Copyright aspects of training and using AI

Training generative AI: Challenges of mining copyrighted works

Caren Kresse
Open Source Automation Development Lab (OSADL) eG

Training generative AI

... requires a large amount of **training data** of the type of material that is to be generated (e.g. text, images).

- Training data are often **collected*** from publicly accessible Internet sources.
- Data must be **prepared*** for use (e.g. scaled, labeled).
- Data are then **used*** in the training and testing process of AI models.

DALL-E 3 via ChatGPT: Please create a picture of a 19th century photographer in the wild west.

***Copyright relevant acts (reproduction, adaption, ...) might be performed in all of these steps.**

Copyright law in Europe

- Copyright law **protects** works of art and literature.
- It grants copyright holders **exclusive rights** (among others)
  - to reproduce
  - to distribute
  - to publicize adaptions of

  their protected works.
- It also allows certain **limitations** / **exceptions** where works can be used without the copyright holders' consent.

Copyright limitations / exceptions

… describe certain cases where works can be used without the copyright holders' consent, such as

- reproductions for private use,
- temporary reproductions as part of a technological process,
- reproductions for the purpose of **text and data mining**.

Text and data mining (TDM)

- TDM is used to **analyze** (large amounts of) data to **extract information**, e.g. on patterns and correlations.
- The European Directive on Copyright in the Digital Single Market (CDSM) introduced a **copyright exception for reproductions of works for TDM** in 2019.
  - In Germany: § 44b UrhG (2021).
- Reasoning: Copyright law only protects the personal intellectual creation and not the information contained within that TDM is aimed at.

Does TDM include AI training? (1)

Certainly not

- AI training does not use only information but also the embodied form.

Yes, of course

- AI training aims to generalize patterns based on probabilities.

Does TDM include AI training? (1)

Certainly not

- AI training does not use only information but also the embodied form.
- It is possible to reproduce identical elements of training data.

Yes, of course

- AI training aims to generalize patterns based on probabilities.
- A fully trained AI does not aim to reproduce training data, but generate something new.

Does TDM include AI training? (2)

Certainly not

- Exceptions to copyright shall not conflict with normal exploitation of a work. However, products of generative AI are in competition with the works used for training.

Yes, of course

- The possibility to opt-out of the TDM exception provides a fair balance of interests between human works and AI creations.

Does TDM include AI training? (3)

Certainly not

- When the TDM exception was included in the CDSM Directive (2019), AI training was not mentioned explicitly, nor was it discussed in detail during the legislative process.

Yes, of course

- The law is worded technology-neutrally and recital 2 stipulates the aim to "stimulate innovation [...] and production of new content".
- There are several documents predating the CDSM directive that mention TDM in connection with AI training.
- The AI Act (2024) references the TDM exception explicitly for training of generative AI.

Does TDM include AI training? (4)

Certainly not

- The court did not decide on the applicability of the TDM exception for training of AI, but only for data collection and analysis.

Yes, of course

- A German court (LG Hamburg) decided that the TDM exception is applicable for analysis of images with AI to provide the information in a database for AI training.
- The court did however comment that the reasons given against are not convincing.

Does TDM include AI training? (result)

Probably, yes

- Stronger arguments (and more literature) point towards the applicability of the TDM exception to AI training.

Does TDM include AI training? (result)

Probably, yes

- Stronger arguments (and more literature) point towards the applicability of the TDM exception to AI training.

So, let's take a look at it.

Further requirements by the TDM exception (1)

- The exception applies **only for reproduction and extractions**, not for distribution or publication.
- The works used for TDM must be **lawfully accessible** (note: not lawfully published).
- Reproductions may only be **retained for as long as is necessary** (???) for the purposes of text and data mining.
- **Modifications** that are merely due to technical requirements (such as scaling) are also allowed.

Further requirements by the TDM exception (2)

- The exception does no apply when the right holder has **expressly reserved** the right for TDM in a **machine-readable** way.

Machine-readable reservation

- Data for TDM is generally collected by Internet "crawlers" that **automatically find and download** relevant content.
- A reservation should be processable by such automatic crawlers and to that end be **machine-readable**.
- Technology is now so far advanced that almost anything can be made "machine-readable", but:
  - **at different costs** and
  - with varying **levels of ease** for a copyright holder.

