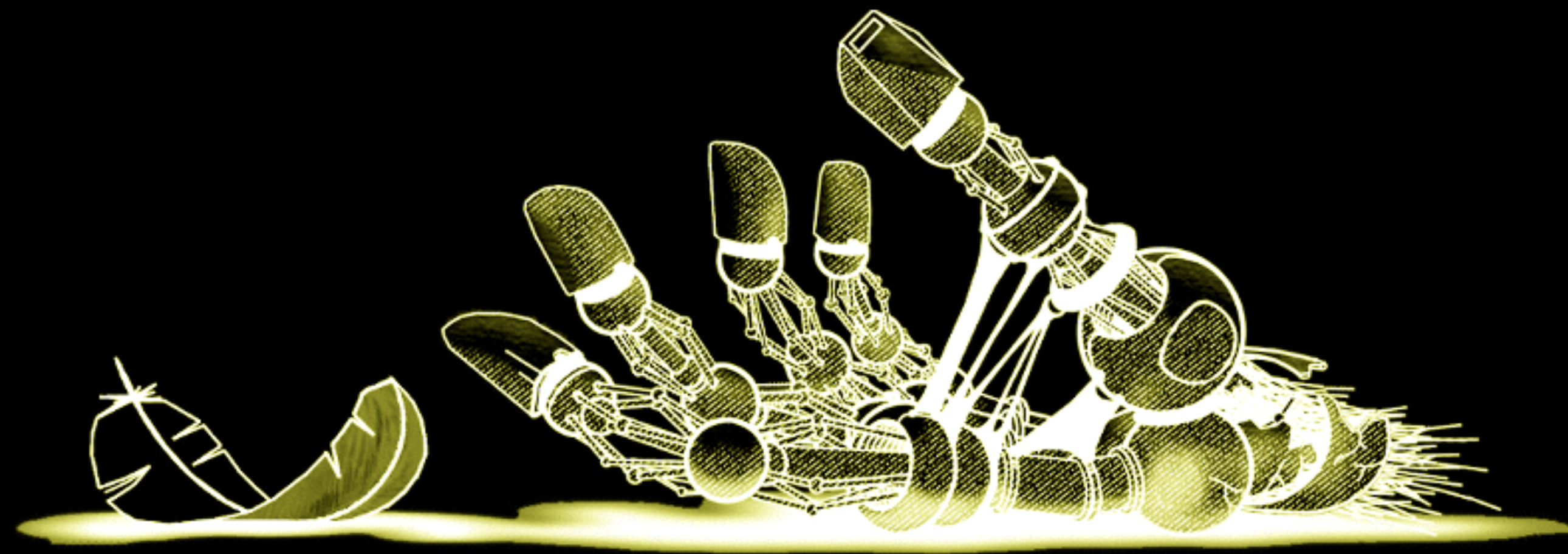
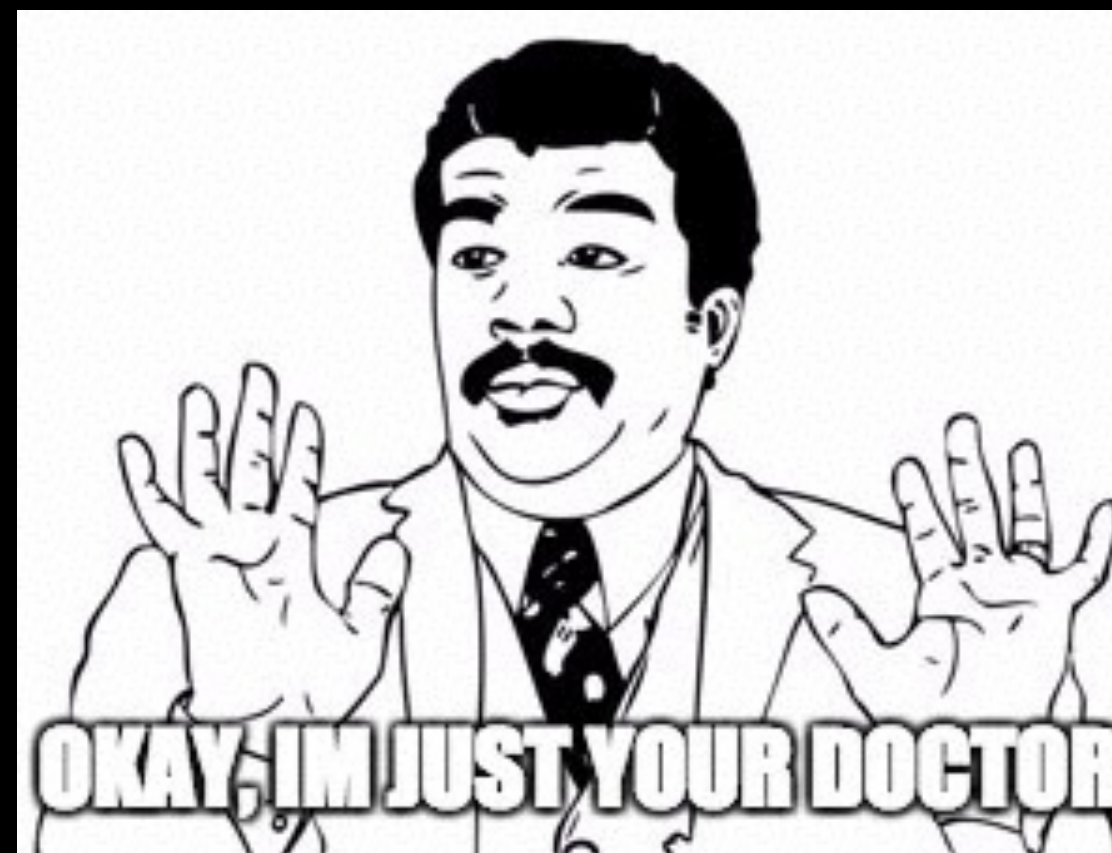


Building a high scale machine learning pipeline with Apache Spark and Kafka



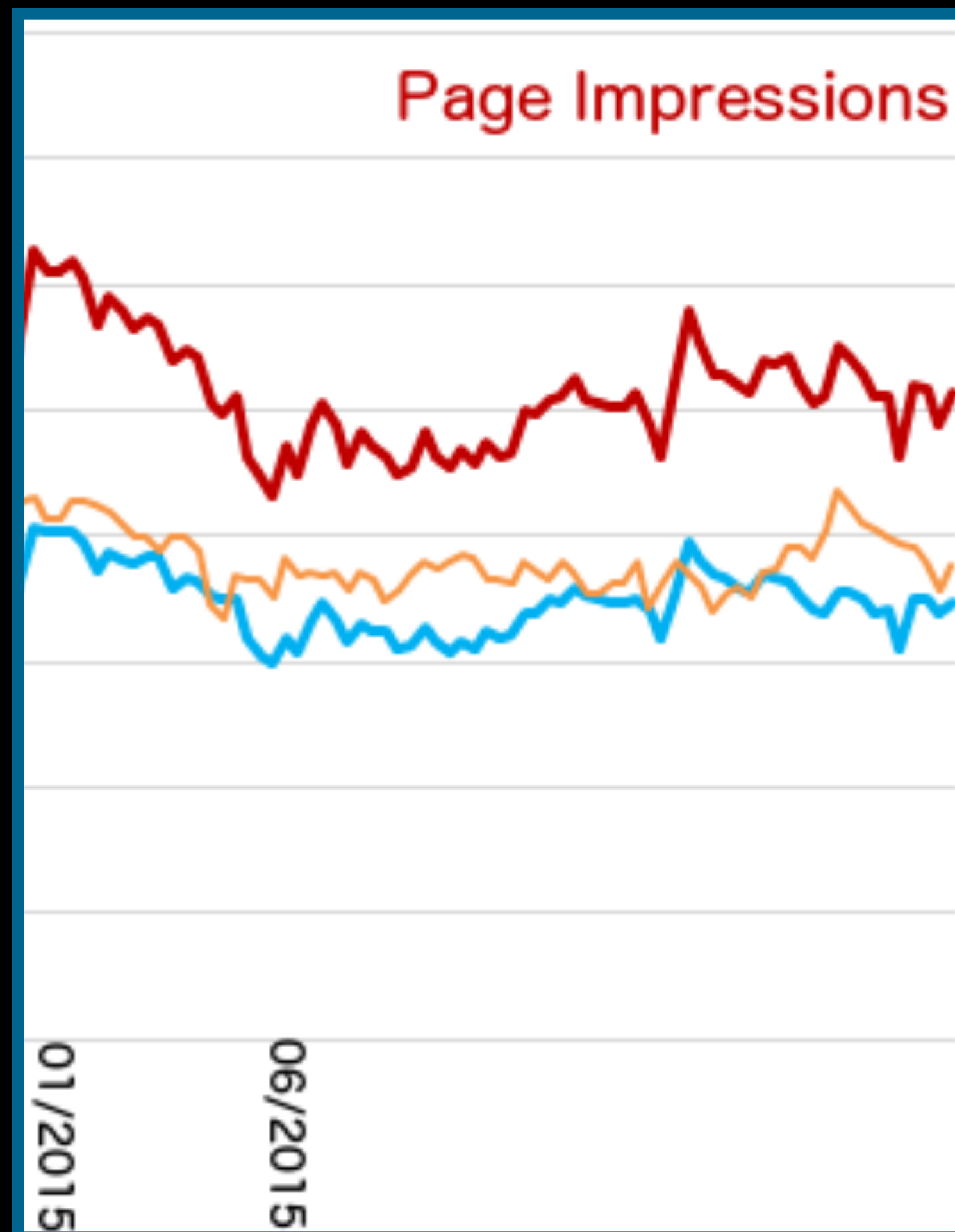
Bedő Dániel

gutefrage.net



- biggest community-driven question & answer website in Germany
- 20 million questions, 70 million answers
- similar to Quora, Yahoo Answers

Google Update Impact



Ordering of answers



Frage von  16.05.2012

Warum lernt man in der Schule so viel Sinnloses?

