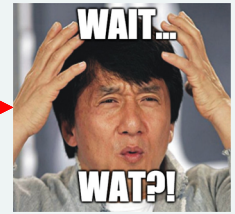
MISSING

ERRORS

CONFLICTS

ENTITY RECOGNITION

Name	Surname	Country	Genre	Role	Act
Roger Meddows	Taylor	UK	Rock	Drums	Queen
R.	Taylor	UK	Rock	Drums	Live Aid ‘85
Roger	Taylor	England	Rock‘n Roll	Percussion	Duran Duran
Roger Meddows	Taylor	UK	Rock	Guitar	The Cross
Roger	Taylor	?	None	-999	
Roger Meadows	Tylor	UK	Rock	Drums	Queen
Roger Andrew	Taylor	UK	Rock	Drums	Queen



HOW CAN WE FIX THIS MESS?!?

“IDEAL” SOLUTION

USE UNIVERSALLY UNIQUE IDENTIFIERS

MusicBrainz ID	Name	Surname	Country	Genre	Role	Act
558302b1-07ae-4be1-9edb-2a2b1f036313	Roger Meddows	Taylor	UK	Rock	Drums	Queen
f2b7fc94-1e8d-4f4d-9a3f-620c1a5ac564	R.	Taylor	UK	Rock	Drums	Live Aid ‘85
f2b7fc94-1e8d-4f4d-9a3f-620c1a5ac564	Roger	Taylor	England	Rock‘n Roll	Percussion	Duran Duran
558302b1-07ae-4be1-9edb-2a2b1f036313	Roger Meddows	Taylor	UK	Rock	Guitar	The Cross
f2b7fc94-1e8d-4f4d-9a3f-620c1a5ac564	Roger	Taylor	?	None	-999	
558302b1-07ae-4be1-9edb-2a2b1f036313	Roger Meadows	Tylor	UK	Rock	Drums	Queen
	Roger Andrew	Taylor	UK	Rock	Drums	Queen

THEN, WHAT IS THE PROBLEM?!?!

THERE IS
NO AGREED, UNIVERSAL
DATA SOURCE
FOR MUSIC…

BUT… CAN'T WE CREATE ONE!?

MANY TRIED...

SO, HOW SHOULD WE PROCEED?!?

START WITH YOUR OWN TURF

DATA OPS

- **Data modelling:** Use the right model for right type of data
  - Tabular information (e.g. client records)?: **Relational Databases**
  - Unstructured/semi-structured data (e.g. tags)?: **NoSQL Document Stores**
  - Relations, communities (e.g. chain of title)?: **Graph Databases**
  - **NOTE:** There are many more things to consider! Just ask any information management person!
- Quality
- Matching
- Enrichment
- Reconciliation
- Master Data Management

DATA OPS

- Data modelling
- **Quality:** Monitor the state of your metadata
  - **Completeness:** How much information is available for a document, field or dataset?
  - **Coverage:** How well does a data source represent others?
  - Does the data fulfil the expected types/constraints?
  - …
  - A nice tool: Amazon Deequ
- Matching
- Enrichment
- Reconciliation
- Master Data Management

DATA OPS

- Data modelling
- Quality
- **Match** and connect metadata as much as possible
  - Deduplicate documents (Sertan Şentürk vs. Sertan Sentuerk)
  - Link relevant documents (recordings → works)
  - Embrace linked data!
- Enrichment
- Reconciliation
- Master Data Management

DATA OPS

- Data modelling
- Quality
- Matching
- **Enrich** your metadata whenever you have the chance
  - Fill missing values
  - Correct/standardize values
  - Fetch information from other sources
  - …
  - A nice tool: http://holoclean.io/
- Reconciliation
- Master Data Management

DATA OPS

- Data modelling
- Quality
- Matching
- Enrichment
- **Reconciliation**
  - Merge information during deduplication
  - Detect and resolve conflicting information**: ontologies are a great help!**
  - Ensure the processes are not creating mess
- Master Data Management