Cost for machine-readability

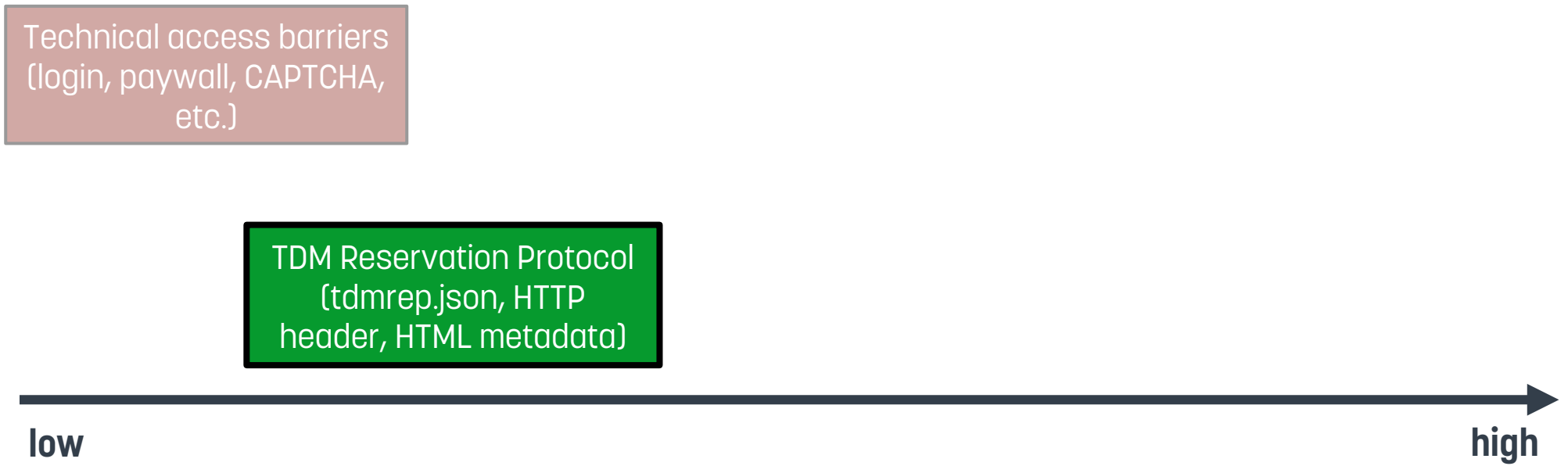
Technical access barriers (login, paywall, CAPTCHA, etc.)

Increasing cost for machine-readability

Technical access barriers

- Content behind a **login mechanism** cannot be accessed by crawlers.
- This is a permitted and effective expression of a TDM reservation **against unprivileged users**.
- However, if a user is privileged to access content, they may use it for TDM if there is no other form of reservation.

Cost for machine-readability



Increasing cost for machine-readability

TDM Reservation Protocol

- Protocol in development by a W3C community group:https://www.w3.org/community/tdmrep/
- Metadata "location", "tdm-reservation" (*true,1* → *reservation* or *false,0* → *no reservation*) and "tdm-policy" (details, e.g. on licensing for use) are sent via
  - HTTP header on request from a client for an entire website
  - HTML tag for a single (HTML conformal) file
  - JSON file (*tdmrep.json*) in the root directory of a website

    The following examples are taken from https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/

HTTP header

```
HTTP/1.1 200 OK
Date: Wed, 26 Mar 2025 15:07:48 GMT
Content-type: text/html
tdm-reservation: 1
tdm-policy: https://provider.com/tdm-policy.json
```