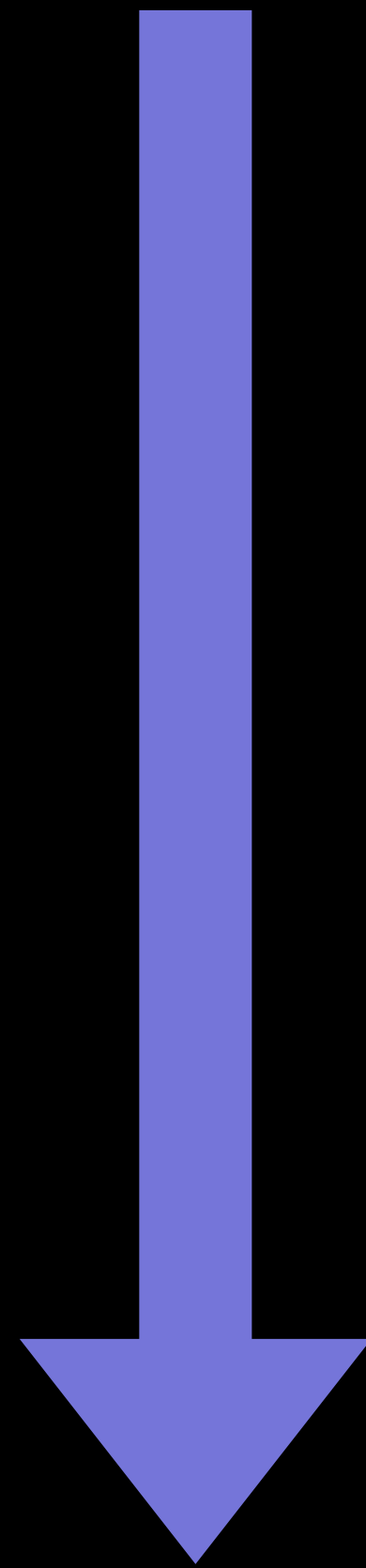
Ich bin der Meinung, das unser Schulsystem weitgehend sinnlos ist. Man lernt zu viele Sachen die man nie mehr brauchen wird. Ein Beispiel sind Parabeln: Wieso muss ich wissen, wie man Parabeln berechnet, wie sie aussehen und so ... [komplette Frage anzeigen](#)

↓ 259 Antworten

 Benachrichtigung erhalten

▸ [Rechtschreibfehler gefunden](#)

