# AI Training for Scientific Research: A Copyright Perspective

Leibniz ScienceCampus
Digital Transformation of Research


**Pascal Sierek**

Moderation: Lea Sophie Singson & Diana Dimitrova

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

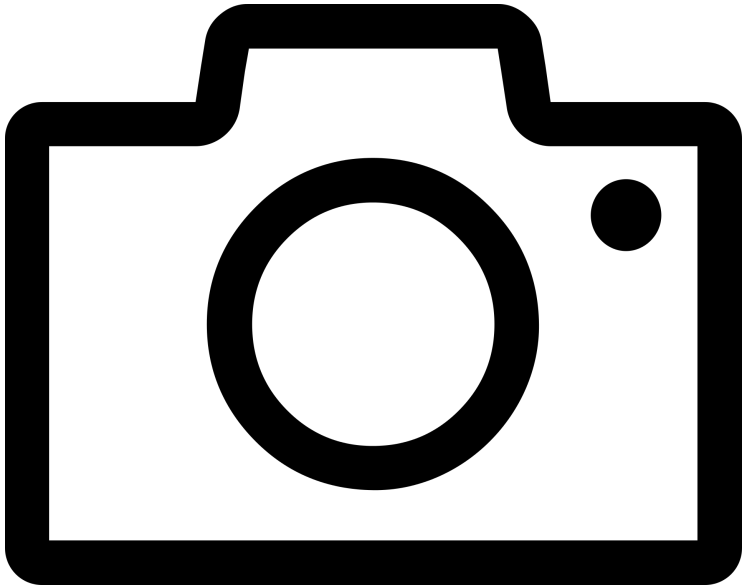
KIT
Karlsruhe Institute of Technology



funded by Leibniz Competition


Interdisciplinary Colloquium on Digitalisation of Research, Zoom, 2nd of April 2026

# Photos and recording



*Pixabay, ste_phania*

YouTube

www.youtube.com/@DiTraRe

Leibniz ScienceCampus
Digital Transformation of Research

Interdisciplinary Colloquium on Digitalisation of Research

2 April | 11:00-12:00

# AI Training for Scientific Research: A Copyright Perspective

Pascal Sierek

Max Planck Institute for Comparative and International Private Law, Hamburg

Register now!

Join us in person
or online urlr.me/!ditrare-colloquium

FIZ Karlsruhe KIT Leibniz Association

MAX PLANCK INSTITUTE
for comparative and international
PRIVATE LAW HAMBURG

# AI TRAINING FOR SCIENTIFIC RESEARCH: A COPYRIGHT PERSPECTIVE

**Dr. Pascal T. Sierek**

**02. April 2026**

MAX PLANCK INSTITUTE
for comparative and international
PRIVATE LAW HAMBURG

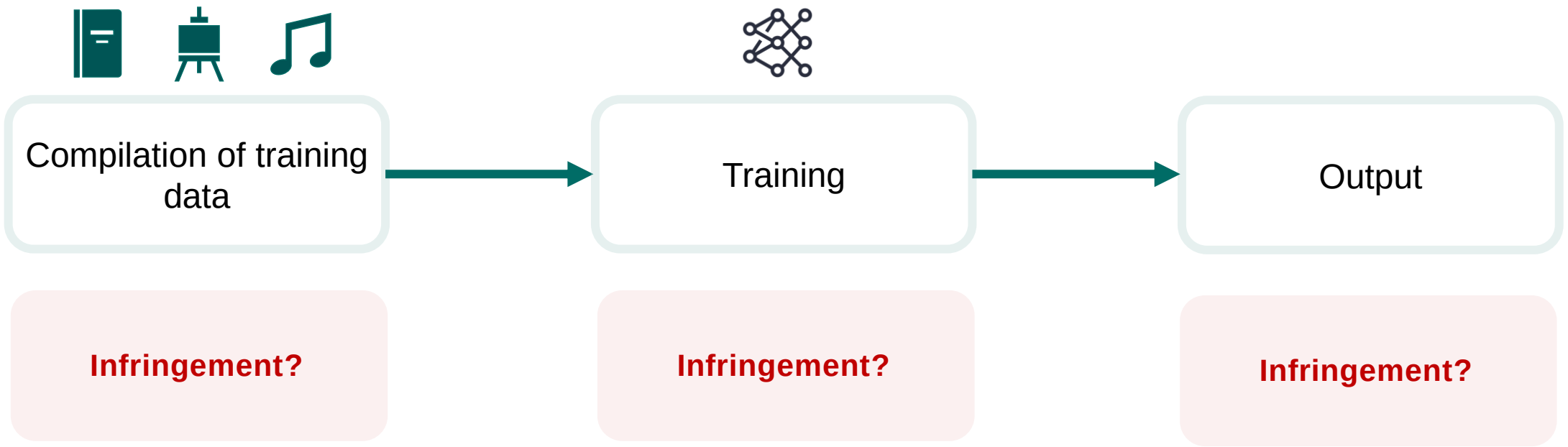
# AI TRAINING FOR SCIENTIFIC RESEARCH: A COPYRIGHT PERSPECTIVE