TDM is allowed under the conditions set out in the tdm-policy.json

HTML metadata

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <meta name="tdm-reservation" content="1">
    <meta name="tdm-policy" content="https://provider.com/tdm-policy.json">
    <title>Document title</title>
  </head>
  <body>
    ...
    <!-- body content -->
    ...
  </body>
</html>
```

TDM is allowed under the conditions set out in the tdm-policy.json

tdmrep.json example

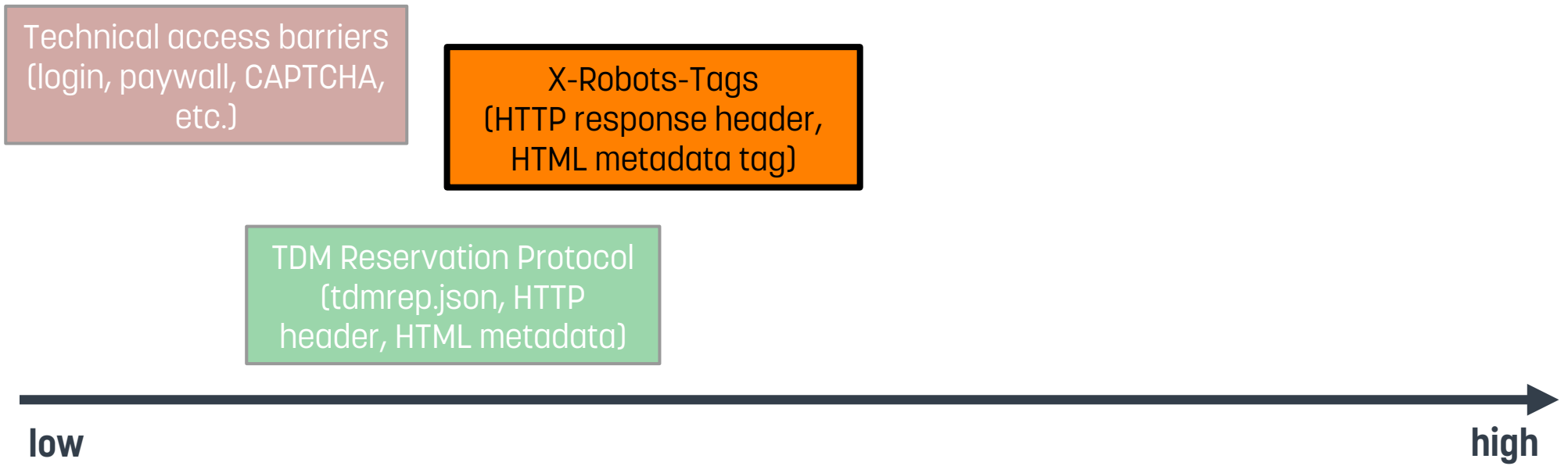
```
[
  {
  "location": "/directory-a/",
  "tdm-reservation": 1
  },
  {
  "location": "/directory-b/",
  "tdm-reservation": 1,
  "tdm-policy": "https://provider.com/tdm-policy.json"
  },
  {
  "location": "/images/*.jpg",
  "tdm-reservation": 0
  }
]
```

No TDM in *directory-a*

TDM in *directory-b* under the conditions set out in the tdm-policy.json

TDM allowed for jpg files in the directory *images*

Cost for machine-readability



Increasing cost for machine-readability

X-Robots-Tags

- Metadata with values like "noai" can be given via HTTP response or in an HTML file.
- Provides the possibility to express TDM reservations for individual content, but does not allow for detailed conditions of use.
- There are no standardized values for this tag, yet.

X-Robots-Tag examples

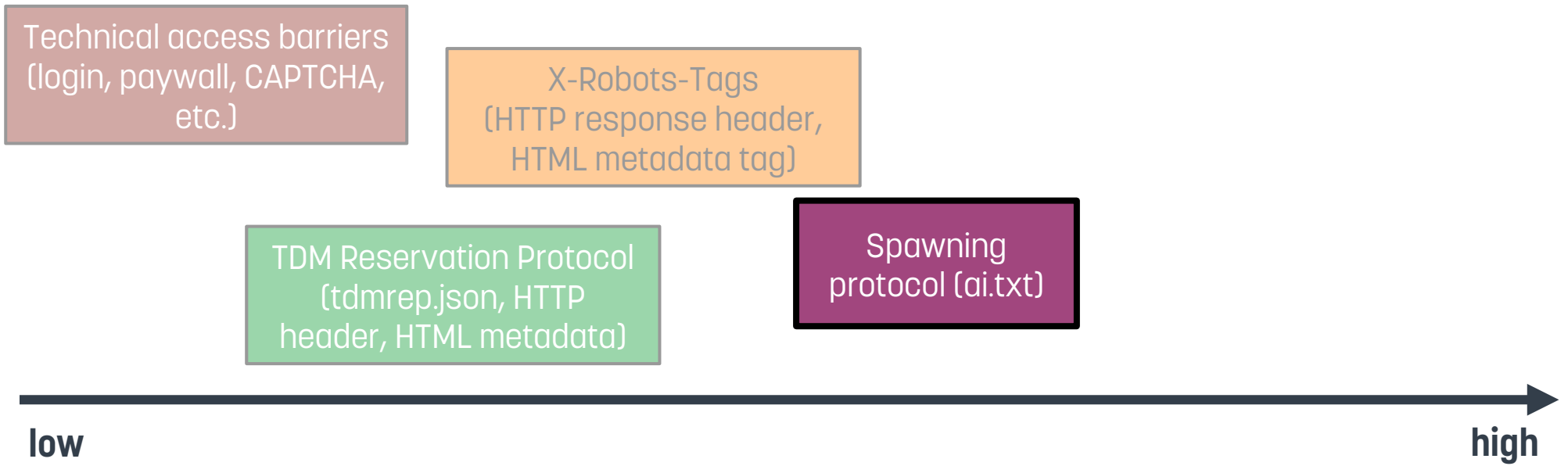
HTTP header