supervised machine learning



determine the type of the training data

gather a training set

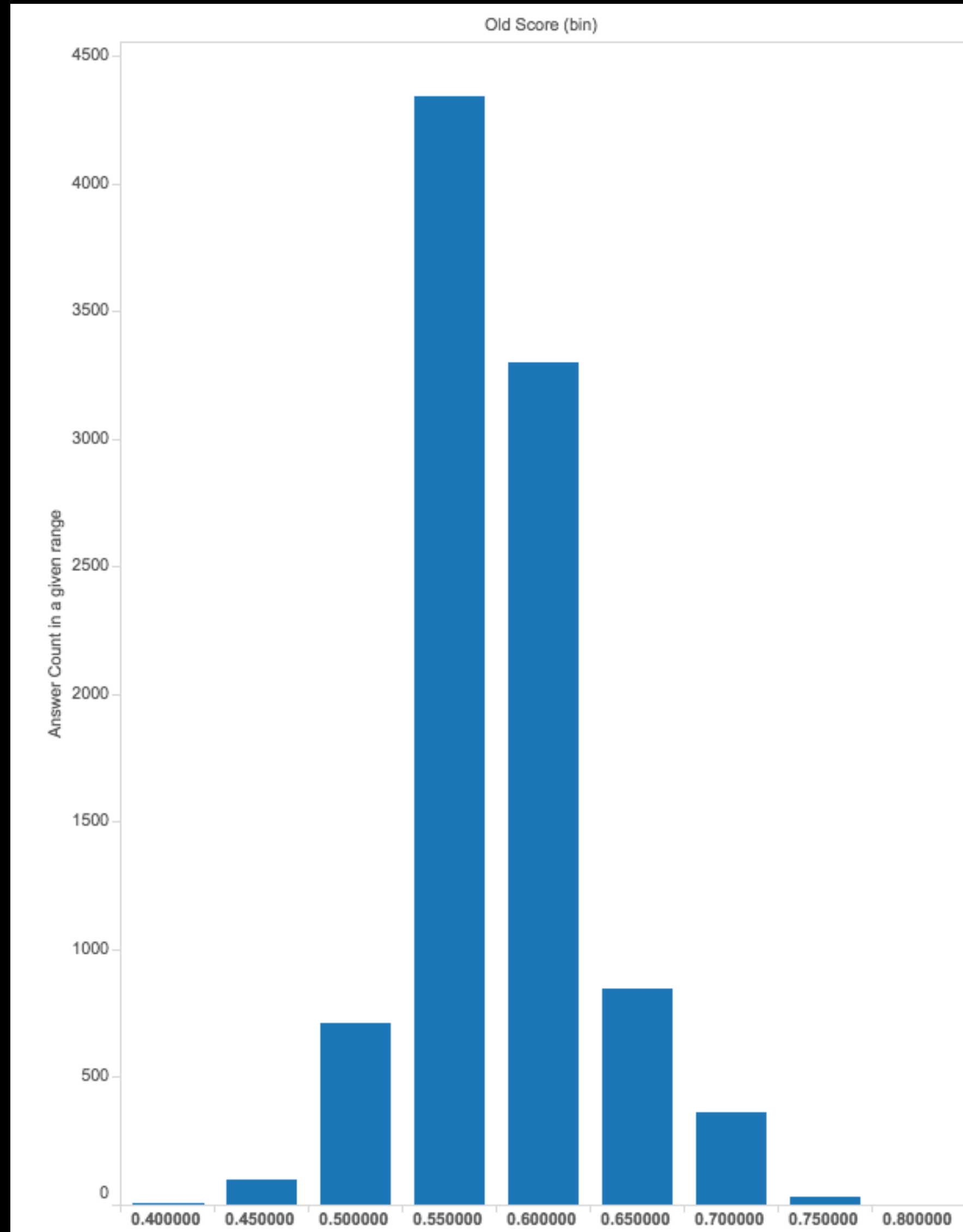
find a representation of the data

pick a learning algorithm

run the training algorithm

evaluate the accuracy

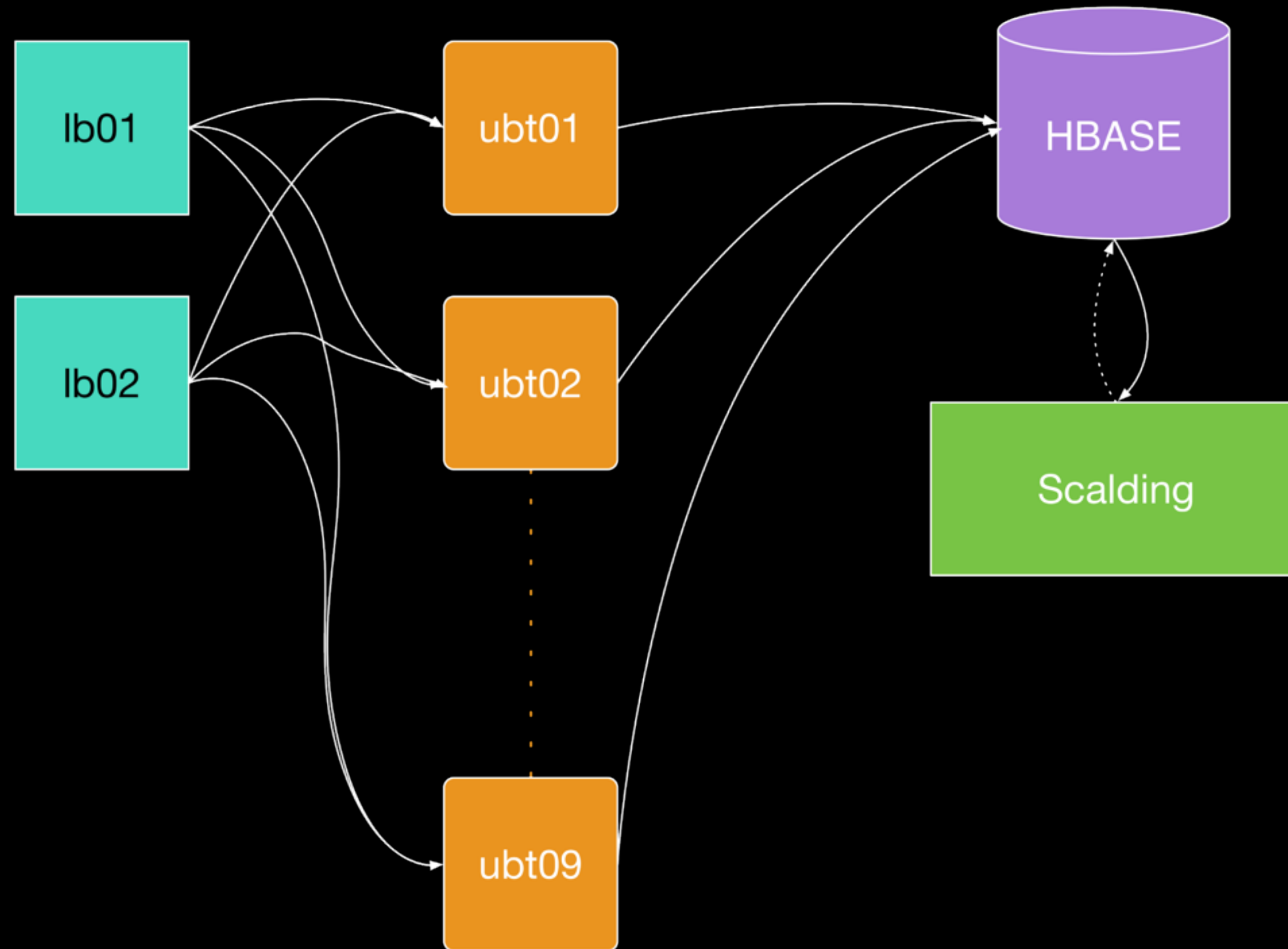
Regression Prototype



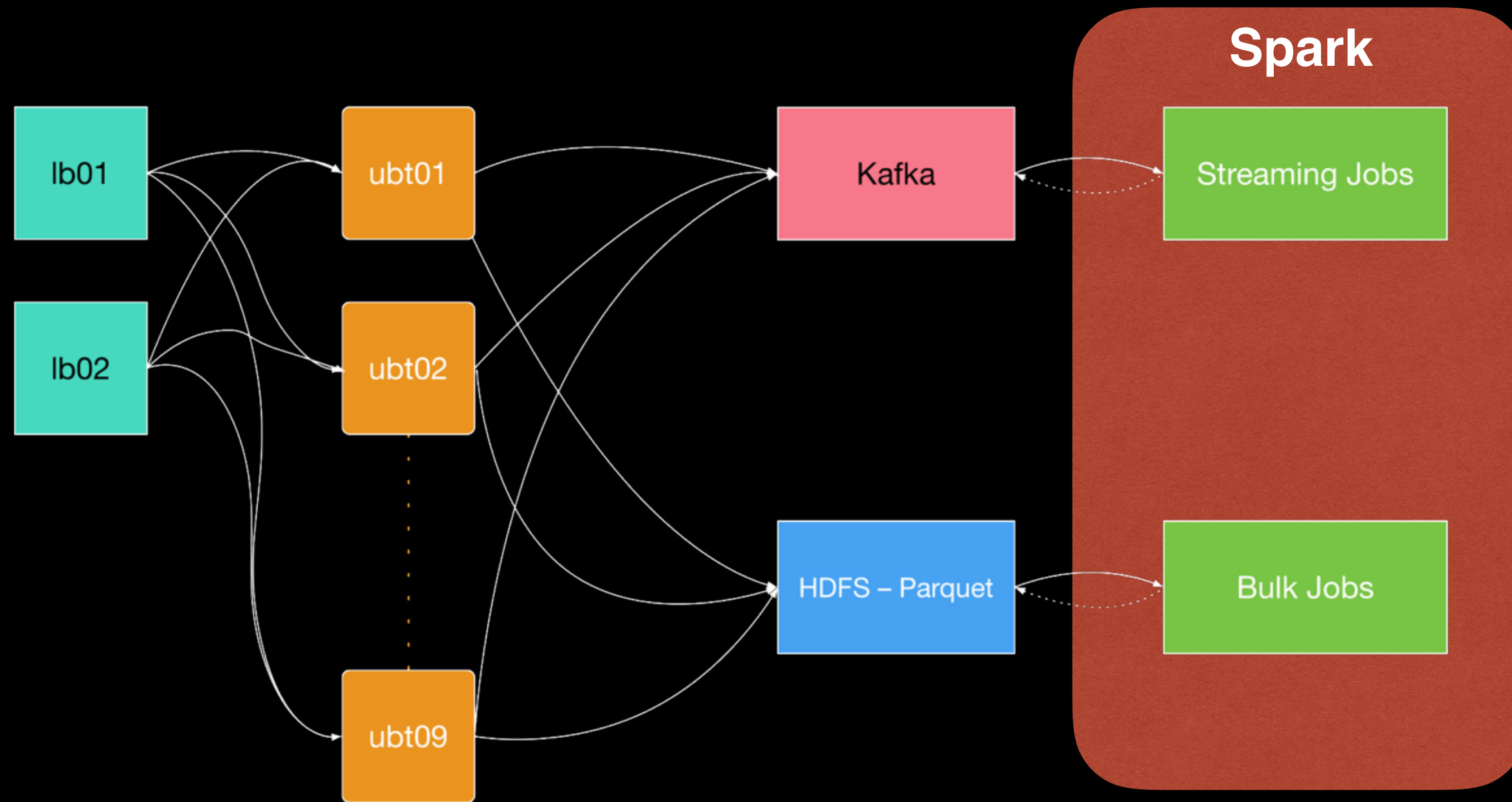
Identify the problems

- Model not complex enough
- Similar inputs, different outputs?
- Not enough training data