**Dr. Pascal T. Sierek**

**02. April 2026**

# BIG PICTURE

- **Fundamental tension:** "AI as innovation engine" vs. "Artpocalypse"
- **Special case:** AI in scientific research



# WHEN IS INFORMATION PROTECTED BY COPYRIGHT?

## Works (§ 2 UrhG): sufficient creative originality

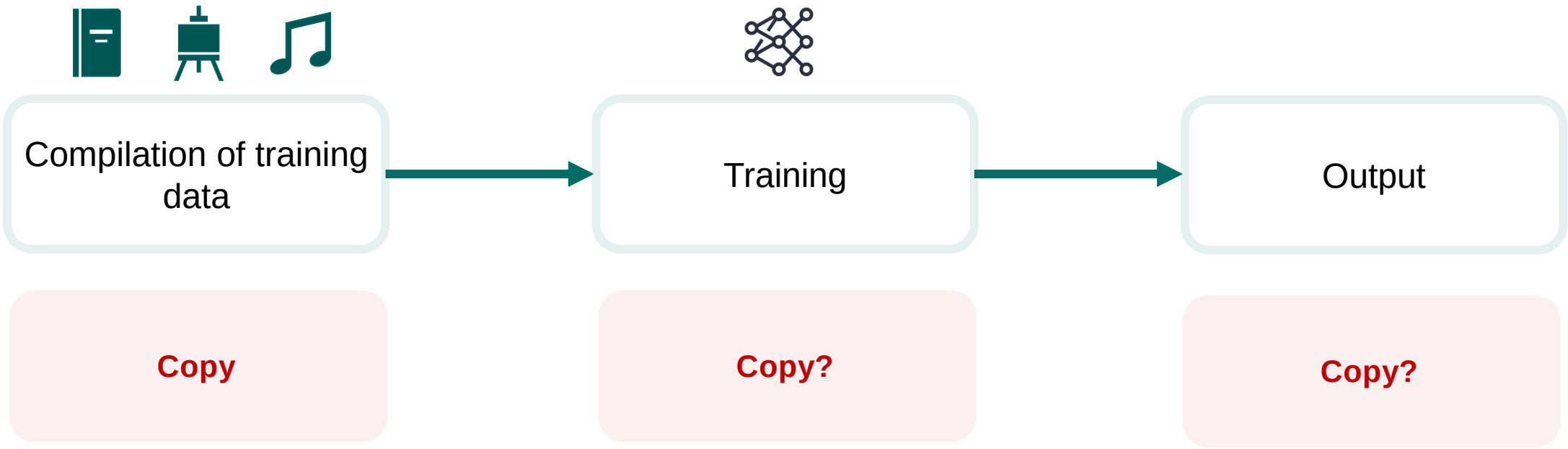
- Photographs, books, texts, videos, music, etc.
- **not:** pure facts, ideas, artistic styles, and scientific findings (however, database protection might be relevant)

## Related rights (even without high creativity)

- Press publications (§§ 87f ff. UrhG)
- Simple photographs (§ 72 UrhG)