```
HTTP/1.1 200 OK
Date: Wed, 26 Mar 2025 15:07:48 GMT
(…)
X-Robots-Tag: noai
(…)
```

HTML metadata

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <meta name="X-Robots-Tag" content="noimageai">
    <title>Document title</title>
  </head>
  <body>
    ...
    <!-- body content -->
    ...
  </body>
</html>
```

Cost for machine-readability



Increasing cost for machine-readability

Spawning protocol

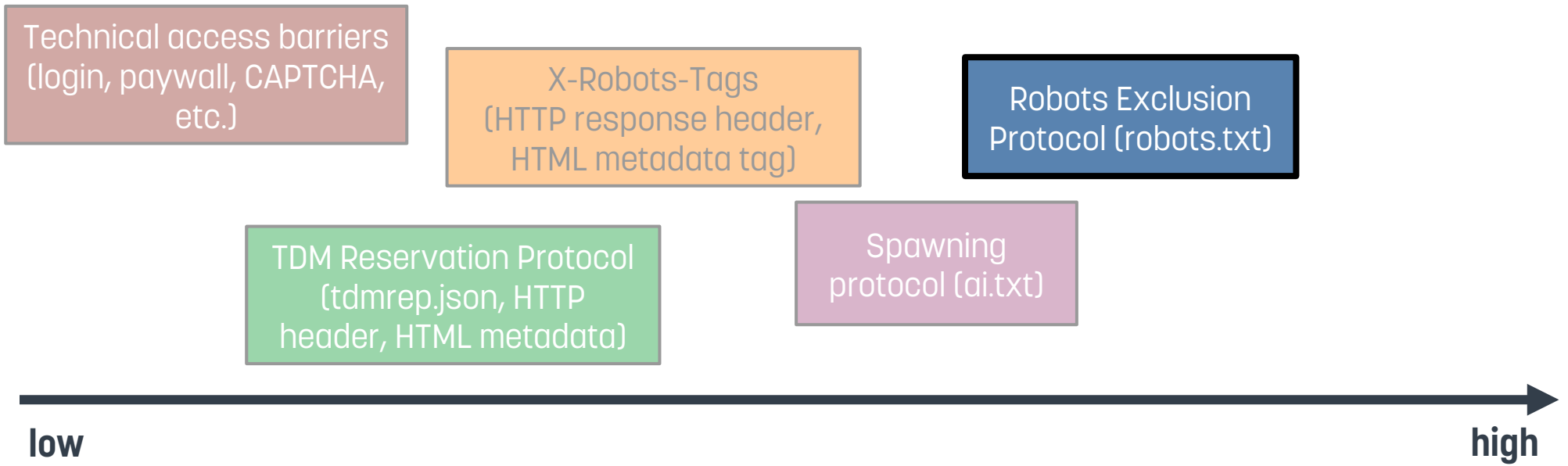
- Created by a US American company in collaboration with AI and content providers.
- TDM reservations are expressed in a *ai.txt* file placed in a website's root directory.
- Individual *ai.txt* can be created at https://spawning.ai/ai-txt#create