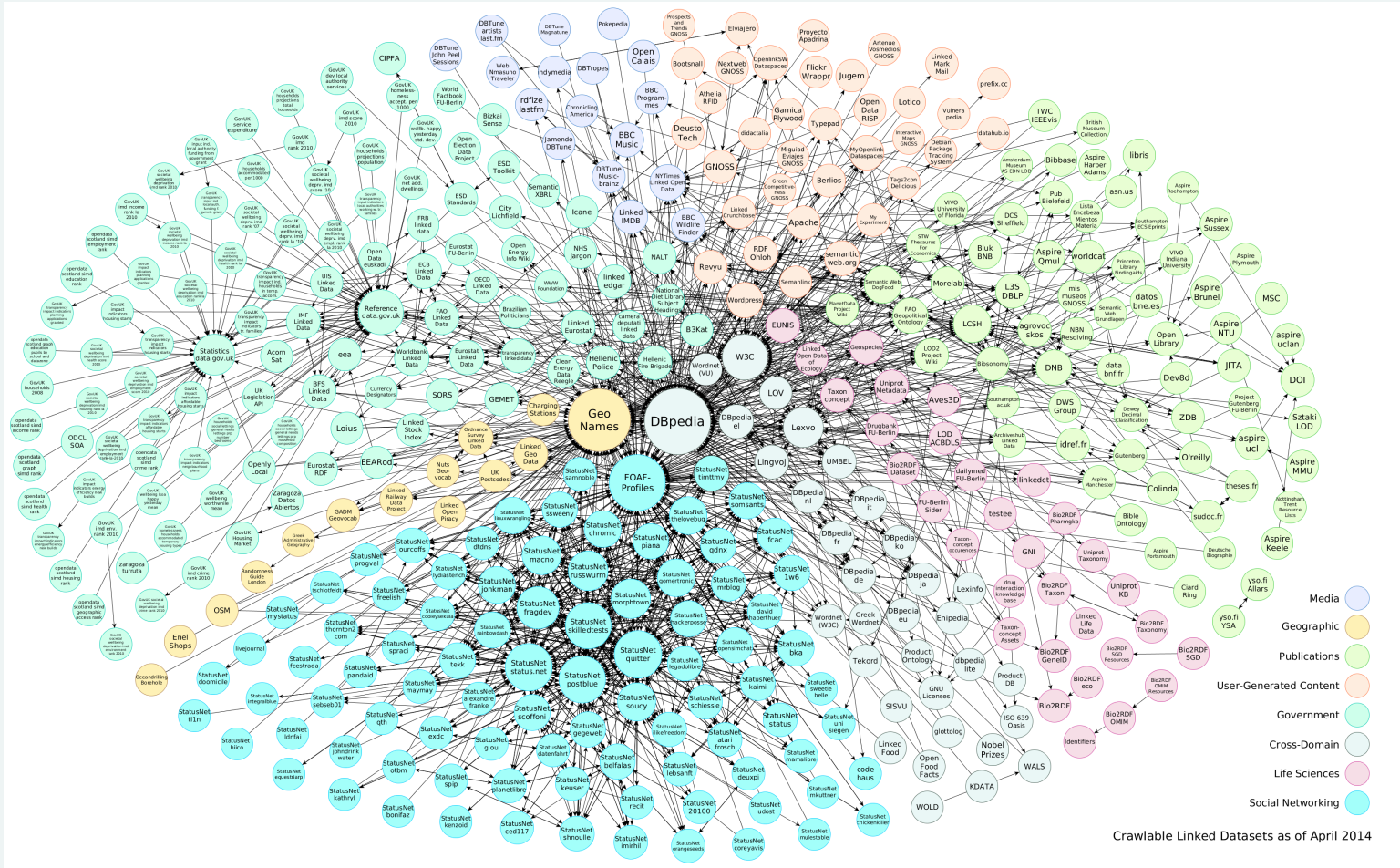
DATA OPS

- Data modelling
- Quality
- Matching
- Enrichment
- Reconciliation
- **Master Data Management**
  - Keep a track of all changes, sources etc.
  - Maintain “golden” records
  - Provide a single point of truth (as much as possible)

