Information Retrieval

Exercise – Winter term 2024/2025

tim.hagen@uni-kassel.de

Agenda

1. Project Status
2. Indexing Process
3. Assignment

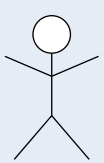
Project Status

Each team, shortly give one example of your topics.
What was the intention behind creating it?

What was the most difficult part?

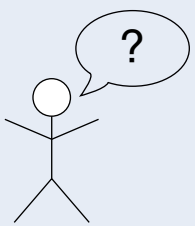
Architecture of a Search Engine

Karaoke Version



Architecture of a Search Engine

Karaoke Version

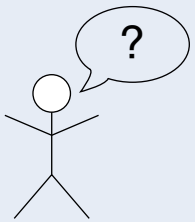


Architecture of a Search Engine

Karaoke Version

Indexing

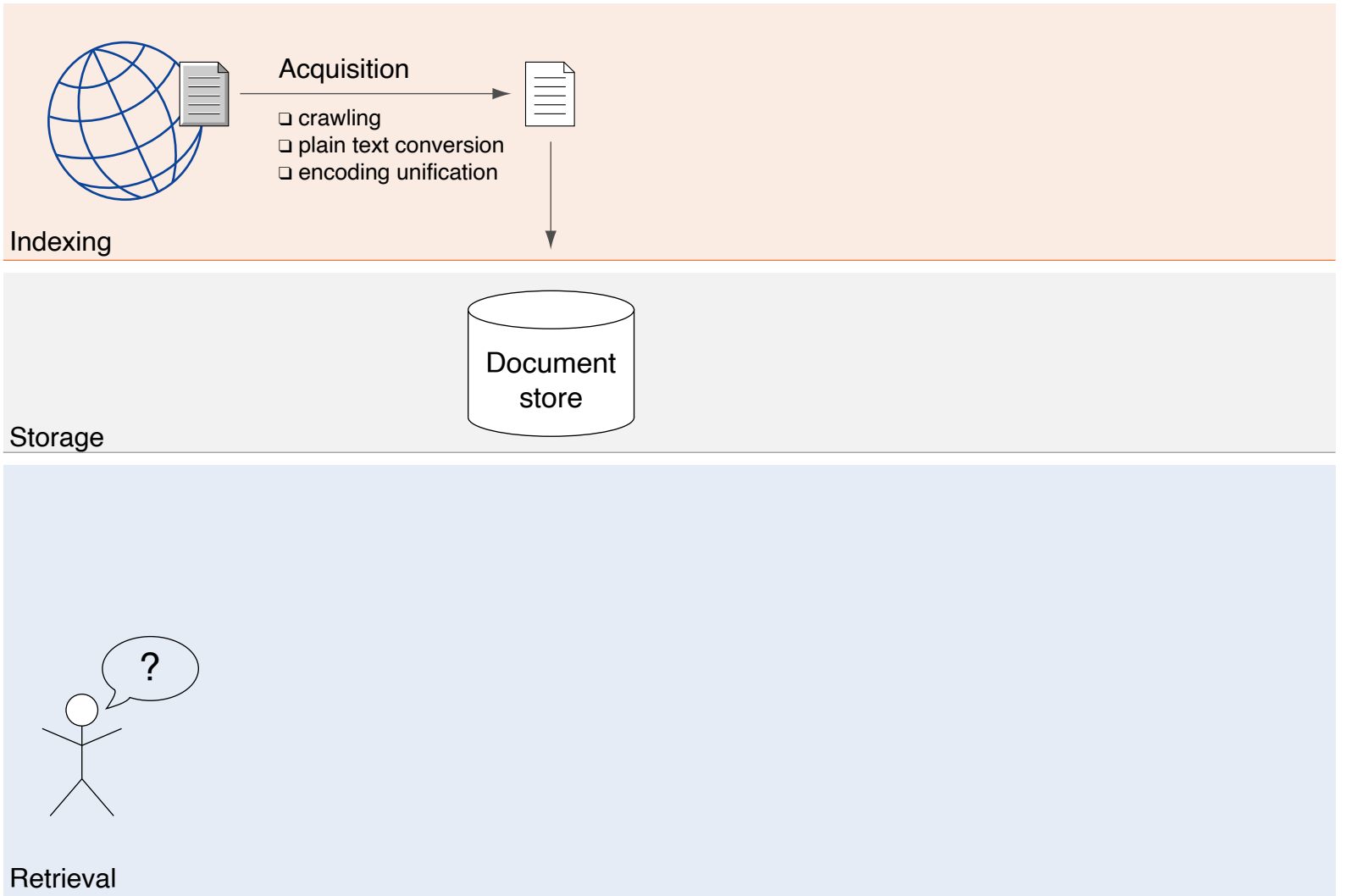
Storage



Retrieval

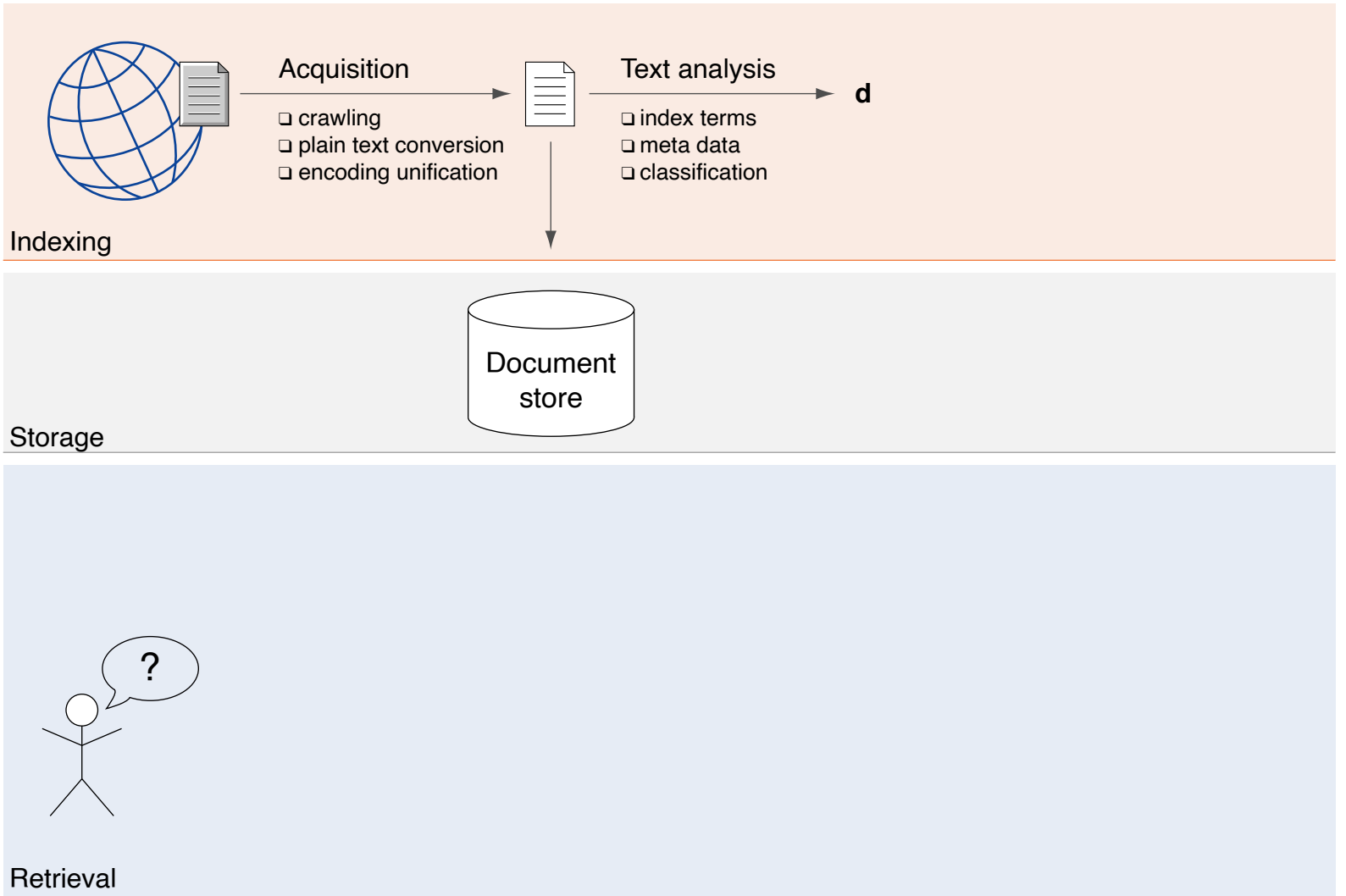
Architecture of a Search Engine

Karaoke Version



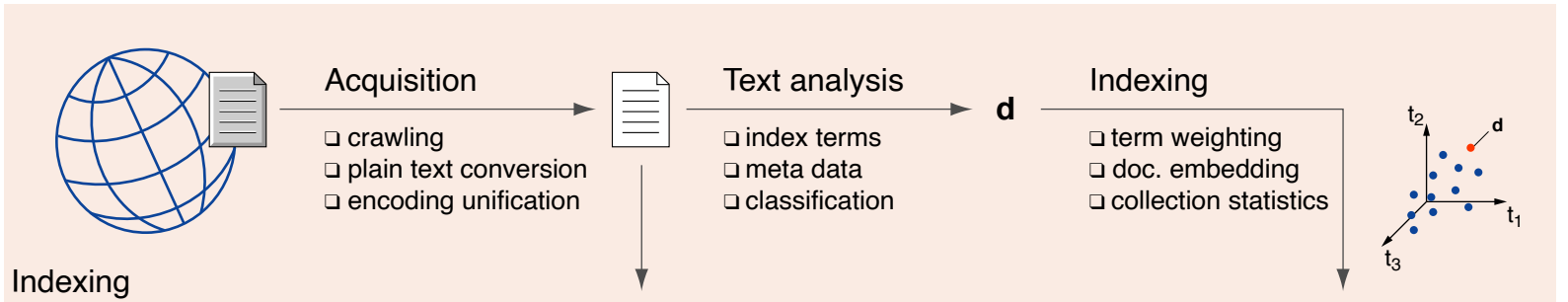
Architecture of a Search Engine

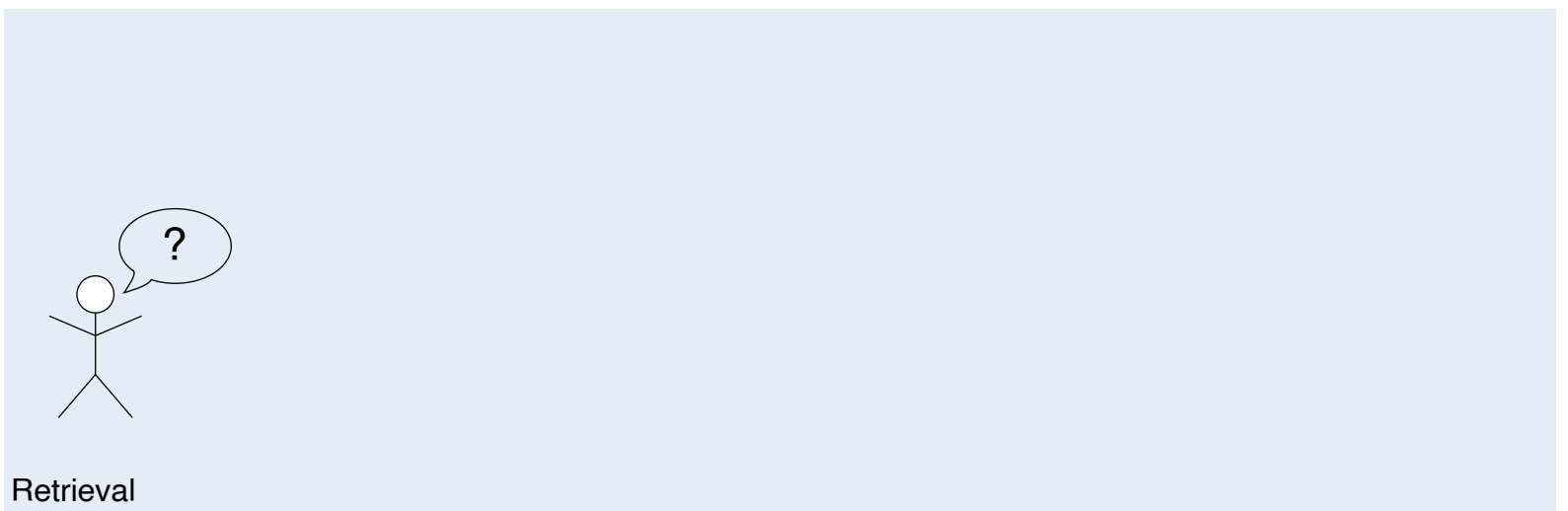
Karaoke Version



Architecture of a Search Engine

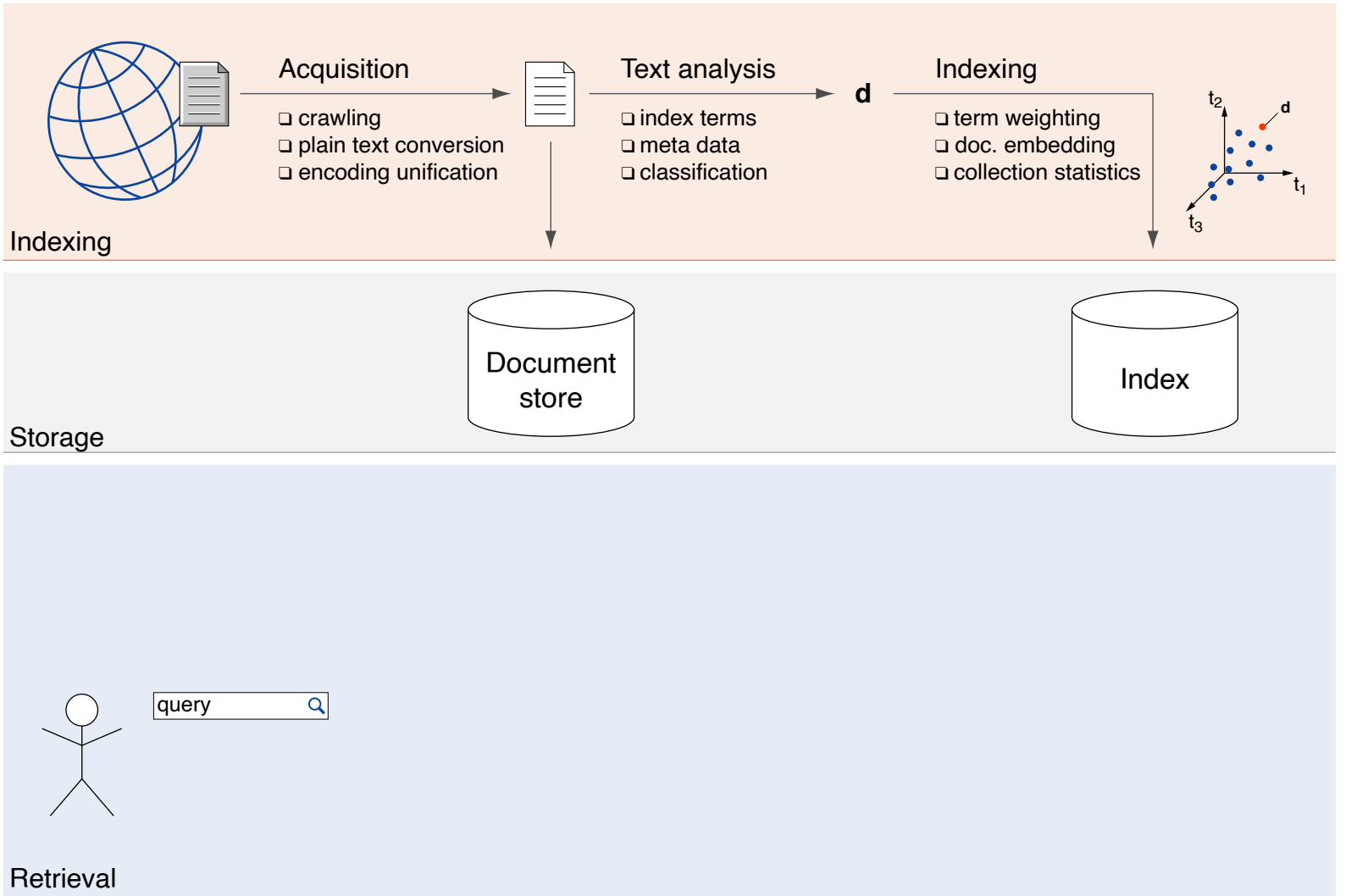
Karaoke Version





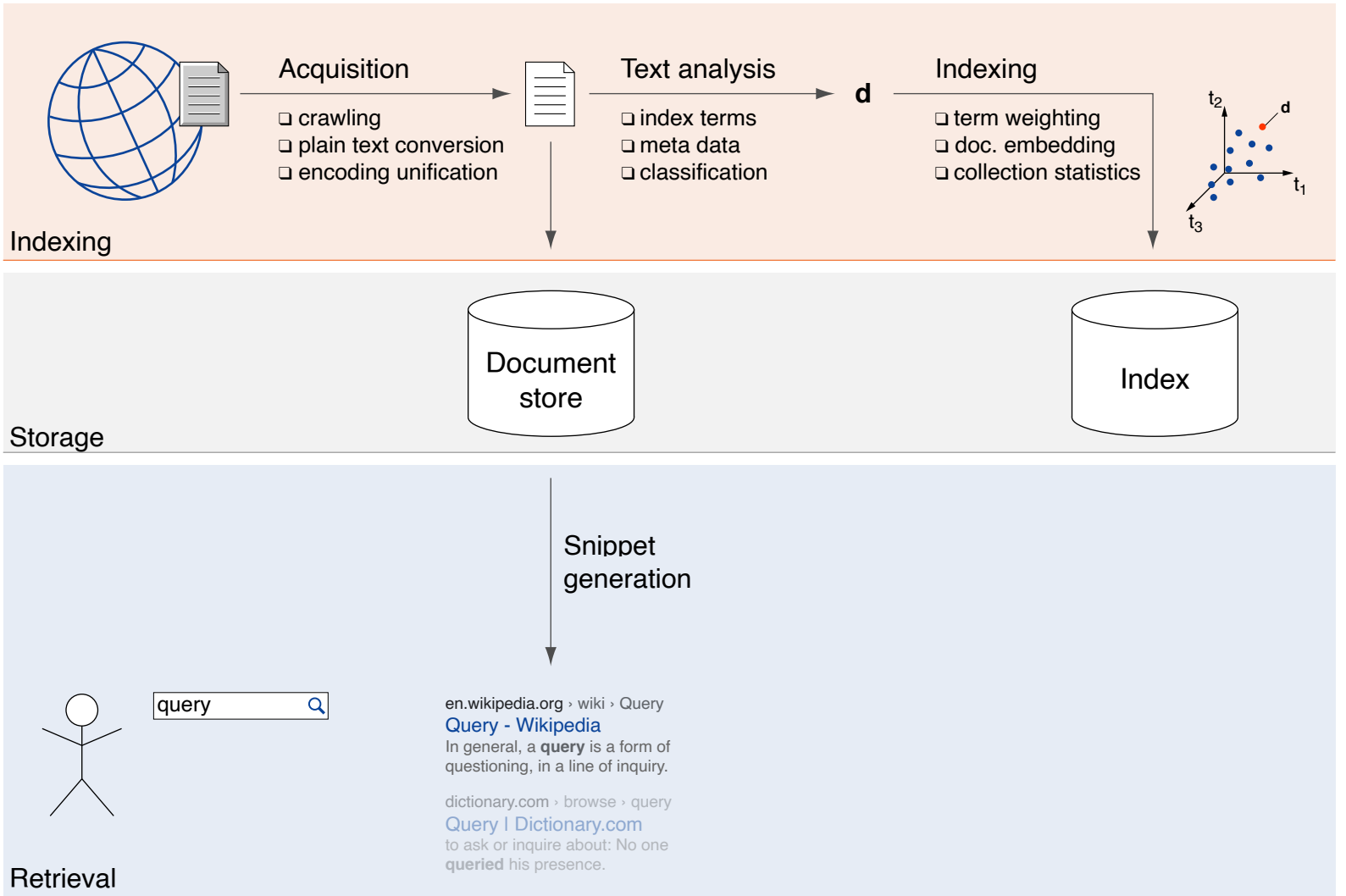
Architecture of a Search Engine

Karaoke Version



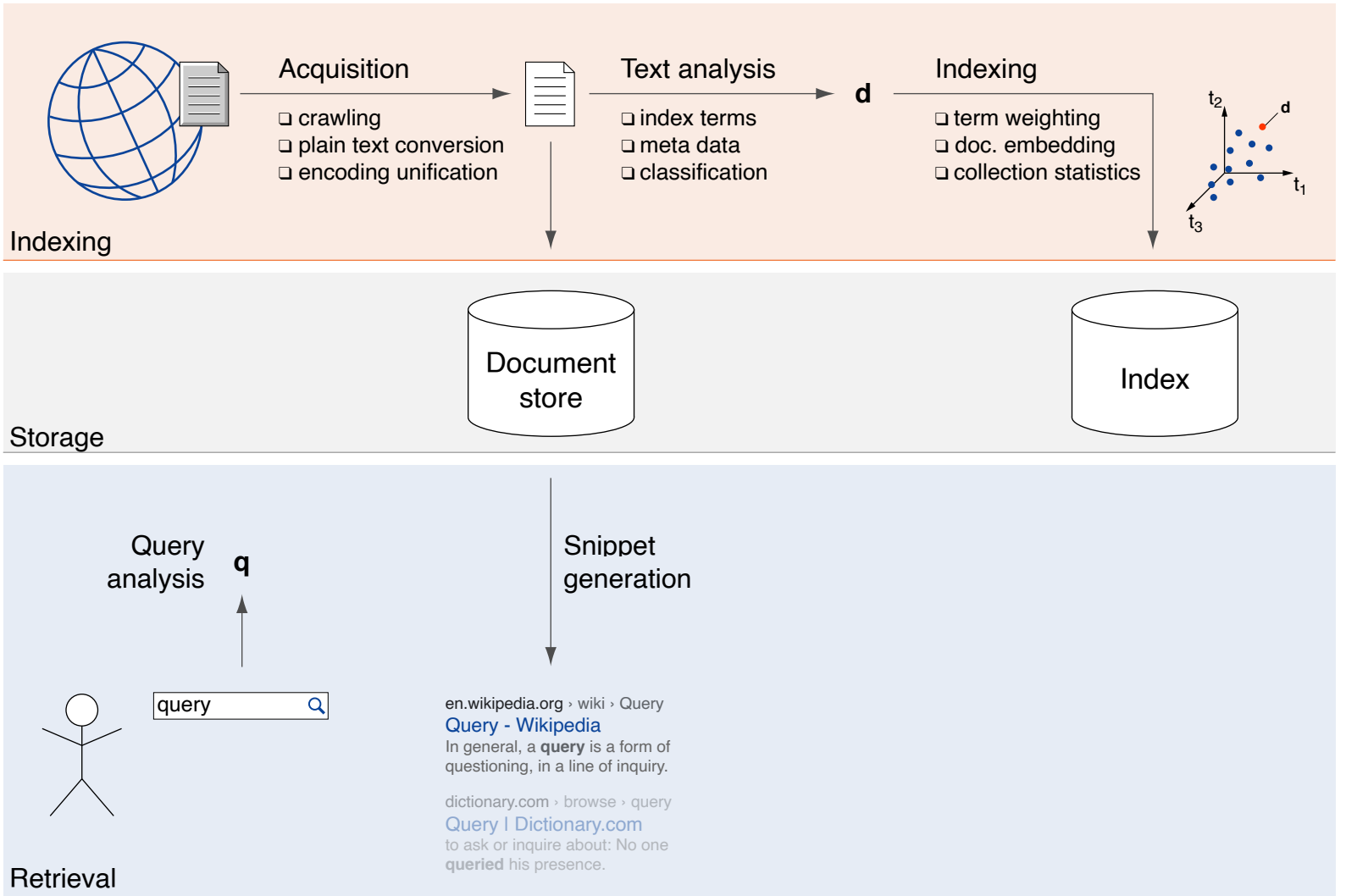
Architecture of a Search Engine

Karaoke Version