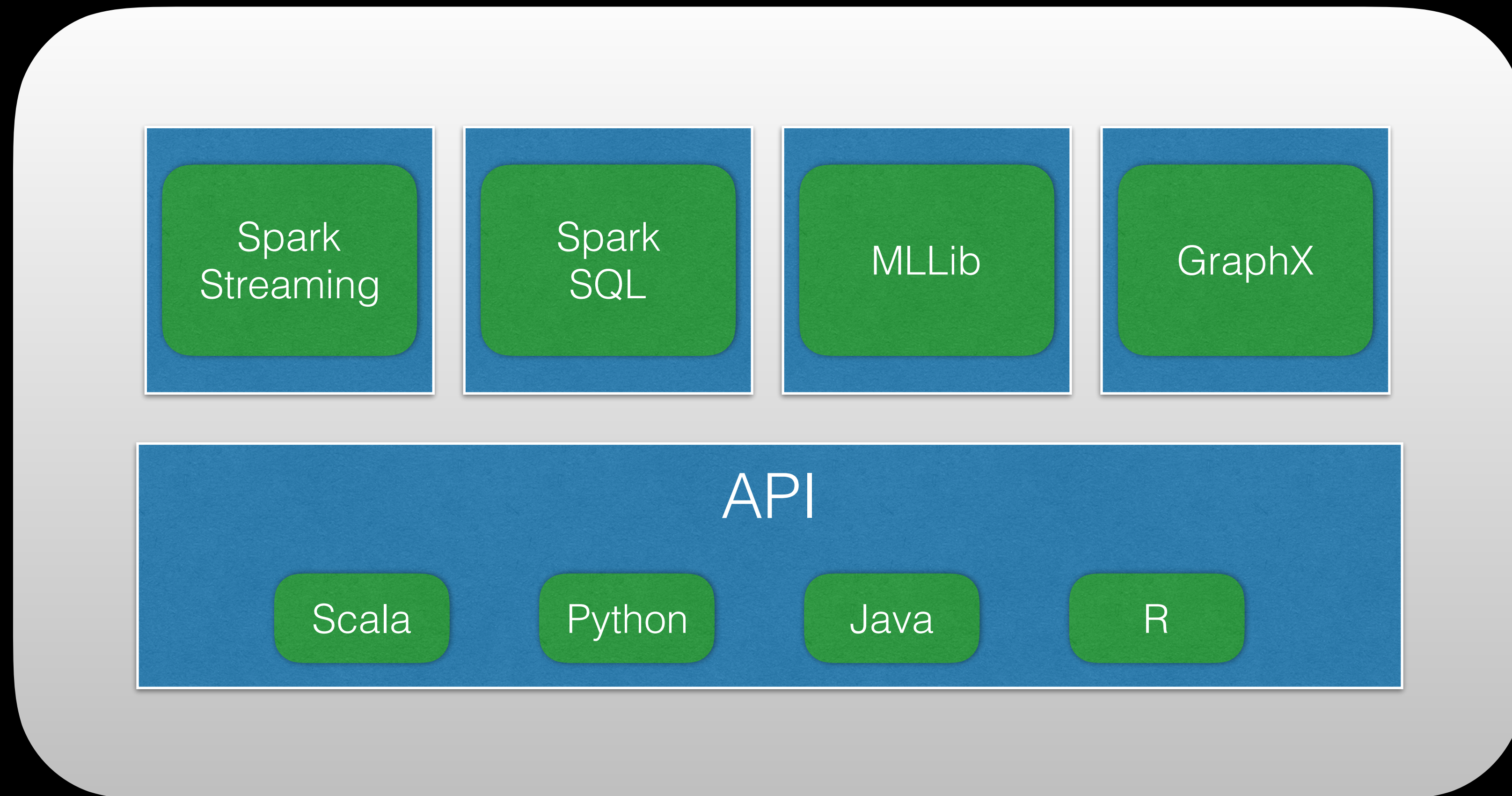
The old ETL pipeline



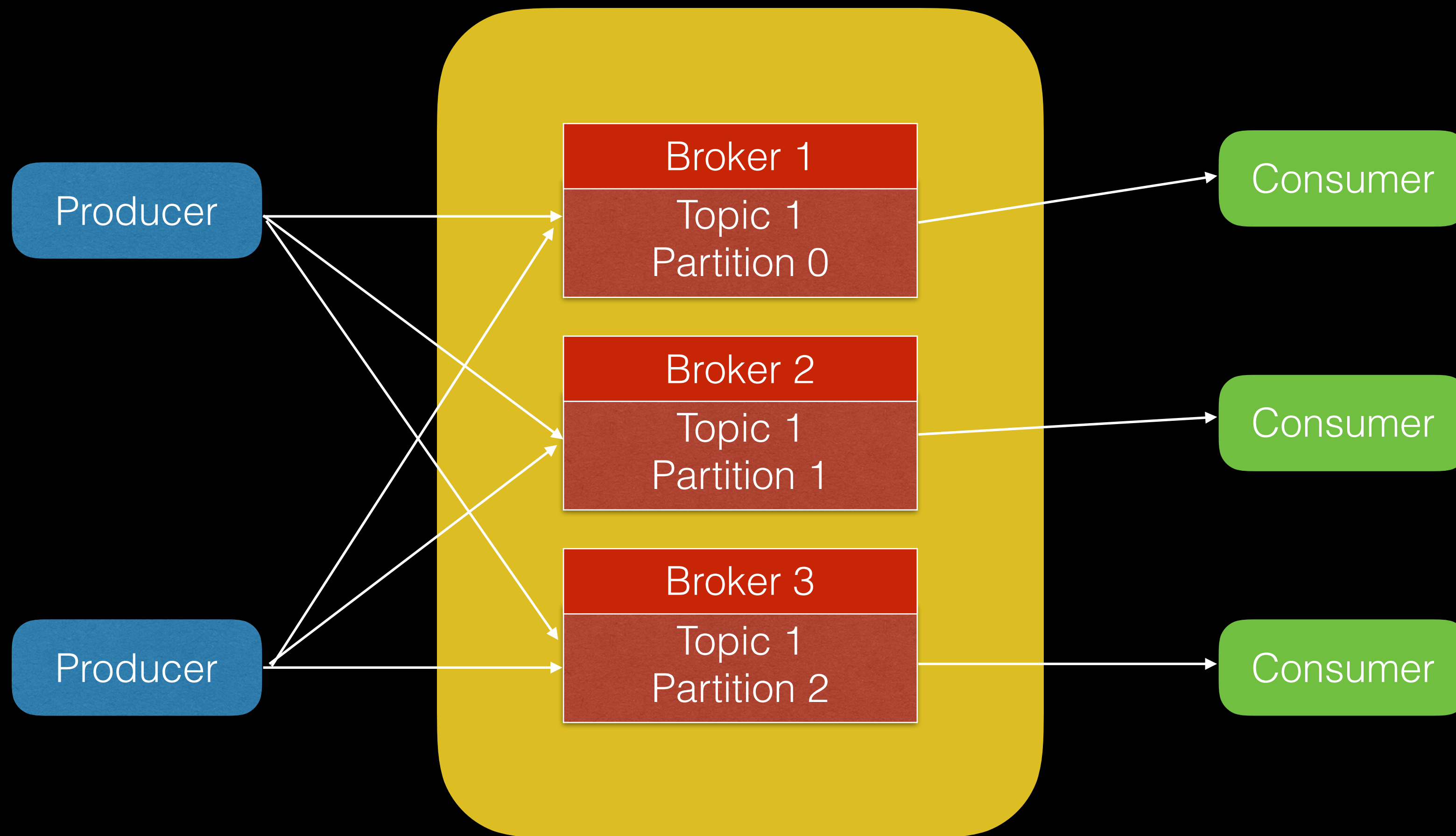
ETL v2



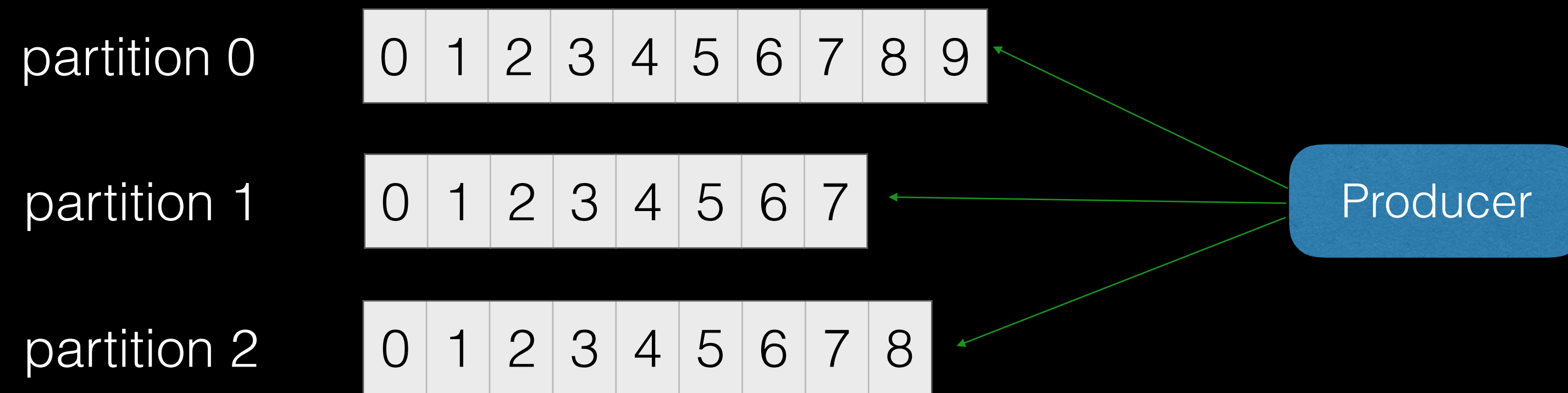
Spark ecosystem



Kafka



Kafka topic



- scale
- parallelism

Parquet

push-down filters

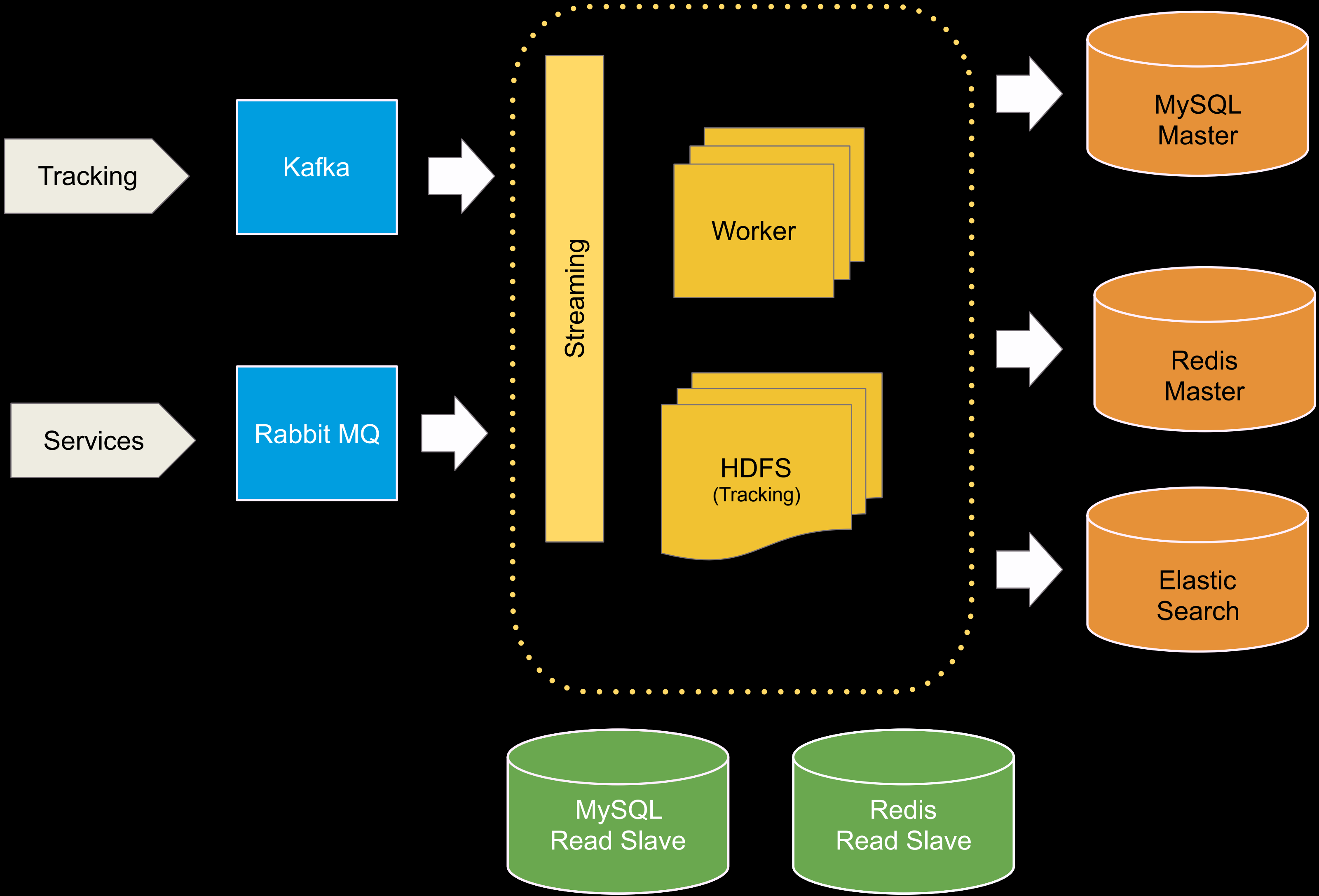
id	cc	votes
1	DE	2
2	DE	3
3	AT	1
4	DE	2

id	cc	votes
1	DE	2
2	DE	3
3	AT	1
4	DE	2


id	cc	votes
1	DE	2
2	DE	3
3	AT	1
4	DE	2

SELECT votes FROM logs WHERE cc = 'AT'

ETL v2



Project Moria


 **A:** Meldungen **3** **Bewertungen** **F:** Änderungen **8** **Hallo** **Abmelden**

Antworten bewerten Bewertete Antworten **2**

Ist zu wenig Watt beim Computer schlecht?


Hallo, ich werde bald eine neue Grafikkarte kaufen, die 60 Watt mehr braucht als meine alte. Kann etwas kaputt gehen oder wird einfach die Leistung limitiert?

Danke im Voraus