SUGGESTIONS

- Cleaning everything by hand is impossible/impractical
  - Heavy monitoring and automation
- Yet keep the humans (experts with domain knowledge) in the loop:
  - to deal with difficult/edge cases
  - to evaluate your automations
- Problems will always linger
  - We have to minimize its impact
- You cannot make your data clean and shiny in a single go
  - Fix iteratively
  - Each improvement will facilitate others
- Play the long game
  - Turn it into a habit/culture of the organization

ULTIMATE GOAL: OPEN & LINKED DATA



METADATA MATCHING

METADATA MATCHING

- Has many names in the literature
  - record linkage, data integration, entity recognition, identity resolution, …
- People mean different things
  - e.g. joining tables, enrichment, conflict resolution, …

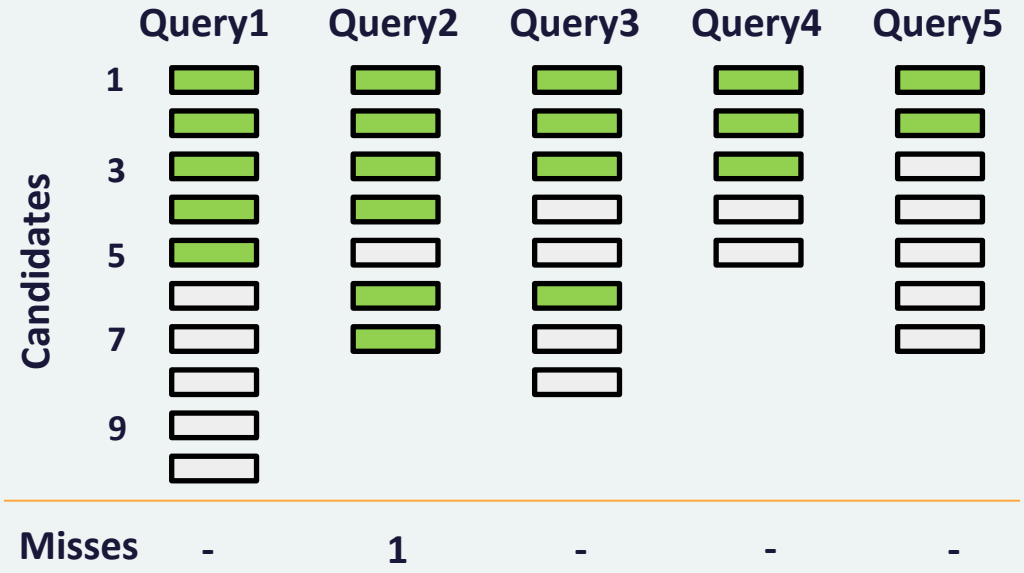
- **Metadata matching:** Creating a (named) link between documents which are represent the same entity (same person in two different datasets) or have a relationship (recording to its work)

METADATA MATCHING (2)

- Retrieval problem
  - Sometimes ranked
- **Targets** vs **Queries**
- (Typically) binary relevance
- Cardinalities vary depending on the task
  - Recording to work: low
  - Work to recording: higher