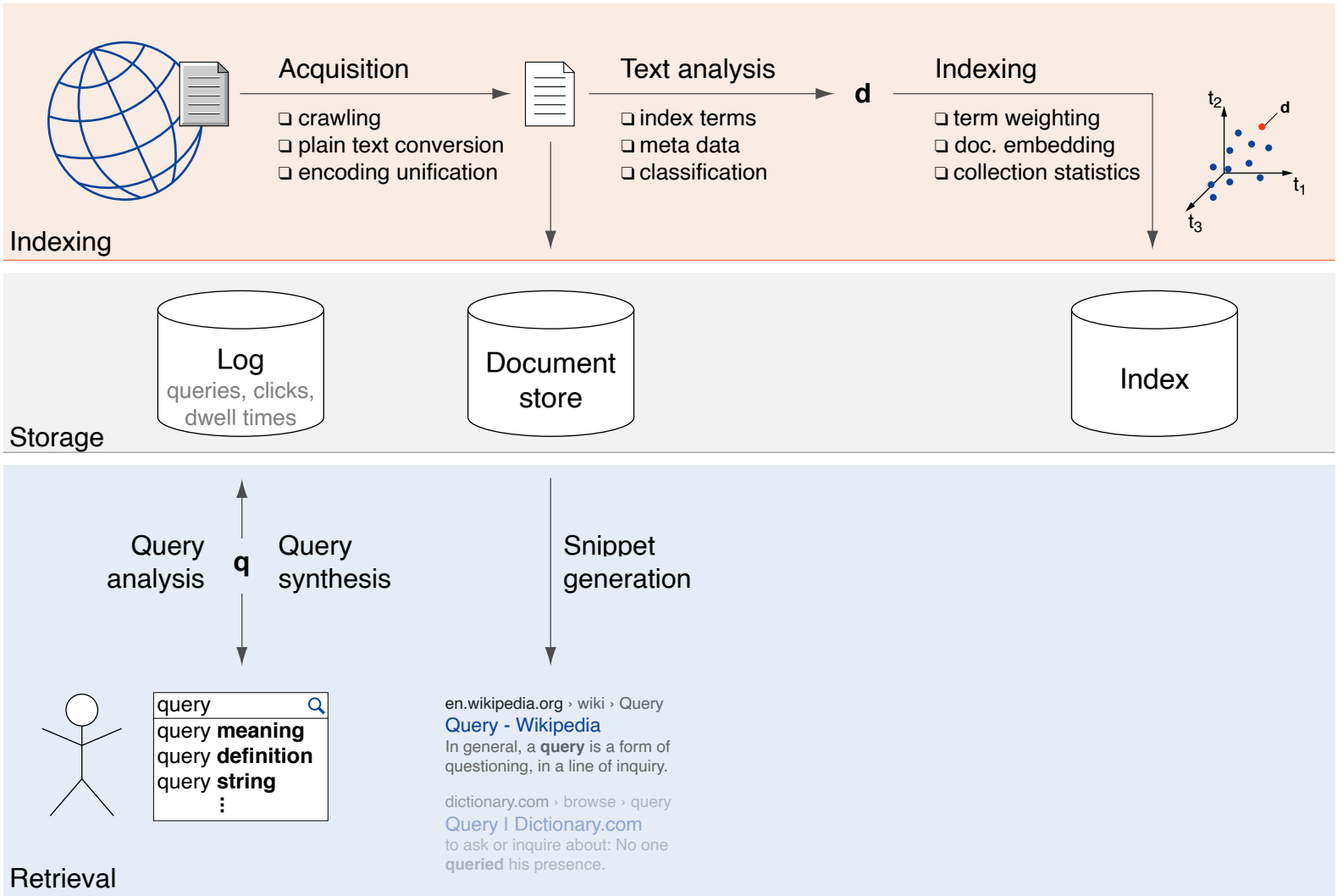
Architecture of a Search Engine

Karaoke Version



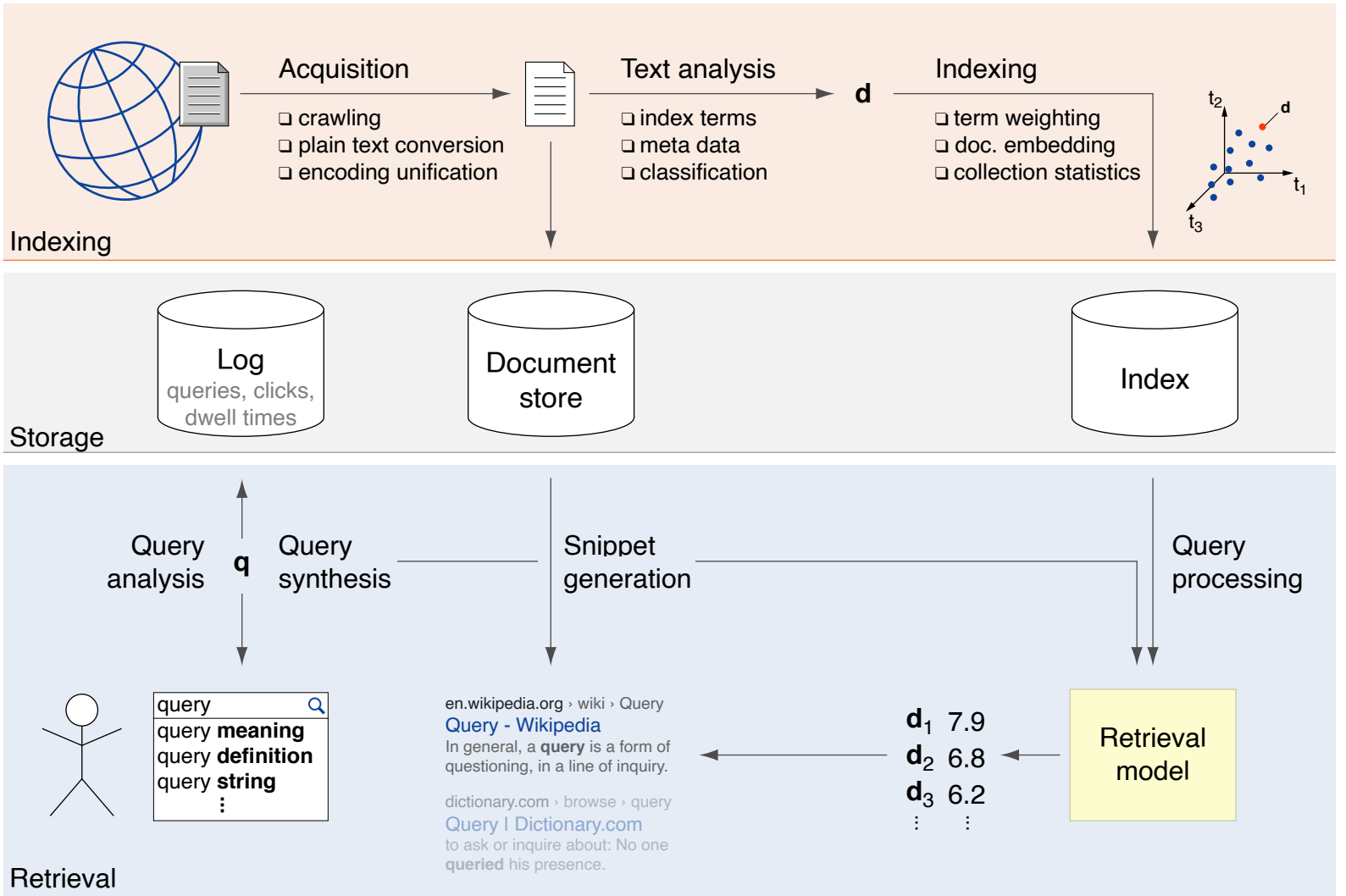
Architecture of a Search Engine

Karaoke Version



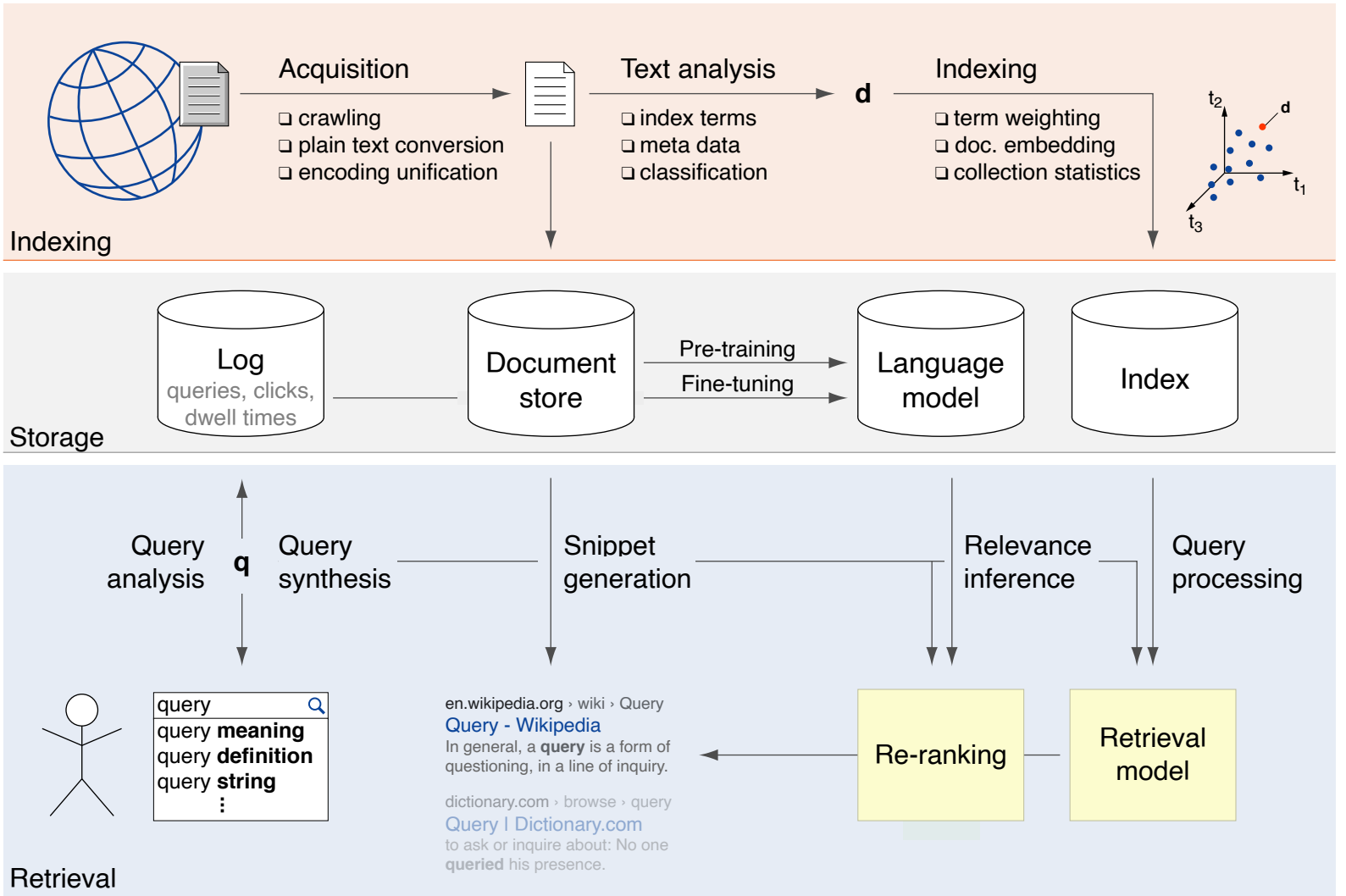
Architecture of a Search Engine

Karaoke Version



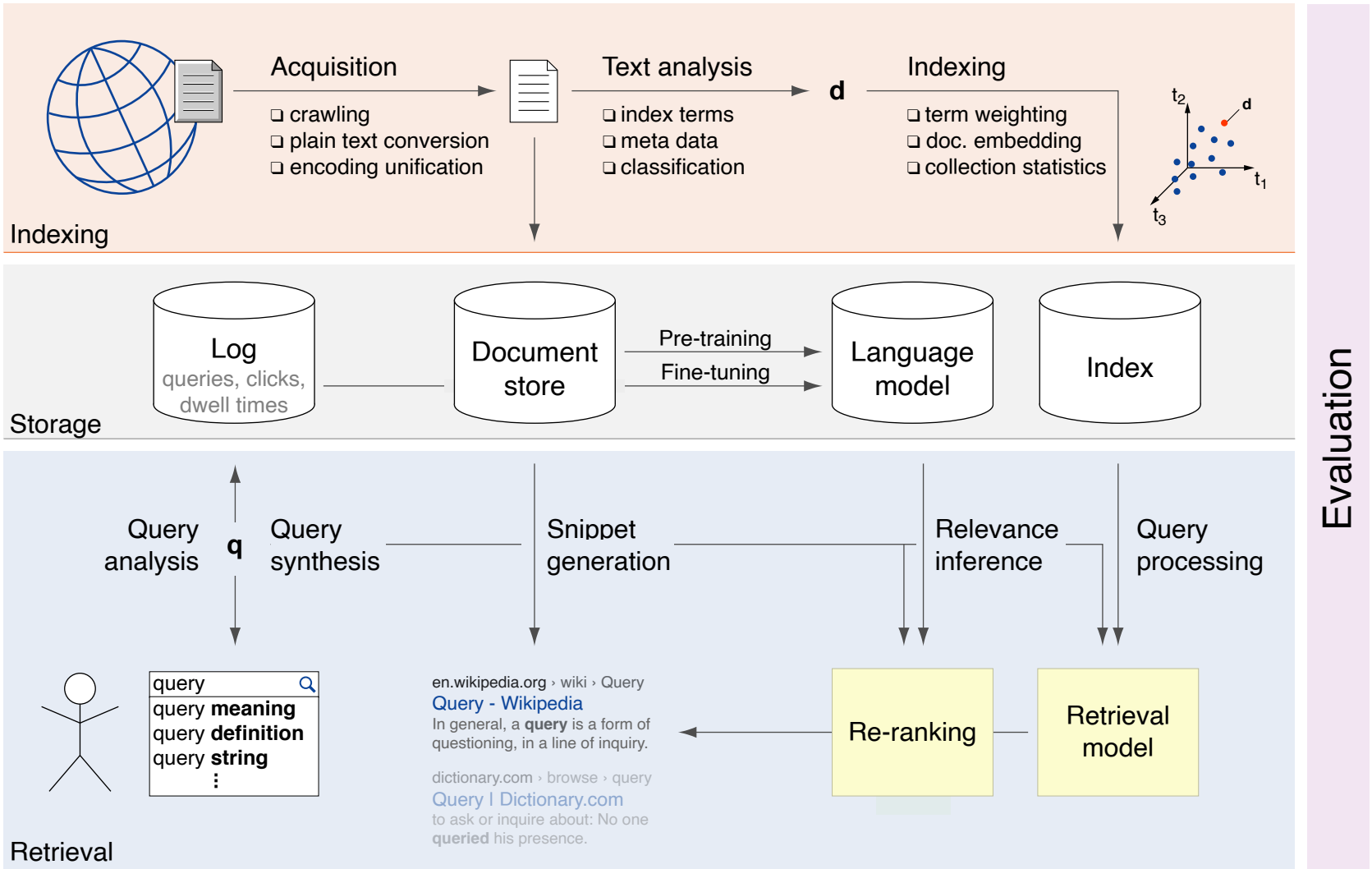
Architecture of a Search Engine

Karaoke Version



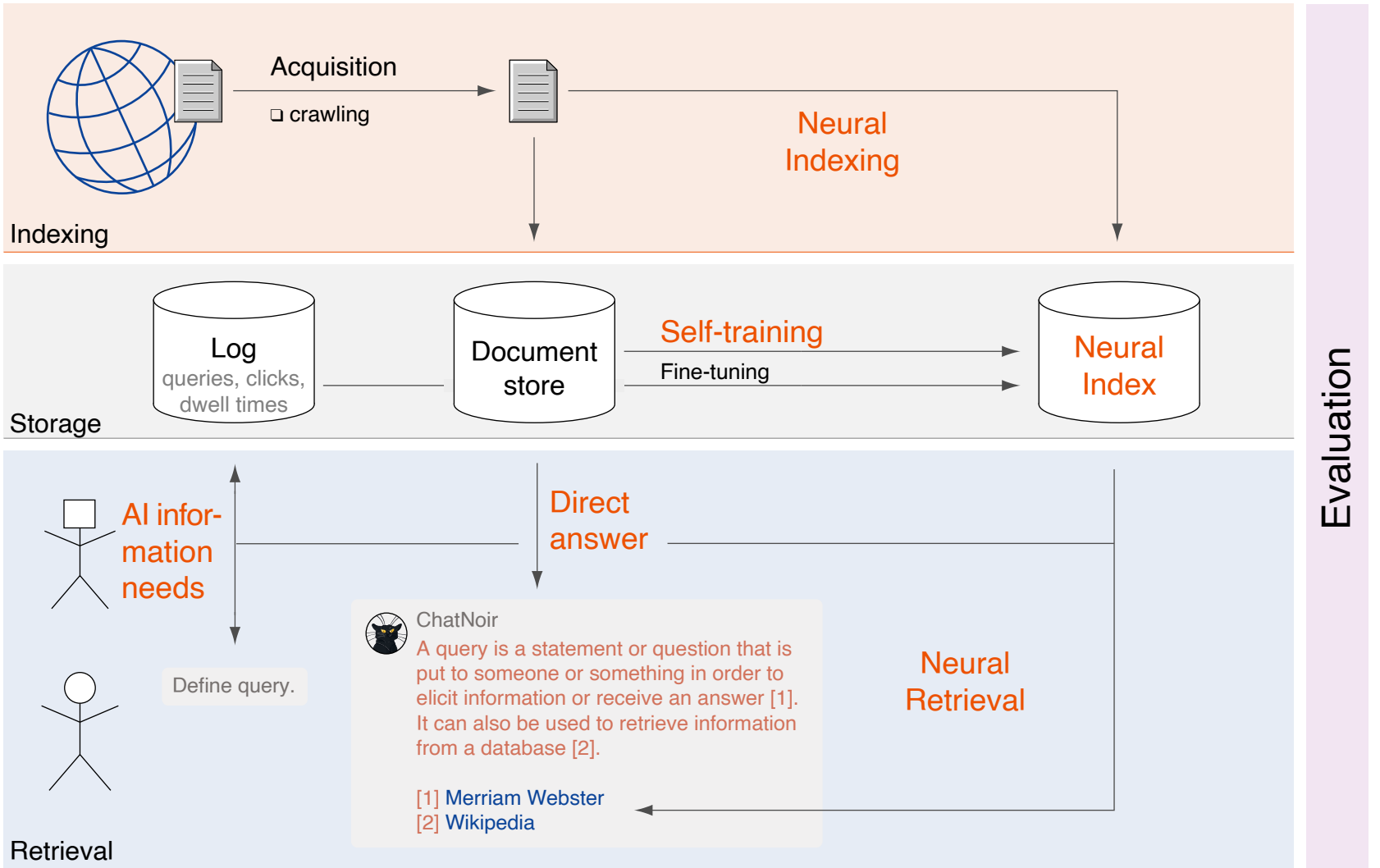
Architecture of a Search Engine

Karaoke Version



Architecture of a Search Engine

Karaoke Version



Document Processing and Indexing

Extract text ✓ → Remove stop words