ai.txt

```
# Spawning AI
# Prevent datasets from using the
# following file types

User-Agent: *
Disallow: *.aac
Disallow: *.aiff
Disallow: *.amr
Disallow: *.flac
Disallow: *.m4a
Disallow: *.mp3
Disallow: *.oga
Disallow: *.opus
Disallow: *.wav
Disallow: *.wma
Disallow: *.mp4
Disallow: *.webm
Disallow: *.ogg
Disallow: *.avi
Disallow: *.mov
Disallow: *.wmv
Disallow: *.flv
Disallow: *.mkv
Disallow: *.py
Disallow: *.js
Disallow: *.java
Disallow: *.c
Disallow: *.cpp
Disallow: *.cs
Disallow: *.h
Disallow: *.css
Disallow: *.php
Disallow: *.swift
Disallow: *.go
Disallow: *.rb
Disallow: *.pl
Disallow: *.sh
Disallow: *.sql
Allow: *.txt
Allow: *.pdf
Allow: *.doc
Allow: *.docx
Allow: *.odt
Allow: *.rtf
Allow: *.tex
Allow: *.wks
Allow: *.wpd
Allow: *.wps
Allow: *.html
Allow: *.bmp
Allow: *.gif
Allow: *.ico
Allow: *.jpeg
Allow: *.jpg
Allow: *.png
Allow: *.svg
Allow: *.tif
Allow: *.tiff
Allow: *.webp
Allow: /
```

Disallows TDM for all users for audio, video and code, but allows TDM for text and images on the entire website

Cost for machine-readability



Increasing cost for machine-readability

Robots Exclusion Protocol (REP)

- Well-established *de-facto* communication protocol for search engines and since 2022 standardized by IETF.
- A *robots.txt* file is placed in a website's root directory.
- Can contain arbitrary text (also in natural language).
- Specific locations can be named, but no fine-grained reservation or conditions possible.
- Does not differentiate between search engine indexing and TDM, because to exclude a specific crawler, its name must be known.

robots.txt example

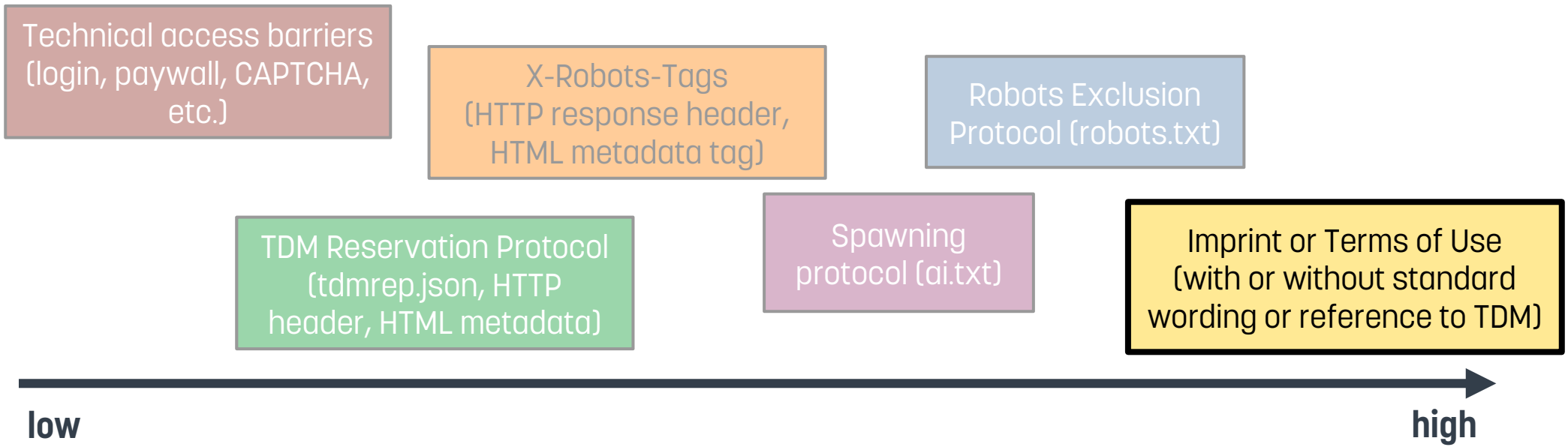
```
# Disallow User-Agents
User-agent: CCBot
Disallow: /

# Disallow public gptbot
User-agent: GPTBot
Disallow: /textfiles/

# Disallow meta externalagent
User-agent: meta-externalagent
Allow: /images/
```

No TDM for CCBot (Common Crawl) on entire website

No TDM in directory *textfiles* for GPTBOT

TDM allowed in directory *images* for Meta bot

Cost for machine-readability



Increasing cost for machine-readability

Imprint or Terms of Use

- Natural language, i.e. many different wordings are possible.
- Can be placed at various locations on a website.
- Would require the use of AI to automatically identify TDM reservations or conditions.