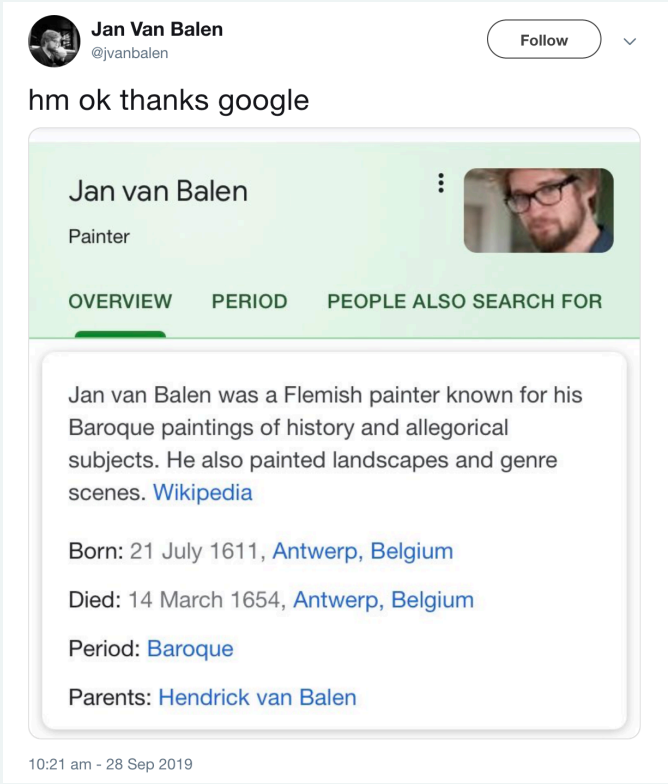
EVALUATION

- Depending on the cardinality and use case, e.g.
  - **Human** (MAP, MRR) vs. **Fully automatic** (F_1); **needle in haystack** (Prefer PR over ROC-AUC); **Low tolerance to false positives** (Weight precision more than recall)



EVERYONE HAS SIMILAR PROBLEMS

- Insurance
- Finance
- Manufacturing
- Medicine
- Online stores
- …



WHERE DO WE NEED MATCHING?

Usage processing: Match consumption reported by streaming services to works

Royalty processing: Match royalty statements to works

Onboarding: Merge entities (creators, works, recordings) in an external catalogue with the internal catalogue

Deduplication: Find duplicate entities inside a catalogue

Enrichment: Match entities in an external catalogue with the internal catalogue to complete missing information

Relation discovery: Identify missing links between entities (e.g. collaborators in an album)

Error resolution: …

SOLUTIONS

RULE-BASED SOLUTIONS

- First choice to go
  - Before big data era (till late 2000s)
  - (many music businesses are still here)
- Based on simple, deterministic decisions
  - e.g. full/sub string match between separate fields and converting the results to a “match percentage” according to some business logic
- Seems interpretable at first, but rules/logic explode as time passes
  - Spurious and/or conflicting rules
- Does not scale beyond small data (>10k documents)
- Does not generalize to different tasks
- Don’t worry if you are at this stage
  - You can use the system to collect initial data to **evaluate** future methods