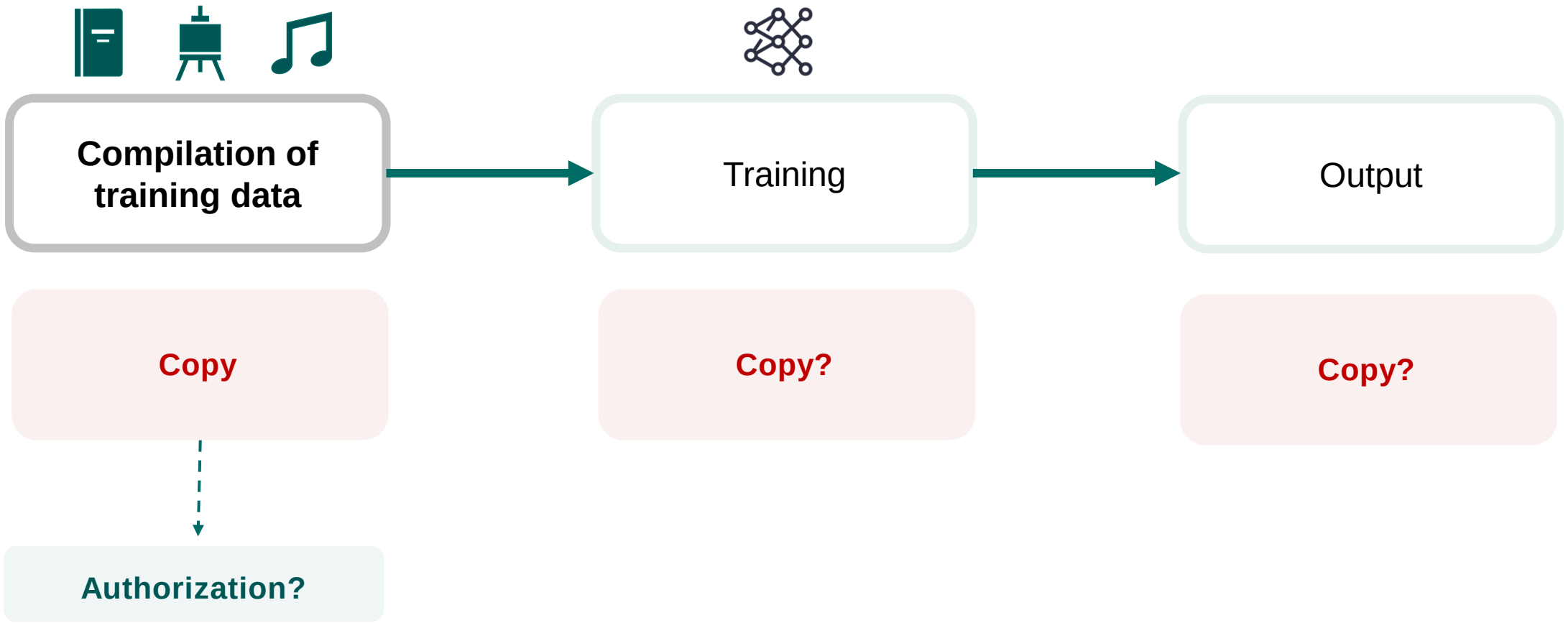
## Databases representing substantial investment (§ 87a UrhG)

# 3 COPYRIGHT-RELEVANT PHASES OF THE AI VALUE CHAIN



# PHASE 1: COMPILATION OF TRAINING DATA

# 3 COPYRIGHT-RELEVANT PHASES OF THE AI VALUE CHAIN



# AUTHORIZATION

## Copying copyrighted material → needs authorization

- **Option A:** Consent from the rightsholder → usually impractical at scale
- **Option B:** Statutory exception (= limitation)

## Two key exceptions:

- **§ 60d UrhG,** scientific TDM
- **§ 44b UrhG,** commercial TDM

# SCIENTIFIC TDM (§ 60D URHG)

## Text and data mining for scientific research purposes

(1) *It is **permitted to make reproductions** to **carry out text and data mining** (section 44b (1) and (2) sentence 1) for **scientific research purposes** in accordance with the following provisions.*

