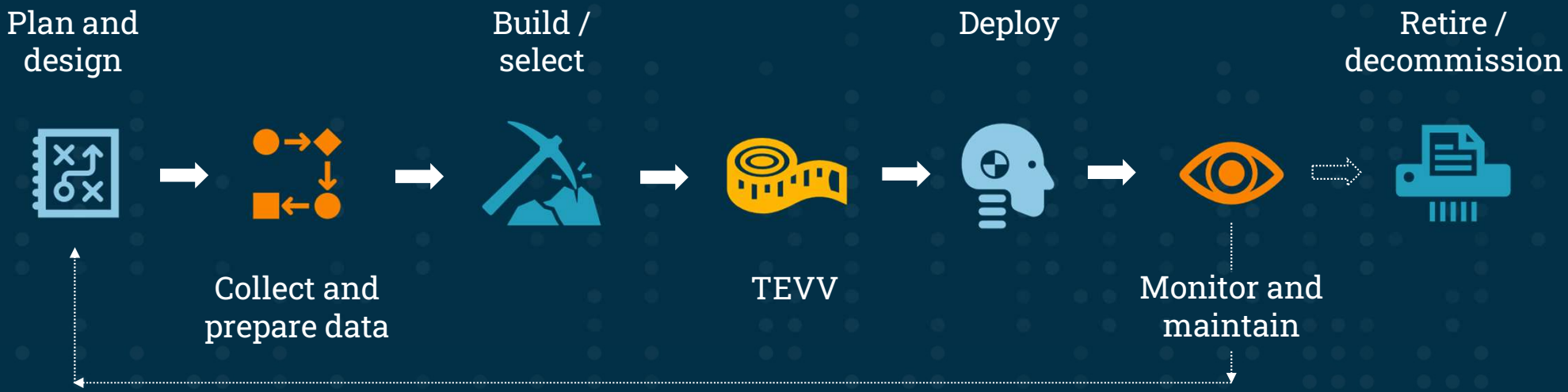
Fruit of the Poisonous Tree

Privacy, AI, and the Scaling Era's Original Sin

AI Development Life Cycle



The Risk

- Foundation models have been trained on massive amounts of publicly available data scraped from the internet.
- This includes unconsented personally identifiable information and sensitive data.
- The collection and processing of this data is non-compliant with global laws and regulations.
- Organizations building or procuring models trained on this data expose themselves to significant legal, financial, and operational risk.

DataComp CommonPool

Hong, R., Hutson, J., Agnew, W., Huda, I., Kohno, T., & Morgenstern, J. (2025). A common pool of privacy problems: Legal and technical lessons from a large-scale web-scraped machine learning dataset (arXiv:2506.17185v1). arXiv. https://doi.org/10.48550/arXiv.2506.17185

A Common Pool of Privacy Problems: Legal and Technical Lessons from a Large-Scale Web-Scraped Machine Learning Dataset

Rachel Hong[1], Jevan Hutson[2], William Agnew[3], Imaad Huda[2], Tadayoshi Kohno[1], Jamie Morgenstern[1]

ABSTRACT

We investigate the contents of web-scraped data for training AI systems, at sizes where human dataset curators and compilers no longer manually annotate every sample. Building off of prior privacy concerns in machine learning models, we ask: What are the legal privacy implications of web-scraped machine learning datasets? In an empirical study of a popular training dataset, we find significant presence of personally identifiable information despite sanitization efforts. Our audit provides concrete evidence to support the concern that any large-scale web-scraped dataset may contain personal data. We use these findings of a real-world dataset to inform our legal analysis with respect to existing privacy and data protection laws. We surface various privacy risks of current data curation practices that may propagate personal information to downstream models. From our findings, we argue for reorientation of current frameworks of "publicly available" information to meaningfully limit the development of AI built upon indiscriminate scraping of the internet.

KEYWORDS

Empirical Studies, Data Protection, Artificial Intelligence, Ethics, Web Scraping, Dataset Audit

1 INTRODUCTION

With the recent popularity in foundation models like ChatGPT and Midjourney [87, 100], machine learning practitioners often rely on data scraped from the web to train large language or vision models [21, 112, 124]. DataComp CommonPool, for instance, is one of the largest publicly available image-text dataset scraped from the web with over 12.8 billion samples [47]. This dataset has been downloaded over 2 million times at the time of writing with half a million downloads in the month of October 2024 alone [45], and its precursor LAION-5B [116] was used to train well-known image generation models like Midjourney, Stable Diffusion, and Google's Imagen [6, 87, 114]. Since machine learning models are a function of their training data, the downstream models trained on DataComp CommonPool may share problematic behavior [17], including the potential leakage of personally-identifiable information (PII) [25, 92]. Just as prior work highlights the importance of data-centric AI governance [56], we emphasize that regulating a dataset with such wide usage may be more effective than addressing the harms of every model one-by-one – in other words, tackling the "root" rather than the "leaves" as illustrated in Figure 11.