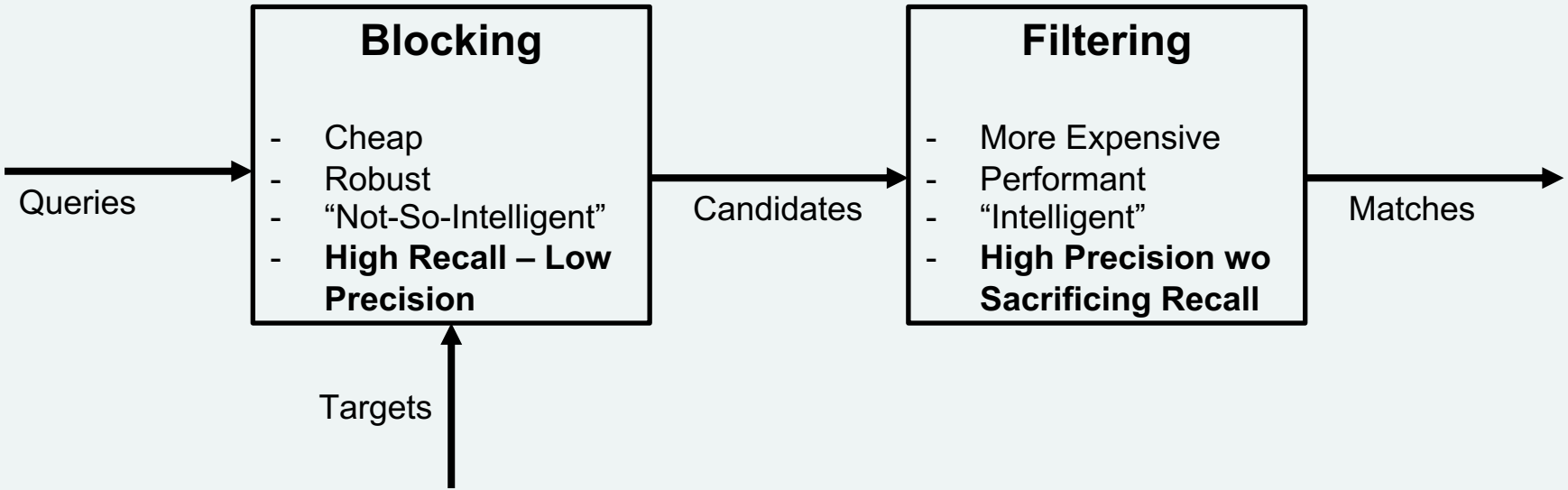
OUT OF BOX SOLUTIONS

- If you don’t have engineering resources or time
- No need to build or maintain much infrastructure

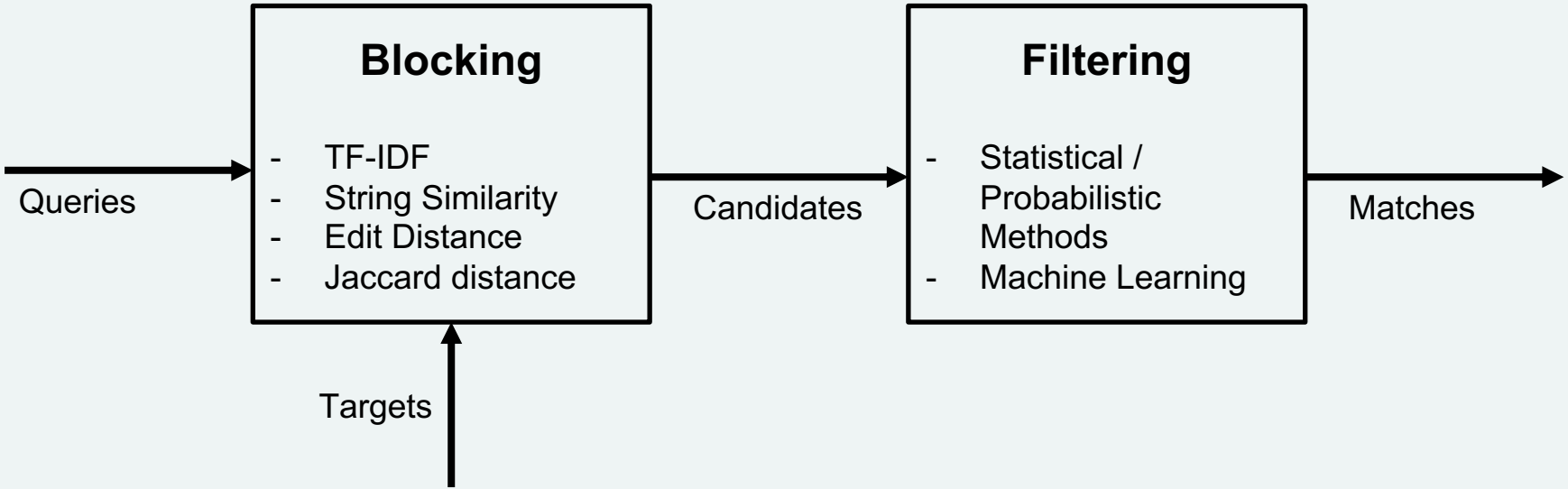
- “FindMatches” ML transform
- Works hand-to-hand with Lake Formation