Imprint or Terms of Use: examples

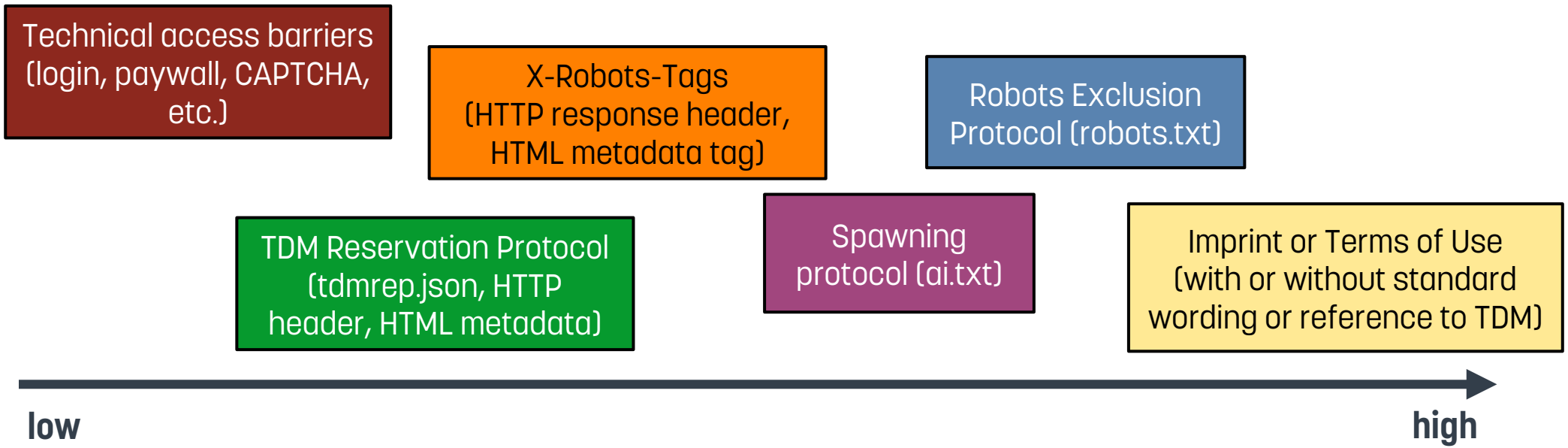
Text und Data Mining gemäß § 44b UrhG | Text and Data Mining according to § 44b UrhG

Der Verlag behält sich das Recht zu Vervielfältigungen für das Text und Data Mining gemäß § 44b UrhG vor.
The publisher reserves the right to reproduce for text and data mining according to § 44b UrhG.

RESTRICTIONS

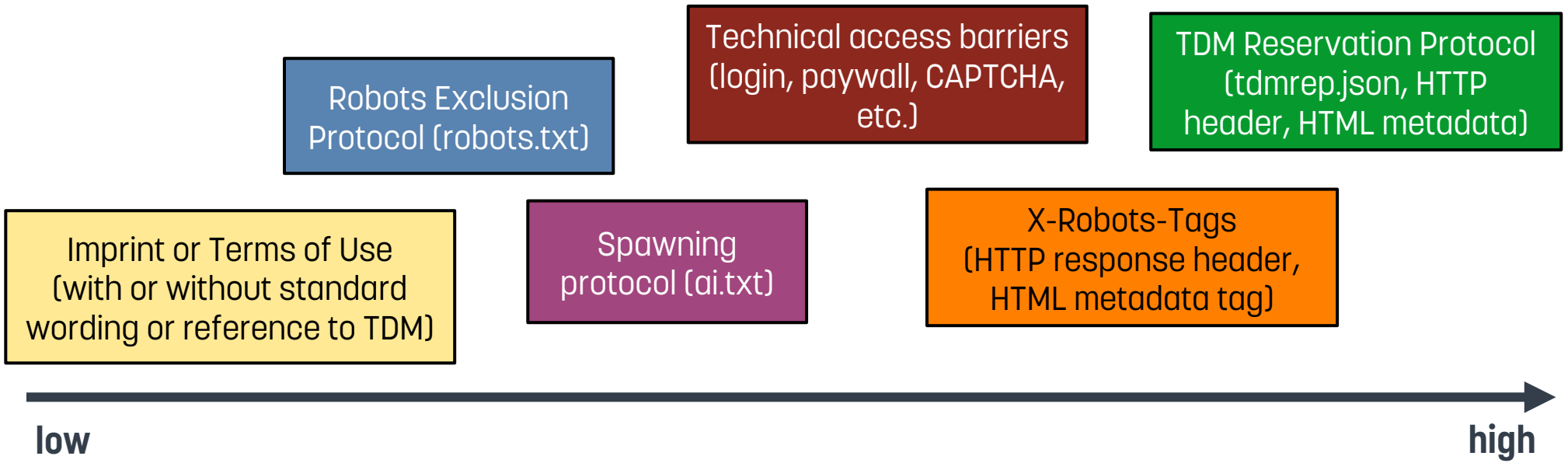
YOU MAY NOT: [...] Use automated programs, applets, bots or the like to access the [...] website or any content thereon for any purpose, including, by way of example only, downloading content, indexing, scraping or caching any content on the website.