***(2) Research organisations are authorised to make reproductions.** 'Research organisations' means universities, research institutes and other establishments conducting scientific research if they*

1. *pursue non-commercial purposes,*
2. *reinvest all their profits in scientific research or*
3. *act in the public interest based on a state approved mandate.*

## Requirements

- Personal Scope: Research Organisation
- Text and Data Mining (TDM) and lawful access
- Purpose Limitation & Project Specificity
  - Reproduction for TDM
  - TDM for scientific research purposes

# SCIENTIFIC TDM (§ 60D URHG)

## Text and data mining for scientific research purposes

(1) *It is permitted to make reproductions to carry out text and data mining (section 44b (1) and (2) sentence 1) for scientific research purposes in accordance with the following provisions.*

**(2) *Research organisations are authorised to make reproductions.*** *'Research organisations' means universities, research institutes and other establishments conducting scientific research if they*

1. *pursue non-commercial purposes,*
2. *reinvest all their profits in scientific research or*
3. *act in the public interest based on a state approved mandate.*

## Requirements

- **Personal Scope: Research Organisation**
- Text and Data Mining (TDM) and lawful access
- Purpose Limitation & Project Specificity
  - Reproduction for TDM
  - TDM for scientific research purposes

# PERSONAL SCOPE

**Must qualify as a research organization (§ 60d(2) UrhG):** *"'Research organisations' means universities, research institutes and other establishments conducting scientific research if they 1. pursue non-commercial purposes, 2. reinvest all their profits in scientific research or 3. act in the public interest based on a state approved mandate."*

- **Research**: methodical, systematic activity aimed at generating new, verifiable knowledge (natural sciences + humanities)
- **Organization**: university, research institute, or other institution; **plus**:
  - Non-commercial purposes, **or**
  - All profits reinvested in research, **or**
  - Government-recognized public-interest mandate

**Special Case: Public-Private Partnerships (§ 60(2) UrhG)**

- Exception does **not** apply when a private company has:
  - **Decisive influence** over the research organization, **AND**
  - **Preferential access** to the research results

# EXAMPLE: LAION E.V.


- **LAION: private nonprofit association**
- **Case**
  - Downloaded data from a US dataset (5+ billion URL image description pairs)
  - Checked whether each image matched its description → filtered mismatches
  - Published a table of URLs with matching images and descriptions
  - A photographer sued LAION (LG Hamburg → OLG Hamburg)
- **Key question: Is LAION a research organization?**

# EXAMPLE: LAION E.V.

- **LAION: private nonprofit association**
- **Case**
  - Downloaded data from a US dataset (5+ billion URL image description pairs)
  - Checked whether each image matched its description → filtered mismatches
  - Published a table of URLs with matching images and descriptions
  - A photographer sued LAION (LG Hamburg → OLG Hamburg)
- **Key question: Is LAION a research organization?**
  - **Research?** Yes.
    - Creating the dataset = methodical, systematic, verifiable (applied research)
    - Even a single preparatory step in a larger research process is sufficient (contested)
  - **Research organization?** Yes.
    - "Other institution" that itself conducts research
    - Non-commercial purposes (bylaws, results published for free); commercial third parties benefiting from the data ≠ commercial purposes of the organization

# SCIENTIFIC TDM (§ 60D URHG)

## Text and data mining for scientific research purposes

(1) *It is permitted to make reproductions to carry out **text and data mining (section 44b (1) and (2) sentence 1)** for scientific research purposes in accordance with the following provisions.*

(2) *Research organisations are authorised to make reproductions. 'Research organisations' means universities, research institutes and other establishments conducting scientific research if they*

1. *pursue non-commercial purposes,*
2. *reinvest all their profits in scientific research or*
3. *act in the public interest based on a state approved mandate.*

## Requirements

- Personal Scope: Research Organisation
- **Text and Data Mining (TDM) and lawful access**
- Purpose Limitation & Project Specificity
  - Reproduction for TDM
  - TDM for scientific research purposes