- Pioneer DataOPS start-up
- Does more than matching

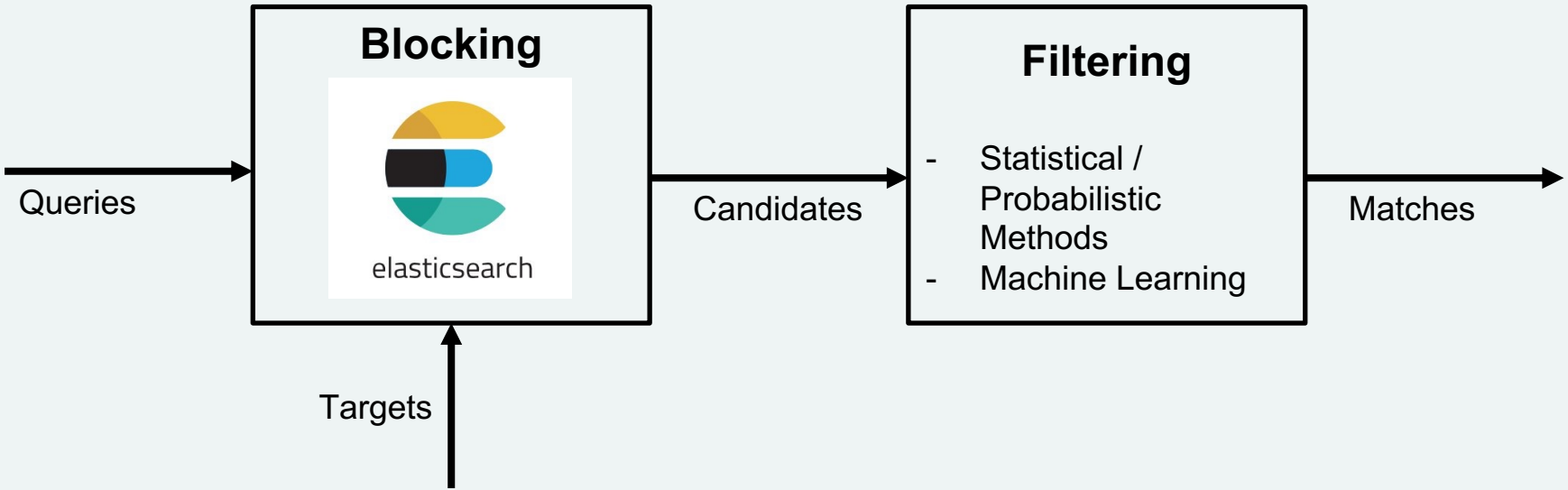
COMMON APPROACHES



APPROACHES

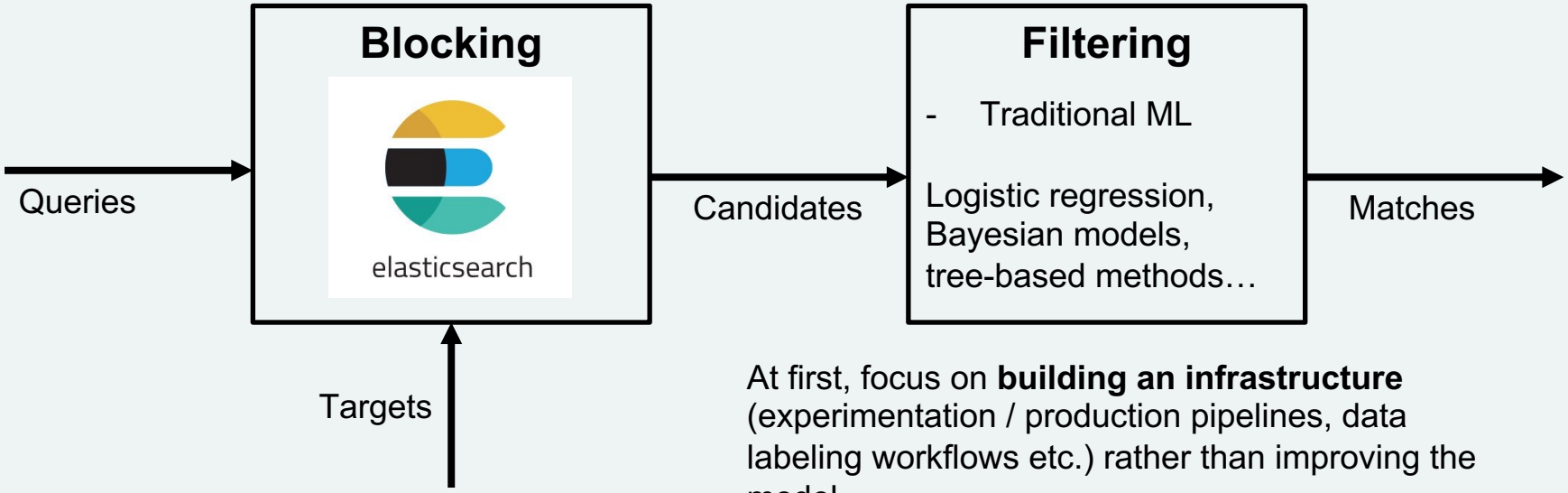

Blocking
- TF-IDF
- String Similarity
- Edit Distance
- Jaccard distance
Filtering
- Statistical / Probabilistic Methods
- Machine Learning
Queries
Candidates
Matches
Targets

ELASTICSEARCH