In our work, we use DataComp CommonPool as a case study of web-scraping and conduct an investigation into data privacy concerns. We perform a *legally-grounded audit*, one of the first to our knowledge, in which our audit findings inform our legal analysis on web-scraping, and vice versa, where recent legal literature on data privacy motivates our audit inquiries [68, 124]. Specifically, our audit asks: *What kinds of personally identifiable information are present in DataComp? How do current data cleaning practices address privacy concerns?* To do so, we draw upon prior frameworks on privacy [88], representation [38], and data filtering [64].

Our legal analysis considers how use of DataComp CommonPool for AI development might trigger application of and compliance obligations under existing privacy laws for developers and downstream deployers, including US state comprehensive privacy laws and international data protection laws. We consider and problematize current interpretations of "publicly available data" under existing privacy laws. We also consider how privacy risks and compliance obligations triggered by the production of DataComp CommonPool propagate to downstream models trained on this dataset. Lastly, we consider ongoing privacy risks that are currently not being addressed sufficiently by data filtering and other responsible data curation and hygiene practices, which informs recommendations on how policymakers might address these risks.

We make the following contributions:

(1) We find instances of personal information present in DataComp CommonPool, revealing various privacy concerns in web-scraped image-text datasets. For example, we uncover examples of personal information including credit card numbers and passport numbers, and we estimate at least 142,000 images depict resumes of individuals.
(2) We argue that no automated cleaning of web-scraped data can remove all PII and that ongoing cleaning methods are not sufficient to tackle privacy and must be scrutinized. Specifically, the DataComp CommonPool creators use a face blurring tool to preserve privacy, and we find that this tool fails to catch an estimated 102 million images of real human faces, demonstrating the importance of privacy tool assessments.
(3) We map these audit results to legal concerns to provide a critique of current data curation practices according to existing privacy laws. We also apply our findings from this widely used dataset to demonstrate shortcomings of existing privacy frameworks, such as the implications of the exemption for publicly available information.

We first present the context for web-scraped machine learning dataset development by detailing the history of DataComp CommonPool in Section 2, the stakeholders and artifacts associated in each step of the curation pipeline in Section 3, and related computer science and legal work in Section 4. We then present our audit methodology in Section 5 and the empirical results in Section 6. We use these findings to inform our legal analysis to determine the application of various data protection laws in Section 7. Finally in

1. University of Washington Paul G. Allen School of Computer Science & Engineering.
2. University of Washington School of Law.
3. Carnegie Mellon University Carnegie Bosch Institute.

DataComp CommonPool

- Publicly available image-text dataset
- 12.8 billion samples from the web
- Subset of Common Crawl
  - Massive, open-source repository of web crawl data

What does DataComp include?