von  vor 4 Stunden [Frage in gutefrage.net](#)
[Editieren](#)

Antwort

Hey ho,
das kommt ganz auf dein Netzteil an^^.
kaputt gehen kann dabei nichts.
MfG

von  vor 4 Stunden [Antwort in gutefrage.net](#)

★ Perfekt **Gute Antwort** **Geht so** **Schlechte Antwort** **Löschen** **Nächste Antwort**

Moria v1.325 © gutefrage.net 2015

Clean training data?



Project Angmar

- tried lots of different supervised learning methods
- feature engineering - most crucial part
- analyse the domain, chart everything

Features

Content
length
syntactic complexity
number of links
probability of deletion

Social
votes
most helpful answer
number of comments
answered by expert

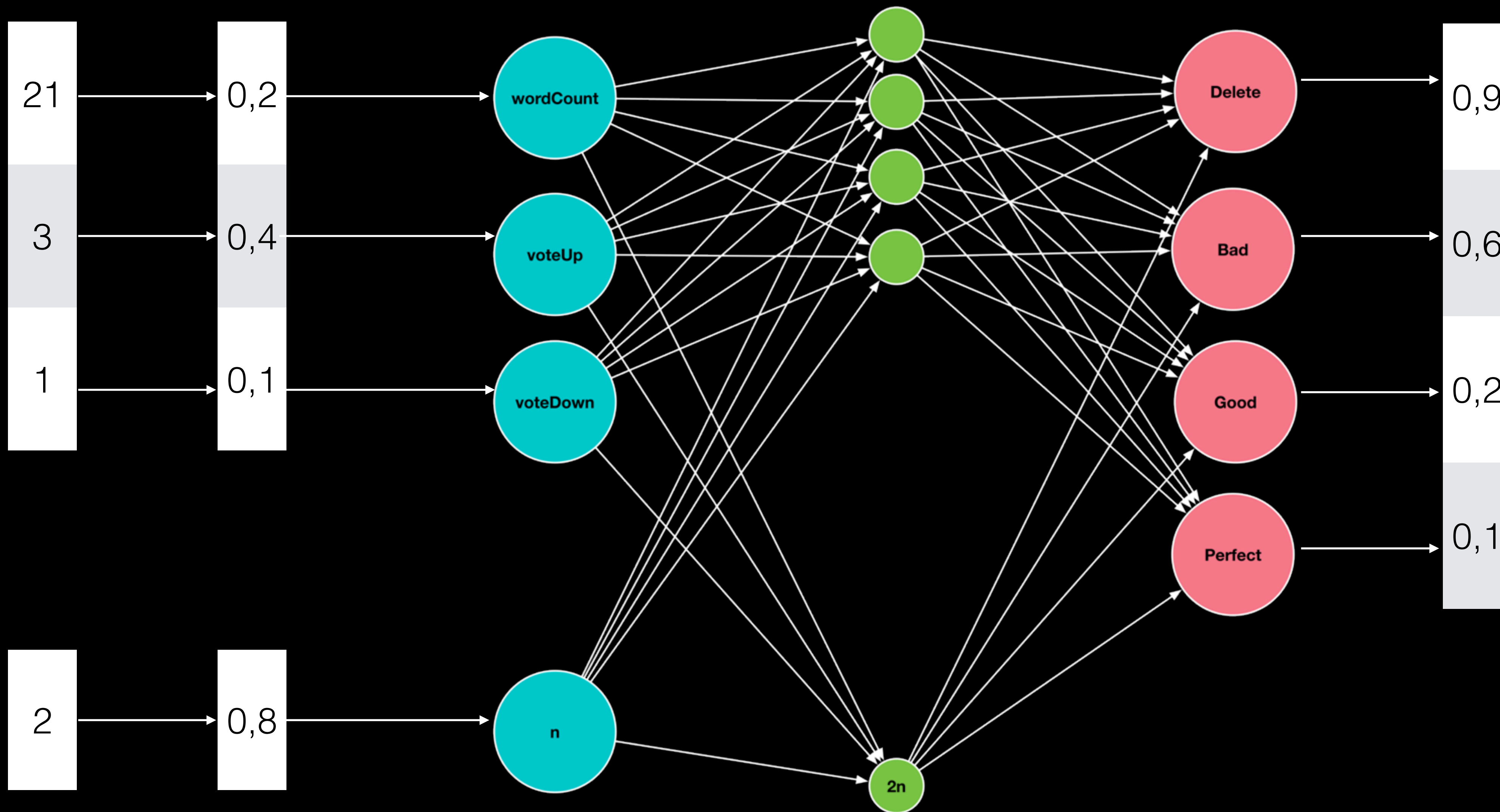
Author
gained votes
credibility score
role
deleted answer ratio
number of answers
number of comments
reported answer ratio