Şentürk, S. (2019). Music metadata matching using ElasticSearch. London ElasticSearch Meetup

FILTERING USING ML

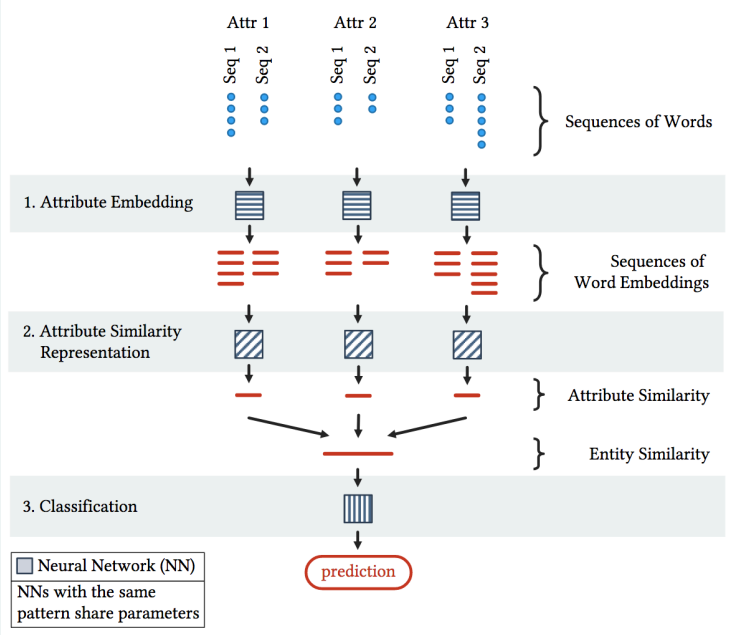


At first, focus on **building an infrastructure** (experimentation / production pipelines, data labeling workflows etc.) rather than improving the model

Nice tool: Snorkel for weakly labelling data

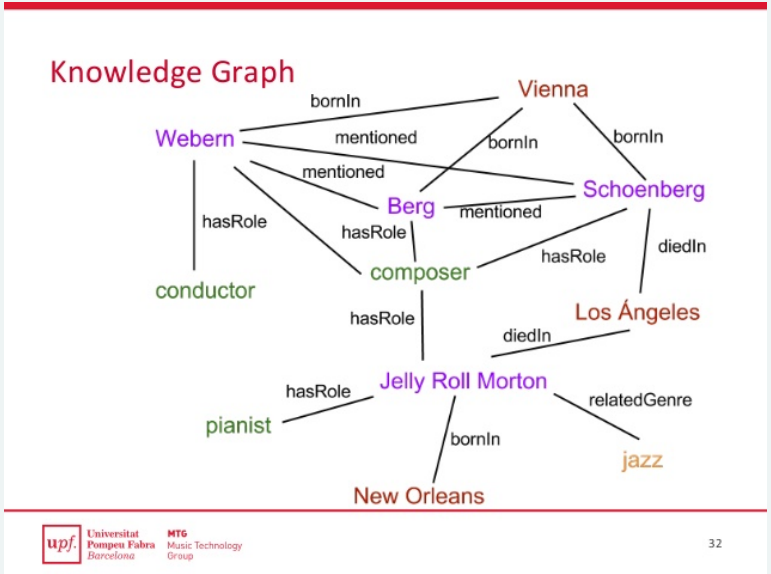
WHERE IS DEEP LEARNING!?