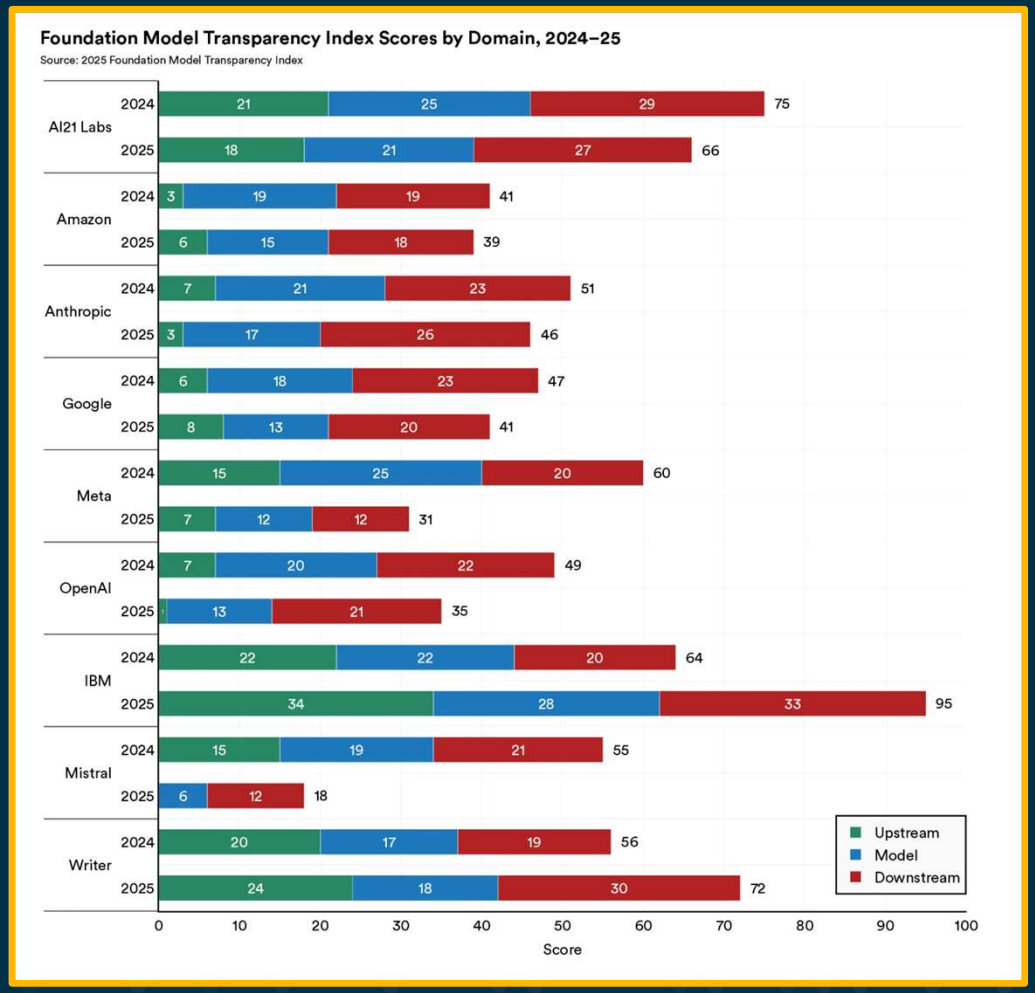
- Estimated
  - 102 million images of human faces
  - 200,000 resumes, including SSNs, disability status, DOB, dependents
- Credit cards (with security codes), driver's licenses, passports, birth certificates
- Children's information (names, faces, birth certificates, health status)
- Precise metadata (names, GPS coordinates)

Popularity

- Over 2 million downloads (June 2025)
- Precursor LAION-5B used to train well-known image generation models like Midjourney, Stable Diffusion, Google's Imagen (Nano Banana)
- Common Crawl used to train GPT, LLaMa, and PaLM
  - Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2025). Large language models: A survey (arXiv:2402.06196v3). arXiv. https://doi.org/10.48550/arXiv.2402.06196

Transparency Scores Declining

Stanford University, The Foundational Model Transparency Index



FTMI Dimensions

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2025

Source: 2025 Foundation Model Transparency Index

Major Dimensions of Transparency	AI21 labs Jamba 1.6	Qwen 3	amazon Nova Premier	Claude 4	deepseek DeepSeek-R1	Google Gemini 2.5	IBM Granite 3.3	Meta Llama 4	Midjourney V7	Medium 3	OpenAI o3	WRITER Palmyra X5	Grok 3	Average
Data Acquisition	92%	17%	17%	25%	17%	33%	100%	33%	0%	0%	8%	58%	0%	31%
Data Properties	0%	20%	0%	0%	20%	0%	100%	20%	0%	0%	0%	40%	0%	15%
Compute	22%	11%	11%	0%	44%	11%	100%	22%	0%	0%	0%	100%	11%	26%
Model Information	75%	75%	0%	25%	75%	0%	100%	75%	0%	0%	0%	75%	0%	38%
Model Access	50%	50%	50%	50%	50%	50%	100%	50%	0%	25%	0%	50%	0%	40%
Capabilities	75%	50%	50%	25%	50%	25%	75%	50%	0%	25%	25%	50%	25%	40%
Risks	60%	0%	40%	60%	20%	20%	100%	20%	0%	0%	60%	40%	0%	32%
Model Mitigations	60%	0%	60%	80%	20%	40%	80%	0%	0%	20%	80%	40%	0%	37%
Release	88%	63%	75%	75%	63%	88%	100%	50%	63%	38%	63%	88%	50%	69%
Usage Data	20%	0%	20%	60%	0%	0%	80%	0%	20%	0%	20%	100%	0%	25%
Impact	71%	0%	0%	29%	0%	29%	86%	14%	29%	14%	14%	86%	0%	29%
Post-deployment Monitoring	71%	0%	57%	57%	0%	43%	100%	29%	0%	43%	71%	86%	0%	43%
Model Behavior Policy	100%	50%	75%	100%	75%	75%	100%	75%	25%	0%	75%	50%	75%	67%
Acceptable Use Policy	80%	60%	80%	100%	60%	80%	80%	40%	60%	60%	60%	80%	60%	69%
Downstream Mitigations	100%	40%	100%	100%	0%	100%	100%	80%	40%	80%	100%	100%	40%	75%
Average	**64%**	**29%**	**42%**	**52%**	**33%**	**40%**	**93%**	**37%**	**16%**	**20%**	**38%**	**69%**	**17%**	