Die Bibliothek von Alexandria war die bedeutendste antike Bibliothek. Sie entstand Anfang des 3. Jahrhunderts v. Chr. in der kurz zuvor in Ägypten gegründeten makedonisch-griechischen Stadt Alexandria. Der Zeitpunkt des Endes der Bibliothek ist ungeklärt. Die Annahmen reichen von 48 v. Chr. bis ins 7. Jahrhundert. Oft geäußert wird die Ansicht, dass sie im 3. Jahrhundert der Zerstörung des gesamten Palastviertels von Alexandria zum Opfer fiel. Bisher sind keine Überreste der Bibliothek gefunden worden, jedoch bieten die Texte antiker Autoren einige Informationen.

Die Bibliothek verfügte über einen für die damaligen Verhältnisse enormen, aber heute unbekannten Bestand an Schriftrollen. Es handelte sich dabei sowohl um literarische Schriften als auch große Mengen an wissenschaftlicher Literatur aus den verschiedensten Fachgebieten. Es ist anzunehmen, dass bereits bald nach der Gründung ein großer Bestand vorhanden war, der danach über Generationen weiter wuchs. Eine kleinere Tochterbibliothek im Serapeion von Alexandria befand sich im Südwesten der Stadt in einem von den einheimischen Ägyptern bewohnten Stadtteil.

Document Processing and Indexing