The structure of the network

Answer
vector

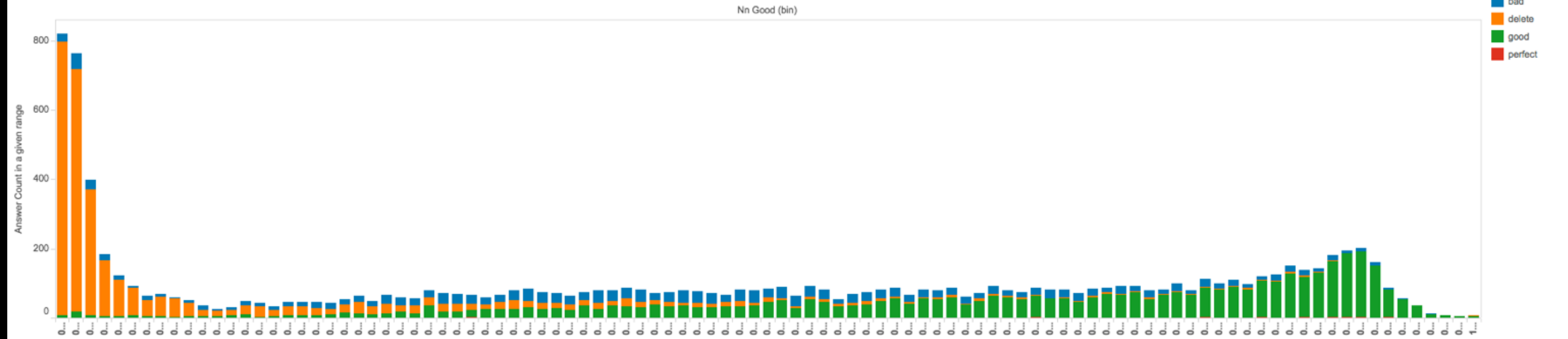
AV
normalized

Output

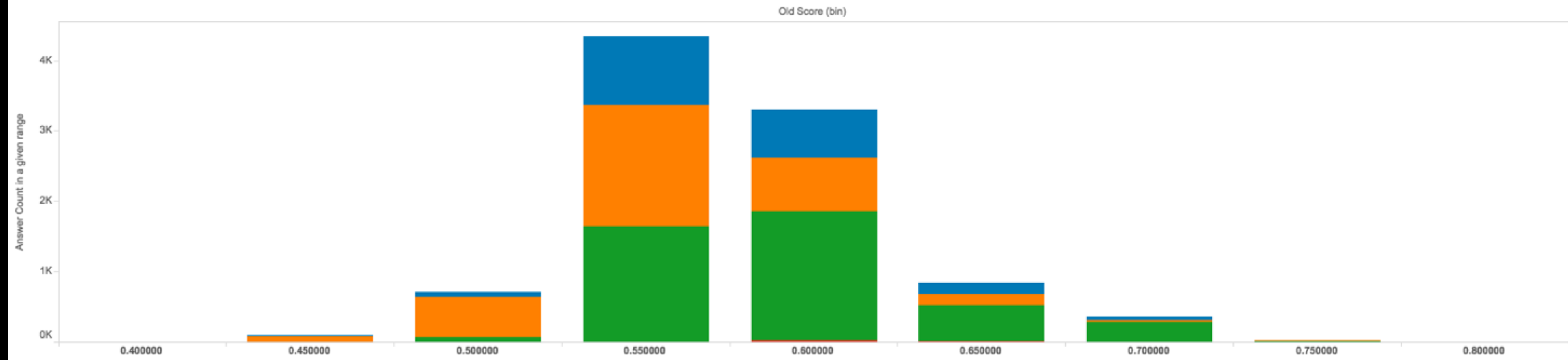


The Result

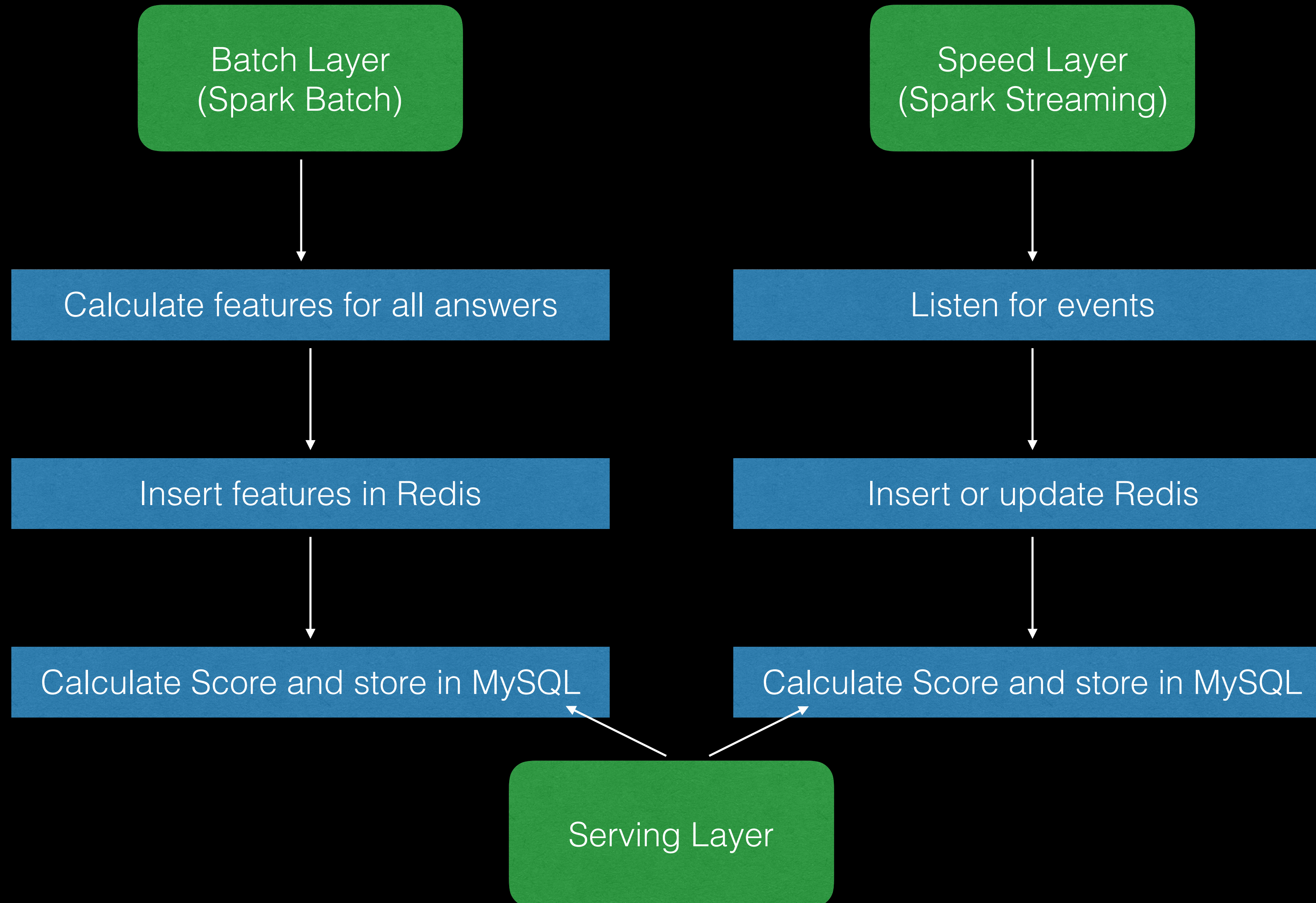
Neural Network based score-prototype



Current score



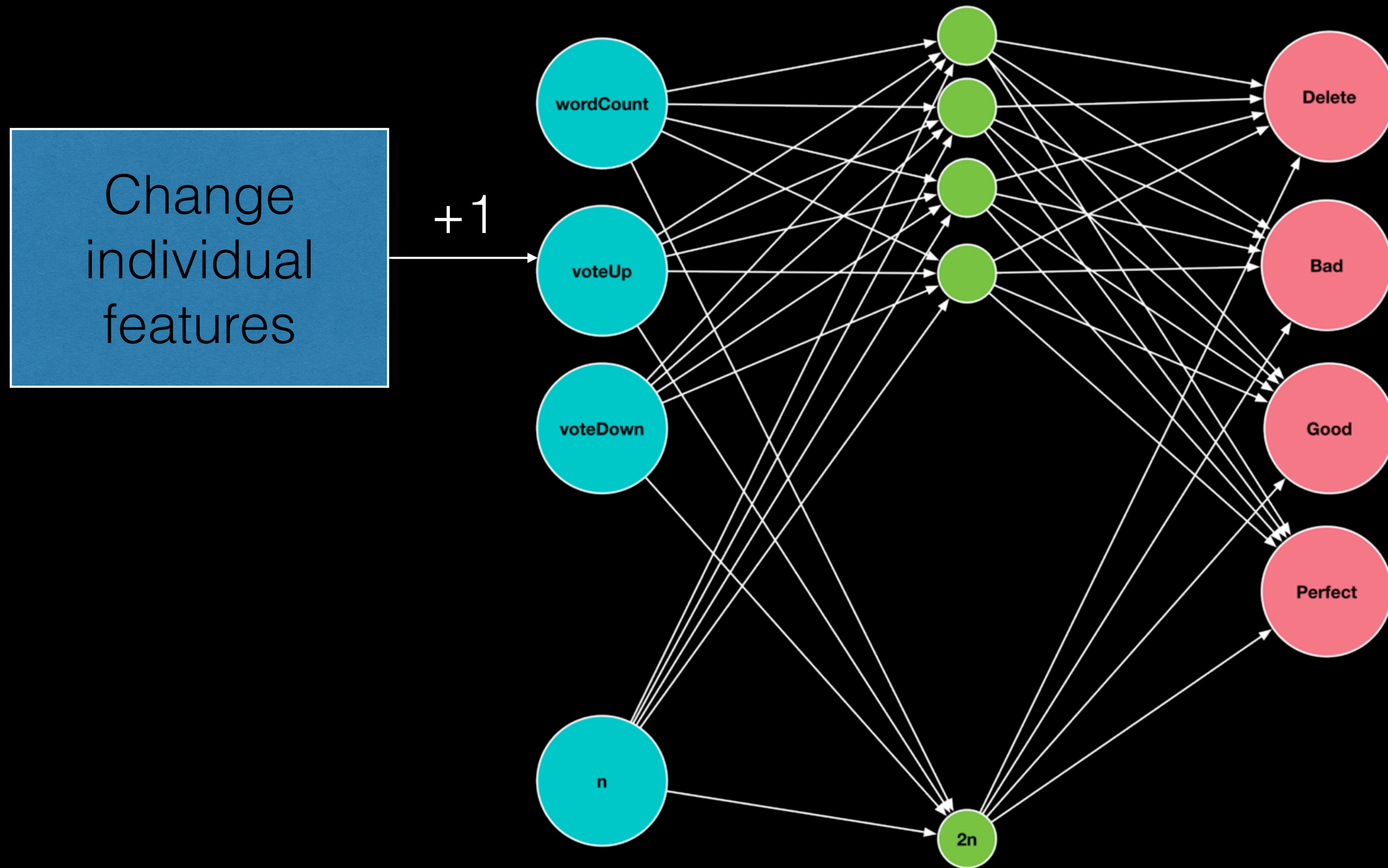