# WHAT IS TDM?

## § 44b UrhG:

*(1) Automated analysis of digital works to* ***derive information,*** *in particular patterns, trends, and correlations.*

*(2) It is permitted to reproduce lawfully accessible works in order to carry out text and data mining.*

## Three scenarios: Derive Information?

- LAION: comparing images with descriptions → match or no match
- University: training AI to detect tumors in radiological images (non-generative AI)
- University: training AI on a book corpus → AI can write books (generative AI)

**Bottom line: in all three cases, AI training = TDM**

# LAWFUL ACCESS

## § 44b UrhG:

(1) Automated analysis of digital works to derive information**,** in particular patterns, trends, and correlations“.

(2) It is permitted to reproduce **lawfully accessible** works in order to carry out text and data mining.

## Lawfully Accessible

- Interpreted **broadly:** Anything freely available on the internet
  - Rightsholder consent irrelevant
  - Terms of service prohibiting scraping irrelevant (prevailing opinion, § 60d(1) UrhG)
- **Not** lawfully accessible:
  - Content behind paywalls or technical protection measures
  - Possibly: content on obviously illegal platforms (undecided)

# SCIENTIFIC TDM (§ 60D URHG)

## Text and data mining for scientific research purposes

(1) *It is permitted to make reproductions to carry out **text and data mining (section 44b (1) and (2) sentence 1) for scientific research purposes** in accordance with the following provisions.*

(2) *Research organisations are authorised to make reproductions. 'Research organisations' means universities, research institutes and other establishments conducting scientific research if they*

1. *pursue non-commercial purposes,*
2. *reinvest all their profits in scientific research or*
3. *act in the public interest based on a state approved mandate.*

## Requirements

- Personal Scope: Research Organisation
- Text and Data Mining (TDM) and lawful access
- **Purpose Limitation & Project Specificity**
  - **Reproduction for TDM**
  - **TDM for scientific research purposes**

# PURPOSE LIMITATION & PROJECT SPECIFICITY

- **TDM purpose**: copies may only be used for TDM / training — not for other purposes (e.g., archiving)

- **Research purpose**: TDM must serve scientific research throughout

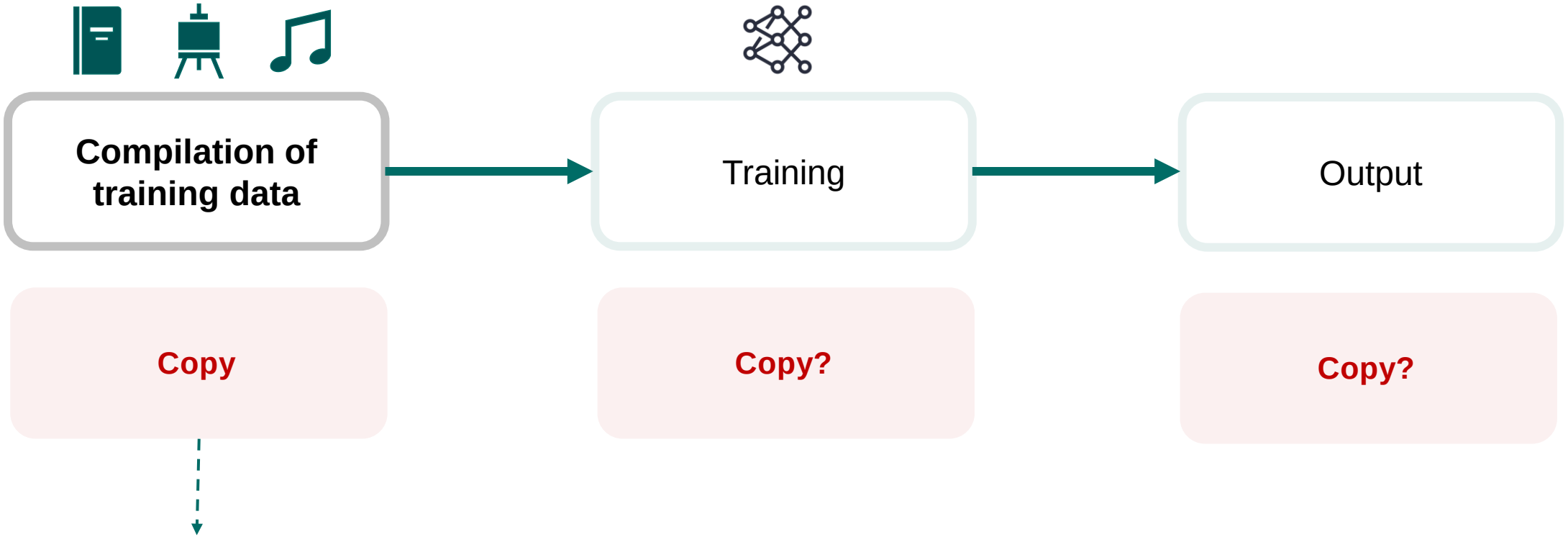
- **Project specificity**
  - § 60d(5) UrhG: *Those authorised … **may retain reproductions** … **for as long as they are needed** for the purposes of the scientific research or the monitoring of the quality … of the scientific findings.*
  - Copies must be deleted if no longer needed for the purpose of the research project
  - **Best practice:** create a deletion plan