Remove punctuation and „unimportant“ words

Die Bibliothek von Alexandria war die bedeutendste antike Bibliothek. Sie entstand Anfang des 3. Jahrhunderts v. Chr. in der kurz zuvor in Ägypten gegründeten makedonisch-griechischen Stadt Alexandria. Der Zeitpunkt des Endes der Bibliothek ist ungeklärt. Die Annahmen reichen von 48 v. Chr. bis ins 7. Jahrhundert. Oft geäußert wird die Ansicht, dass sie im 3. Jahrhundert der Zerstörung des gesamten Palastviertels von Alexandria zum Opfer fiel. Bisher sind keine Überreste der Bibliothek gefunden worden, jedoch bieten die Texte antiker Autoren einige Informationen.

Die Bibliothek verfügte über einen für die damaligen Verhältnisse enormen, aber heute unbekannten Bestand an Schriftrollen. Es handelte sich dabei sowohl um literarische Schriften als auch große Mengen an wissenschaftlicher Literatur aus den verschiedensten Fachgebieten. Es ist anzunehmen, dass bereits bald nach der Gründung ein großer Bestand vorhanden war, der danach über Generationen weiter wuchs. Eine kleinere Tochterbibliothek im Serapeion von Alexandria befand sich im Südwesten der Stadt in einem von den einheimischen Ägyptern bewohnten Stadtteil.