Lambda Architecture



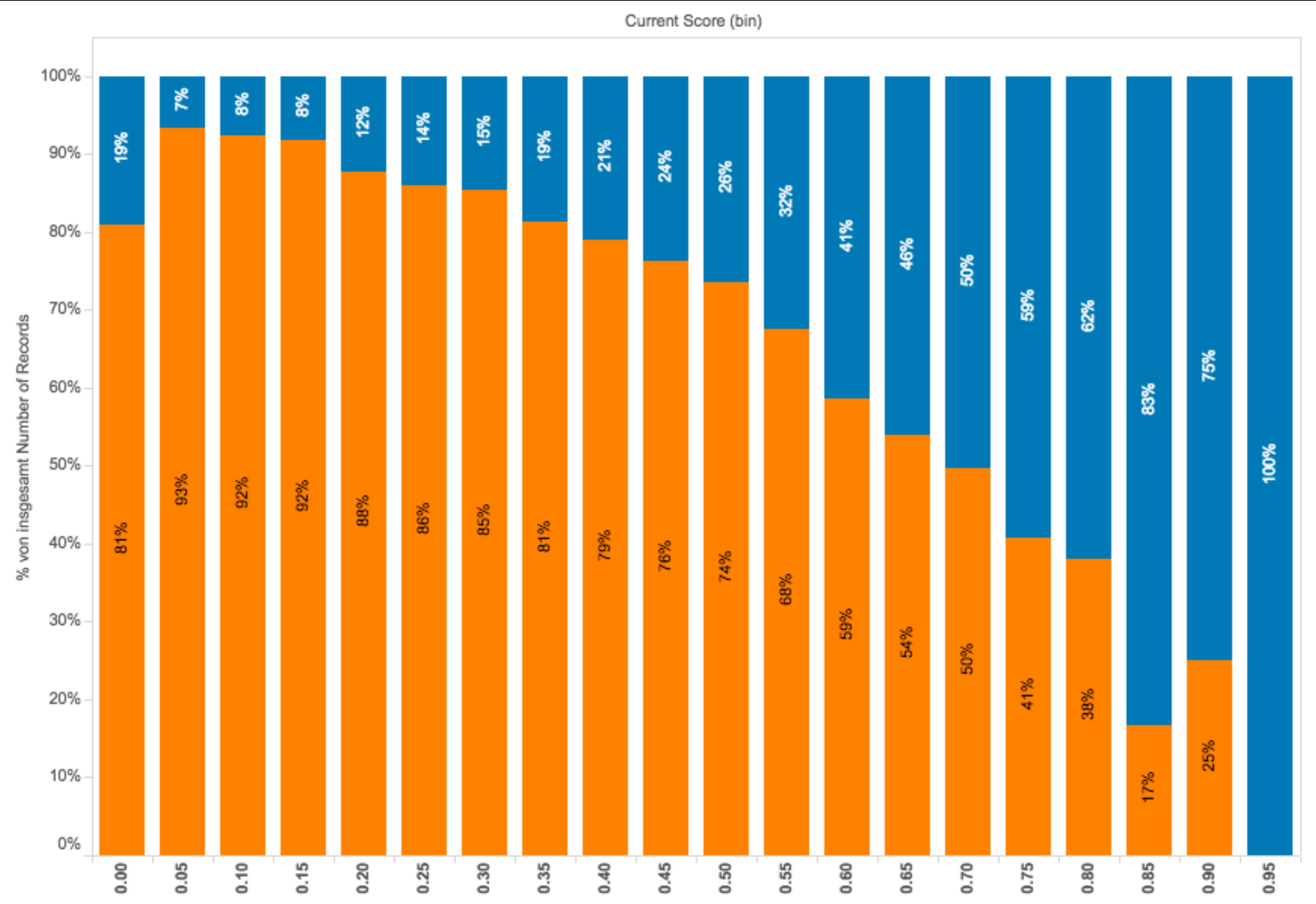
Back pressure

- Bulk jobs insert too fast
- MySQL: sendQueue size, threads connected
- Elasticsearch: load on the instance creating the new index

Debugging the network



real world test (deleted vs non-deleted)