# COMMERCIAL TDM (§ 44B URHG)

- Applies when § 60d does not (i.e. not a research organization)
- **Same** requirements: reproduction + TDM + lawful access
- **Additional** requirement: rightsholder must not have opted out in a **machine-readable format**
  - Usually via robots.txt
  - "Machine-readable" = not fully settled

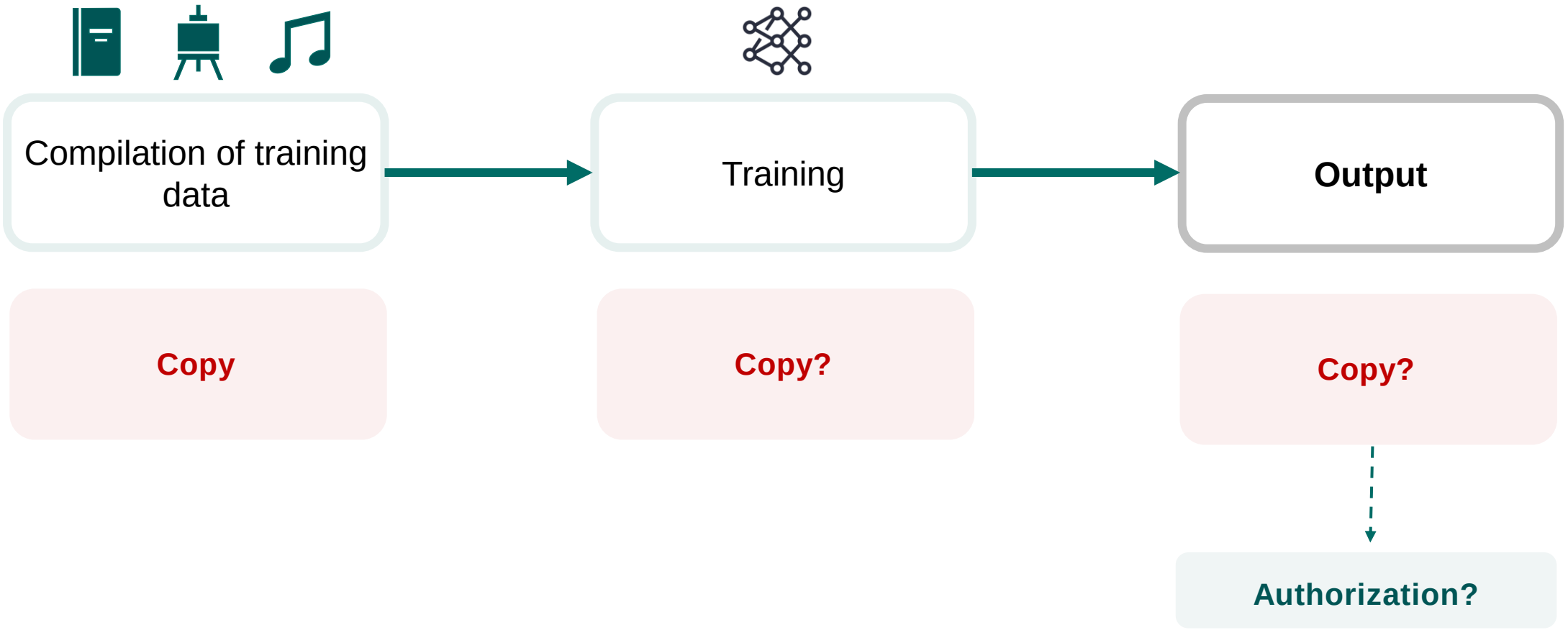
- **Key difference: § 60d is the stronger privilege → no opt-out possible for scientific TDM**