Document Processing and Indexing

Remove punctuation and „unimportant“ words ✓ → Normalize

Bibliothek Alexandria bedeutendste antike Bibliothek entstand Anfang 3 Jahrhunderts Chr kurz zuvor Ägypten gegründeten makedonisch griechischen Stadt Alexandria Zeitpunkt Endes Bibliothek ungeklärt Annahmen reichen 48 Chr 7 Jahrhundert geäußert Ansicht 3 Jahrhundert Zerstörung gesamten Palastviertels Alexandria Opfer fiel Bisher keine Überreste Bibliothek gefunden worden jedoch bieten Texte antiker Autoren einige Informationen

Bibliothek verfügte damaligen Verhältnisse enormen heute unbekannten Bestand Schriftrollen handelte literarische Schriften große Mengen wissenschaftlicher Literatur verschiedensten Fachgebieten anzunehmen Gründung großer Bestand vorhanden Generationen wuchs kleinere Tochterbibliothek Serapeion Alexandria befand Südwesten Stadt einheimischen Ägyptern bewohnten Stadtteil

Document Processing and Indexing

Normalize remaining words

bibliothek alexandria bedeutend antik bibliothek entstehen anfang 3 jahrhundert chr kurz zuvor ägypten gründen makedonisch griechisch stadt alexandria zeitpunkt ende bibliothek kläre n annahme reichen 48 chr 7 jahrhundert äußern ansicht 3 jahrhundert zerstörung gesamt palastviertel alexandria opfer fallen bisher kein überrest bibliothek finden werden jedoch bieten text antik autor einige information

bibliothek verfügen damalig verhältnis enorm heute unbekannt bestand schriftrolle handeln literarisch schrift groß menge wissenschaftlich literatur verschieden fachgebiet an nehmen gründung groß bestand vorhanden generation wachsen klein tochterbibliothek serapeion alexandria befinden südwesten stadt einheimisch ägypter wohnen stadtteil

Document Processing and Indexing

Normalize remaining words ✓ → Count and order

bibliothek alexandria bedeutend antik bibliothek
entstehen anfang 3 jahrhundert chr kurz zuvor
ägypten gründen makedonisch griechisch stadt alexandria
zeitpunkt ende bibliothek klären annahme reichen
48 chr 7 jahrhundert äußern ansicht
3 jahrhundert zerstörung gesamt palastviertel
alexandria opfer fallen bisher kein überrest bibliothek
finden werden jedoch bieten text antik autor einige
information

bibliothek verfügen damalig verhältnis
enorm heute unbekannt bestand schriftrolle handeln
literarisch schrift groß menge
wissenschaftlich literatur verschieden fachgebiet
annehmen gründung groß bestand
vorhanden generation wachsen klein
tochterbibliothek serapeion alexandria befinden südwesten
stadt einheimisch ägypter wohnen stadtteil

Document Processing and Indexing

Count and order lexicographically ✓ → Prune and merge

3:	2	bisher:	1	heute:	1	stadtteil:	1
48:	1	chr:	2	information:	1	südwesten:	1
7:	1	damalig:	1	jahrhundert:	3	text:	1
ägypten:	1	einheimisch:	1	jedoch:	1	tochterbibliothek:	1
ägypter:	1	einige:	1	kein:	1	überrest:	1
äußern:	1	ende:	1	klein:	1	unbekannt:	1
alexandria:	4	enorm:	1	klären:	1	verfügen:	1
anfang:	1	entstehen:	1	literarisch:	1	verhältnis:	1
annahme:	1	fachgebiet:	1	literatur:	1	verschieden:	1
annehmen:	1	fallen:	1	makedonisch:	1	vorhanden:	1
ansicht:	1	finden:	1	menge:	1	wachsen:	1
antik:	2	generation:	1	opfer:	1	werden:	1
autor:	1	gesamt:	1	palastviertel:	1	wissenschaftlich:	1
bedeutend:	1	griechisch:	1	reichen:	1	wohnen:	2
befinden:	1	groß:	2	schrift:	1	zeitpunkt:	1
bestand:	2	gründen:	1	schriftrolle:	1	zerstörung:	1
bibliothek:	5	gründung:	1	serapeion:	1	zuvor:	1
bieten:	1	handeln:	1	stadt:	1		

Document Processing and Indexing

Prune single occurrences and merge words

3:	2	bisher:	1	heute:	1	[illegible]:	1
48:	1	chr:	2	information:	1	südwesten:	1
7:	1	damalig:	1	jahrhundert:	3	text:	1
ägypt[illegible]:	2	einheimisch:	1	jedoch:	1	[illegible]bibliothek:	1
[illegible]:	1	einige:	1	kein:	1	überrest:	1
äußern:	1	ende:	1	klein:	1	unbekannt:	1
alexandria:	4	enorm:	1	klären:	1	verfügen:	1
anfang:	1	entstehen:	1	litera[illegible]:	2	verhältnis:	1
annahm[illegible]:	2	fachgebiet:	1	[illegible]:	1	verschieden:	1
[illegible]:	1	fallen:	1	makedonisch:	1	vorhanden:	1
ansicht:	1	finden:	1	menge:	1	wachsen:	1
antik:	2	generation:	1	opfer:	1	werden:	1?
autor:	1	gesamt:	1	palastviertel:	1	wissenschaftlich:	1
bedeutend:	1	griechisch:	1	reichen:	1	wohnen:	2
befinden:	1	groß:	2	schrift:	2	zeitpunkt:	1
bestand:	2	gründ[illegible]:	2	[illegible]:	1	zerstörung:	1
bibliothek:	6	[illegible]:	1	serapeion:	1	zuvor:	1
bieten:	1	handeln:	1	stadt:	2		

Document Processing and Indexing

Prune single occurrences and merge words ✓

```
3:           2
ägypt:       2
alexandria:  4
annahm:      2
antik:       2
bestand:     2
bibliothek:  6
chr:         2
groß:        2
gründ:       2
jahrhundert: 3
litera:      2
schrift:     2
stadt:       2
wohnen:      2
```

Document Processing and Indexing

Prune single occurrences and merge words ✓

```
3:            2
ägypt:        2
alexandria:   4
annahm:       2
antik:        2
bestand:      2
bibliothek:   6
chr:          2
groß:         2
gründ:        2
jahrhundert:  3
litera:       2
schrift:      2
stadt:        2
wohnen:       2
```

For real use cases

- Repeat process for each document

Then. . .

- Save output in fast searchable structure
- Inverted index $\rightarrow$ analogous to index in books

Next Steps

Assignment

- Exercise sheet on the website [temir.org]

- Familiarize (again) with:
  - TREC campaign
  - Evaluation following the Cranfield paradigm