deleted

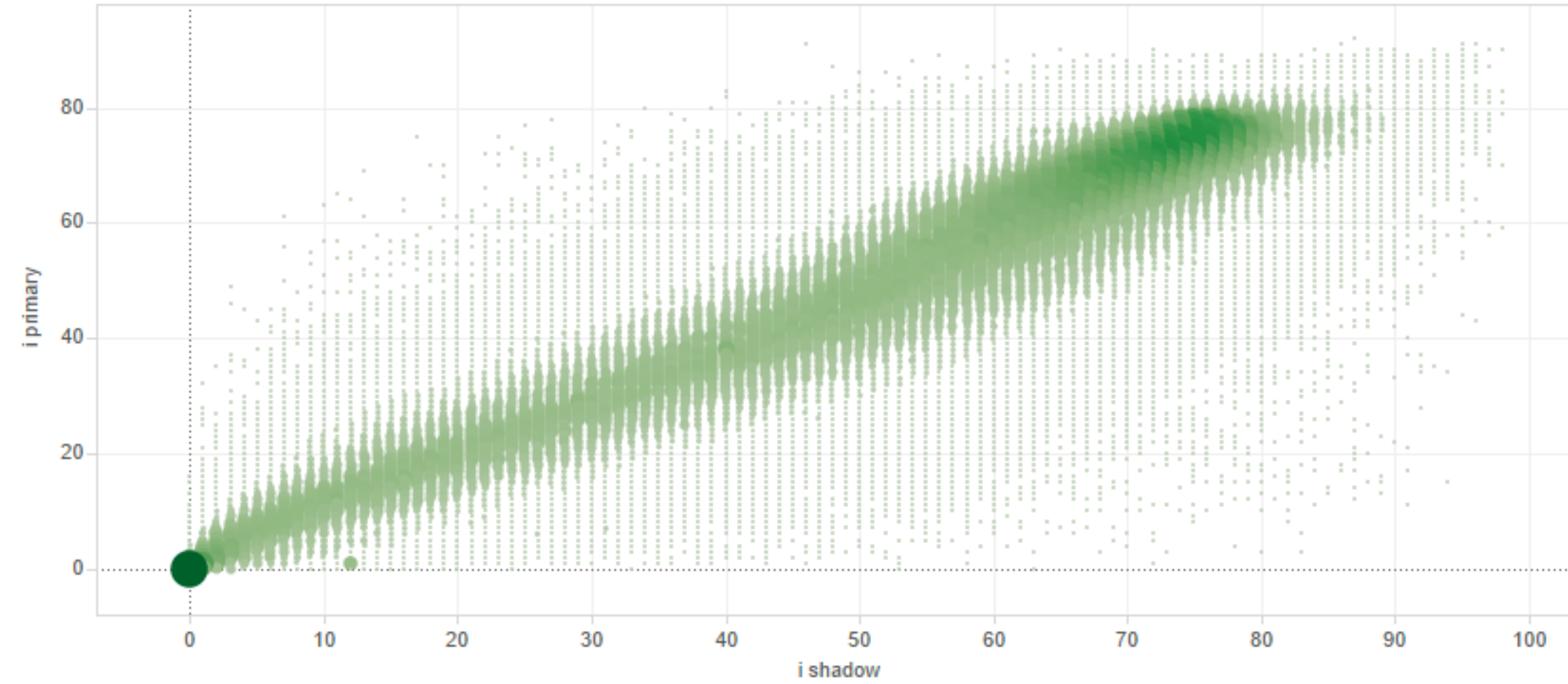
non-deleted

Switching models

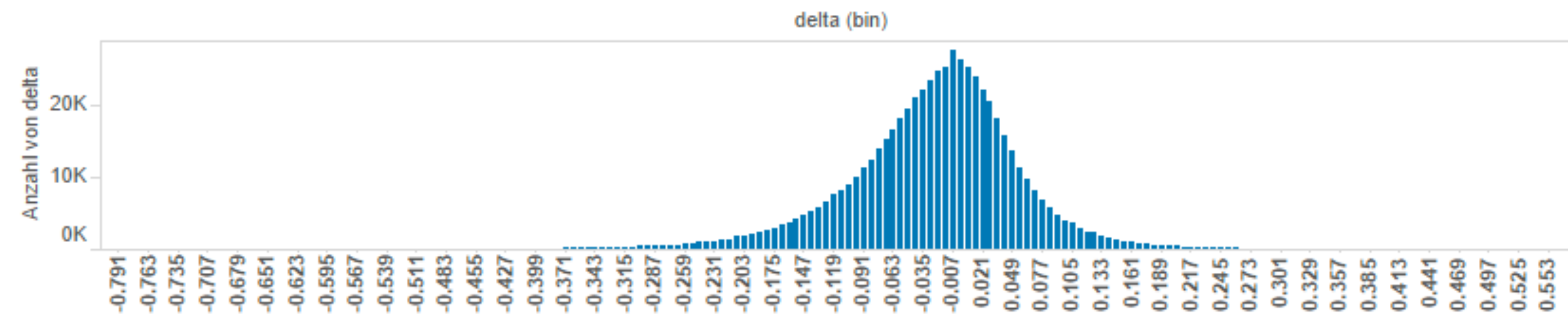
Amount of questions for a score range



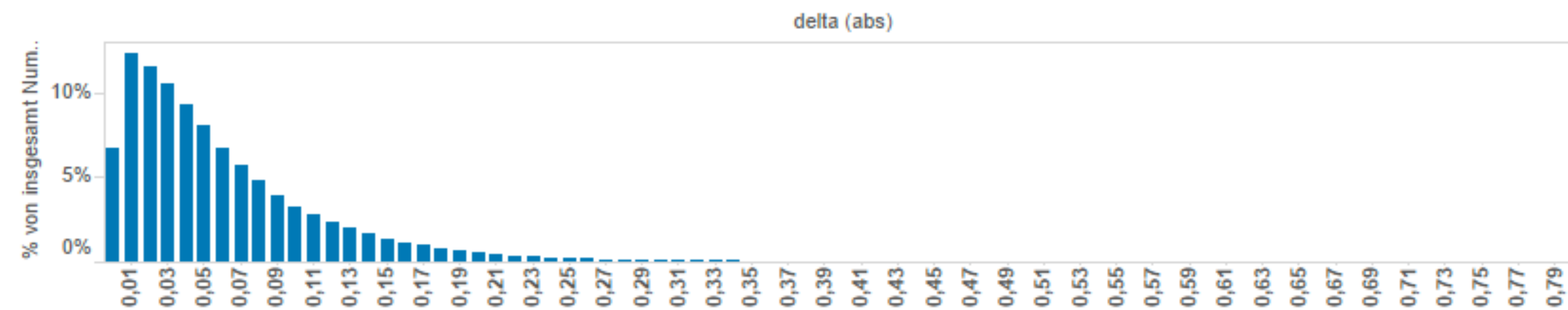
Primary vs Shadow Scatter



Pos Neg Change



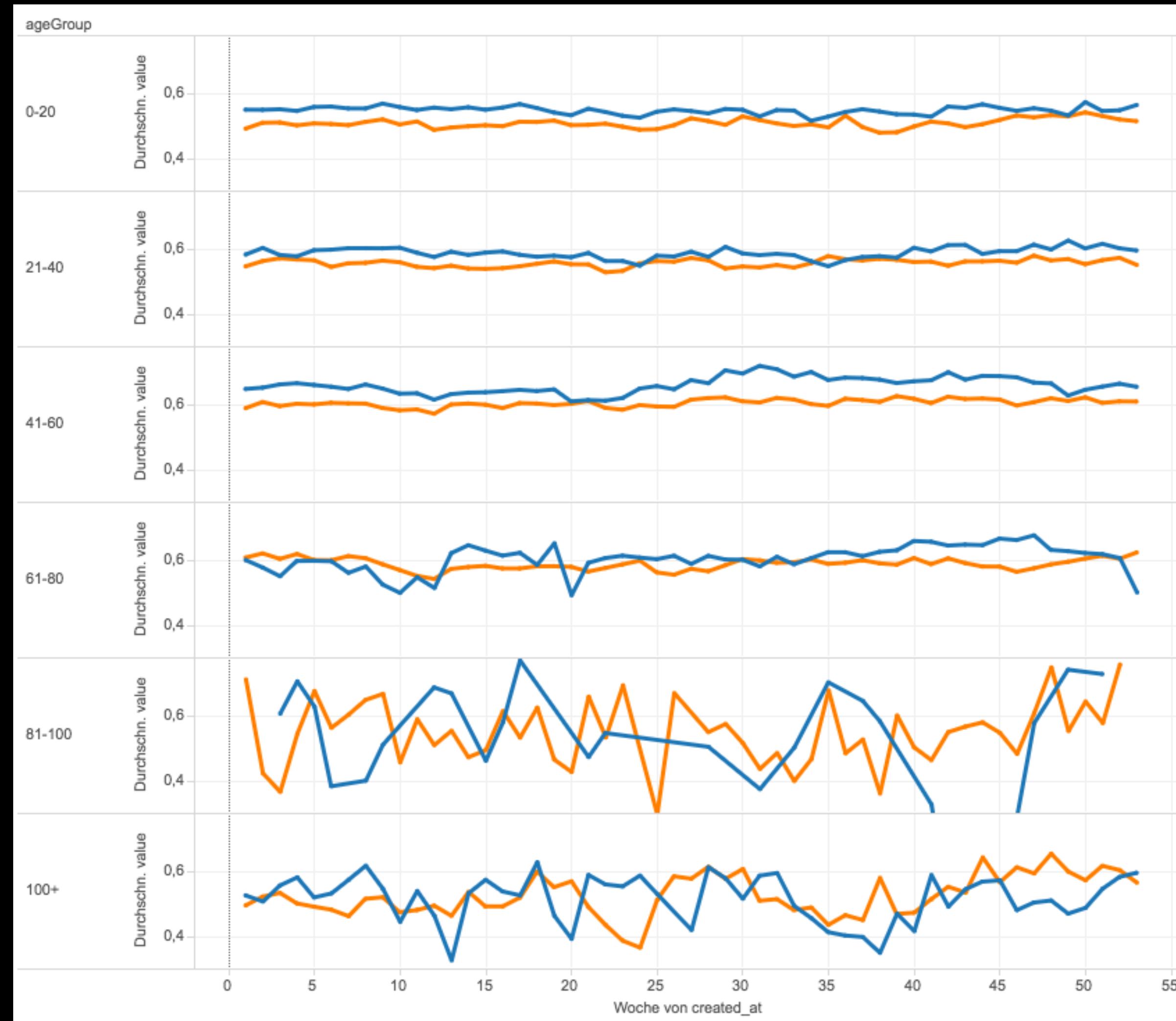
Abs change



Insights

Women

Men



Learnings

- If your use case is complex, you need a complex model
- If you have a complex model, you need lots of data
- If you have lots of data, you need an ETL pipeline that can process huge amounts of data fast
- Think about your use case first, then design the pipeline

Questions?

You can ask them on [gutefrage](#) too :)