# 3 COPYRIGHT-RELEVANT PHASES OF THE AI VALUE CHAIN



# PHASE 3: OUTPUT

# 3 COPYRIGHT-RELEVANT PHASES OF THE AI VALUE CHAIN



# OUTPUT

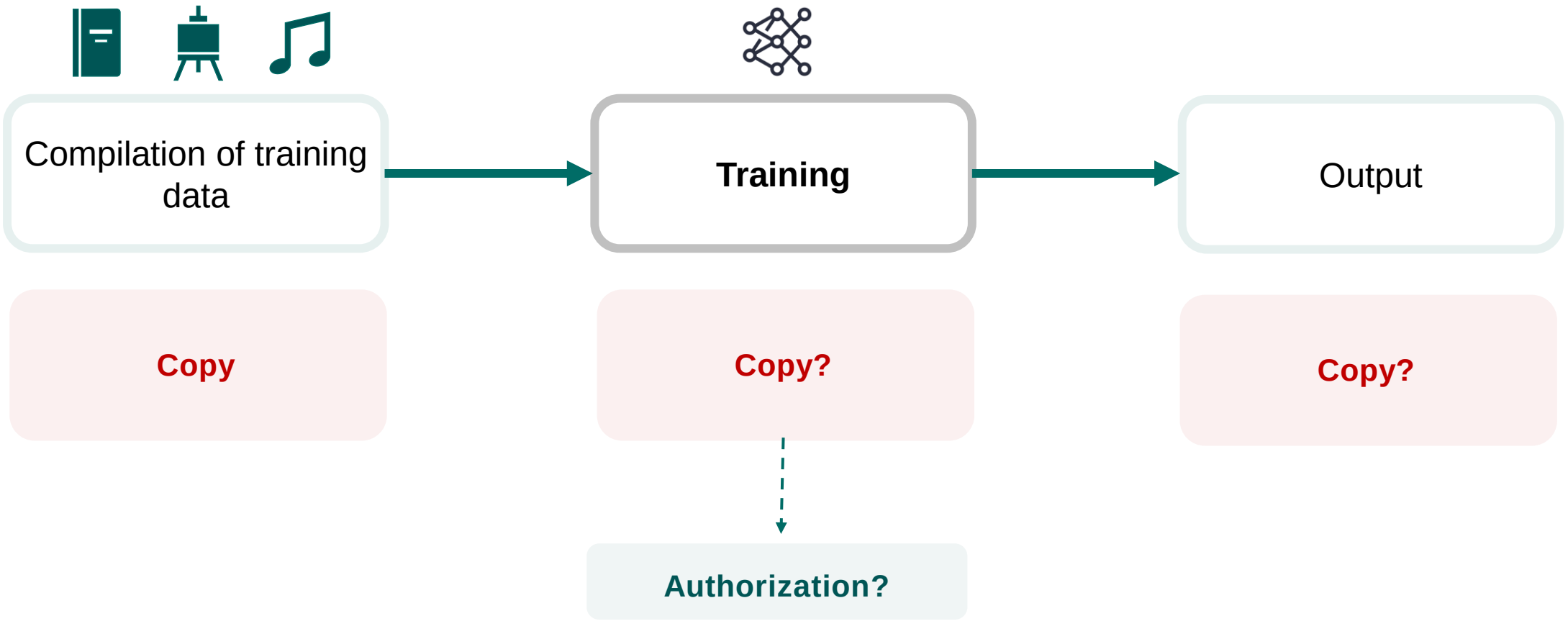
If the output reproduces training data (exactly or slightly modified) → **copyright infringement**