- Late adoption in the academia & industry
- “deepmatcher” (Sigmod 2018)
  - How does DL architectures fare against SOTA? (Does not outperform for structured data yet)
  - SIF, RNN, attention-based & hybrid models using word embeddings
  - Code (Github)
  - Also evaluated on (rather small-sized) music metadata (Demo)



Mudgal, Sidharth, et al. "Deep learning for entity matching: A design space exploration." *Proceedings of the 2018 International Conference on Management of Data*. ACM, 2018.

GRAPH-BASED APPROACHES



- node2vec: http://snap.stanford.edu/node2vec/
- Graph Convolution Neural Networks

Oramas, S., & Sordo, M. (2015) Knowledge acquisition from music digital libraries. IAML/IMS Congress

SUMMARY

- Music metadata is (to put it kindly) bad…
- Why?!?
  - uncoordinated supply chains, data quality issues, absence of trust, lack of data literacy, human errors, and national politics…
- What should we do?
  - First, clean your own turf
  - Follow **information management best practices** and embrace **linked data principles**
  - Connect your metadata with others

SUMMARY (2)

- Metadata matching is inevitable in the current ecosystem (and till eternity)!
  - i.e. symptom of lack of interconnectivity between metadata resources
  - If you sweep your data problems under the carpet, they will hunt you later in the form of matching…
  - Remedy your problems before they become a sickness requiring a surgery…
- Where should I start?
  - Focus on a single use case; generalize later
  - Spend time on understanding the use case, pick a relevant eval measure, collect data
  - Blocking using ElasticSearch; use simple ML models (e.g. tree-based) based on TF-IDF to filter candidates

RESOURCES

- Conferences: Sigmod, VLDB, SigKDD
- Academics: Michael Stonebraker (MIT CSail), Ihab Francis Ilyas (Uni Waterloo), AnHai Doan & Theodoros Rekatsinas (UW Madison), Xu Chu (Georgia Tech), Christopher Ré (Stanford)
- Course Material: Duke Uni CompSci590.01, UW Madison CS 838
- Companies: Tamr, Trifacta
- Tools: AWS Lake Formation, ElasticSearch, Apache TinkerPop

  Magellan, deepmatcher, py_stringmatching, py_stringsimjoin

  Snorkel, HoloClean

  Deequ

FURTHER READING

- Stonebraker, M., & Ilyas, I. F. (2018). Data integration: The current status and the way forward. IEEE Data Eng. Bull., 41(2), 3-9.
- Hellerstein, J. M. (2008). Quantitative data cleaning for large databases. United Nations Economic Commission for Europe (UNECE).
- Oramas, S., & Sordo, M. (2015) Knowledge acquisition from music digital libraries. IAML/IMS Congress
- Allemang, D., & Hendler, J. (2011). Semantic web for the working ontologist: effective modeling in RDFS and OWL. Elsevier.
- Kipf, T. (2016). Graph convolutional networks. Self-Published
- Şentürk, S. (2019). Music metadata matching using ElasticSearch. London ElasticSearch Meetup

WE ARE GROWING

We are hiring!!!

Lots of positions and roles, incl. data engineers at all levels

Speak with me

https://www.kobaltmusic.com/company/careers


QUESTIONS?