Cost for machine-readability



Increasing cost for machine-readability

Ease of implementation for copyright holders



Increasing difficulty of implementation for copyright holders

Declaring TDM reservation: Overview

	Standardized expressions	Individual content	Express conditions	Easy to use for a not technically adept right holder	easily machine-readable
Technical access barriers	n.a.	++	+	+	+
TDM Reservation Protocol	++	++	++	– –	+
X-Robots-Tag	not yet	+	– –	– –	+
Spawning protocol	++	+	– –	+	+
Robots Exclusion Protocol	0	0	– –	–	++
Imprint or Terms of Use	– –	+	++	++	– –

Current developments

- Ongoing court cases (small selection):
  - Robert Kneschke v. LAION (LG Hamburg, appealed)
  - GEMA v. OpenAI & GEMA v. Suno Inc. (both LG Munich)
  - Getty Images v. Stability AI (High Court of Justice of England & Wales)
  - NYT v. Microsoft and OpenAI (U.S. District Court for the Southern District of New York)
- Licensing:
  - In December 2023, Axel Springer and OpenAI agree on terms for use of content.

Outside the EU

- Although copyright is harmonized internationally to a fair extend, it remains a **territorial right**.
- The country where the act relevant to copyright law takes place determines the applicable law.
- For data collection and AI training this means the country **where the downloaded data are stored**.

Outside the EU (2)

- United States of America
  - Training of generative AI might be "fair use" considering:
    - purpose of use and whether it is commercial or nonprofit,
    - nature and amount of the copyrighted work,
    - effect on the market for or value of the copyrighted work.
  - "Fair use" cannot be reserved.
- Japan
  - Generous legal permissions, but some restrictions for generative AI.
- China
  - Interim provisions (2023) require providers of generative AI to ensure that data from legal sources is used.

What does this mean for my practice ...

...as a copyright holder?

- Use a (or better: several) standard protocol to reserve TDM rights expressly. Ideally attached to individual works.
- Adherence to any protocol is "voluntary" for crawlers.
- To absolutely ensure that data is not used for AI training, impose technological access barriers.

What does this mean for my practice ...

...as a machine learning engineer?

- Consider all "easily" machine-readable reservation protocols.
- For large data sources, also consider natural language reservations or licenses.
- Document data sources.

Sources (selection)

Regulations:

- Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC

Literature:

- Käde, Lisa: Wann ist ein Nutzungsvorbehalt gegen über Text- und Data Mining maschinenlesbar und dem Rechtsinhaber zurechenbar? CR 2024, 598–603
- Dornis, Tim W./Stober, Sebastian: Urheberrecht und Training generativer KI-Modelle, Technologische und juristische Grundlagen, Nomos Verlagsgesellschaft 2024
- Käde, Lisa: Training generativer KI-Modelle ist (auch) Text- und Data-Mining, KIR 2024, 126–169
- Hamann, Hanjo: Nutzungsvorbehalte für KI-Training in der Rechtsgeschäftslehre der Maschinenkommunikation, ZGE 2024, 113–168
- LG Hamburg mit Anm. Malte Grützmacher: LG Hamburg: Urheberrechtsschutz für KI-Input-Material, CR 2024, 751–758
- Leistner, Matthias: TDM und KI-Training in der Europäischen Union, GRUR 2024, 1665–1675
- de la Durantaye, Katharina: Garbage In, Garbage Out. Regulating Generative AI Through Copyright Law, ZUM 2023, 645–660