- § 16 UrhG (reproduction): the output is a new copy
- § 23 UrhG (adaptation): modified versions covered if original elements remain recognizable
- § 19a UrhG (making available to the public): when output is delivered via chatbot / online system

**No exception applies: Reproduction is not *for* TDM but *a consequence of* TDM**

# PHASE 2: TRAINING

# 3 COPYRIGHT-RELEVANT PHASES OF THE AI VALUE CHAIN



# TRAINING

**Temporary copies during training:** permitted under § 44a UrhG (transient copies)

**Problem:** are works stored **inside the model?**

- Computer science view: training data is not “stored”, but broken into tokens and neuronal connections are weighted
- Legal view (prevailing): irrelevant **how** information is represented (binary, tokens, etc.)
  - decisive factor is whether work can be made perceptible to human senses
  - **Link to phase 3:** if the model regularly outputs training data → courts infer the works are stored inside the model (memorization)

# CONSEQUENCES

**No exception available: not reproduction *for* TDM but *result of* TDM (§§ 60d, 44b do not apply )**

**Consequences:**

- Injunction → must remove the “reproduction” from the model
  - Machine unlearning
  - Retrain model
- Claim for damages → lost licensing opportunity

# CONCLUDING REMARKS

# SUMMARY

| Phase | General Rule | Risk |
|---|---|---|
| 1. Compilation of Training Data | Reproductions are permissible as TDM (§§ 60d, 44b UrhG) | If commercial TDM, right holders‘ opt-out |
| 2. Training | Generally Permissible | Memorization → inferred from output |
| 3. Output | Generally permissible | Reproducing copyrighted content |

# OUTLOOK

## Pending Cases

- **LAION** → German Federal Court of Justice (BGH)
- **GEMA v. OpenAI** → appeal at OLG München
- **Like Company v. Google Ireland** → CJEU preliminary reference

## EU legislature becoming active

# MANY THANKS FOR YOUR ATTENTION!

Please contact me if you have any questions.

**Dr. Pascal T. Sierek**

Max Planck Institute for
Comparative and International Private Law

Mittelweg 187, 20148 Hamburg
Phone: (+49 40) 41900 – 374
E-Mail: sierek@mpipriv.de
Internet: www.mpipriv.de

# Discussion Session

# Next Colloquium

07.05.26 | 11:00-12:00

## From Knowledge Graphs to AI Assistants: Transforming Scholarly Communication with ORKG and TIB AIssistant

Sören Auer
TIB - Leibniz Information Centre for Science and Technology and University Library, Leibniz University Hannover

### Outline

- The vision and architecture of ORKG for structuring and comparing research contributions.
- ORKG ASK as an interface for querying scholarly knowledge graphs using AI.
- The TIB AIssistant as a companion for discovery, summarization, and knowledge synthesis.
- Opportunities and challenges in integrating knowledge graphs and AI for open, interoperable science.

# Save the date!

## DiTraRe Interdisciplinary Symposium 2: 6-7 April 2027


Save the date for our second Interdisciplinary Symposium on Digitalisation of Research, which will take place in Karlsruhe at KIT from 6 to 7 April 2027!

Leibniz ScienceCampus
Digital Transformation of Research

# Thank you for joining!

## Stay connected

- DiTraRe
  - Website: www.ditrare.de/en
  - Email: ditrare@fiz-karlsruhe.de
  - LinkedIn: www.linkedin.com/company/ditrare
  - Mastodon: social.kit.edu/@DiTraRe
  - YouTube: www.youtube.com/@DiTraRe
  - Zenodo: zenodo.org/communities/ditrare
- Discussion forum: www.ditrare.de/en/forum
- Newsletter: www.ditrare.de/en/newsletter

www.ditrare.de/en


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

KIT
Karlsruhe Institute of Technology

funded by Leibniz Competition