Privacy Risk (Jan. 2025)

- 144 countries w/ comprehensive national data privacy laws
- 82% of the world's population; 6.64 billion people
  - 100% Europe
  - 70-85% Americas
  - 76% Africa
- US: 19 states with comprehensive privacy legislation (Oct. 25)

Legal Precarity

- Publicly available ≠ free from legal obligations
  - I.e., personal data does not lose protected status because it's on the internet
- What is the lawful basis for processing?
- How will data subjects exercise rights?

GDPR: Lawful Basis

- Consent
- Contract
- Vital interest
- Legal obligation
- Public interest
- Legitimate interest

GDPR: Data Subject Rights

- Informed
- Access
- Rectification
- Erasure
- Restrict processing
- Data portability
- Object
- Automated decision-making, profiling

Laws with Similar Requirements

- US, California Consumer Privacy Act (CCPA)
- Brazil, LGPD
- PRC, PIPL
- India, DPDPA
- South Africa, POPIA

AI Risk (Jan. 2026)

- EU AI Act, Article 53
- CA AB 2013, Gen AI Training Data Transparency Act
- PRC Interim Measures on the Management of Gen AI Services
- US Federal Government
  - FY26 NDAA Sec. 1533, AI Model Assessment and Oversight
  - OMB M-26-04, “Increasing Public Trust in AI through Unbiased AI Principles”

EU AI Act

- Article 53, General Purpose AI (GPAI)
  - Data provenance documentation
  - Curation methodologies
  - Number of data points, main characteristics
- “Template for the Public Summary of Training Content for General-Purpose AI Models”

Article 53

"Template for the Public Summary of Training Content for General-Purpose AI Models"

2. List of data sources

This Section requires information about specific sources of data used to train the general-purpose AI model. In this section "dataset" should be understood as a single, pre-packaged collection of data. The filtering and pre-processing of data collected from the same pre-packaged collection should not be considered a new dataset to be disclosed separately in the sections below. If a particular dataset can be assigned to more than one of the categories below, providers should select the most relevant category and only report the dataset in that category, except in the case of synthetic data (see Section 2.5).

2.1. Publicly available datasets

This Section requires information about datasets that were used to train the model and which have been compiled by a third party, are made available publicly for free, and are readily downloadable as a whole or in predefined chunks, such as datasets and collections available on public repositories and online platforms, specialised websites, or snapshots of common crawl. The public availability of the datasets for free does not

[1] Excluding audio that is part of video, as this should be reported under the "video" modality instead. Furthermore, the Commission understands the modality of 'audio' to include 'speech'.

mean that the content at issue is necessarily free of rights since it may be subject to licensing arrangements or conditions of use (e.g., certain free and/or open licenses may determine the scope of the uses, including prohibiting uses relating to model training).

A dataset is considered to be "large" if the total data size for any one of the modalities contained in the dataset exceeds 3% of the size of all publicly available datasets for that modality used for training. The size of the dataset should be based on its size after pre-processing (for example filtering), and without splitting the dataset to prevent reporting circumvention.

Draft

Have you used publicly available datasets to train the model? ☐ Yes ☐ No

If yes, specify the modality(ies) of the content covered by the datasets concerned: ☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify…*

List of large publicly available datasets: *For each large dataset, provide the identifier/name of the dataset and a link through which the dataset can be accessed. If a link is not available, provide a general description of the dataset, including the approximate start and end dates of the data collection if known (otherwise indicate "not known"). If only part of the datasets has been used for the training, indicate the general approach to selecting those parts.*

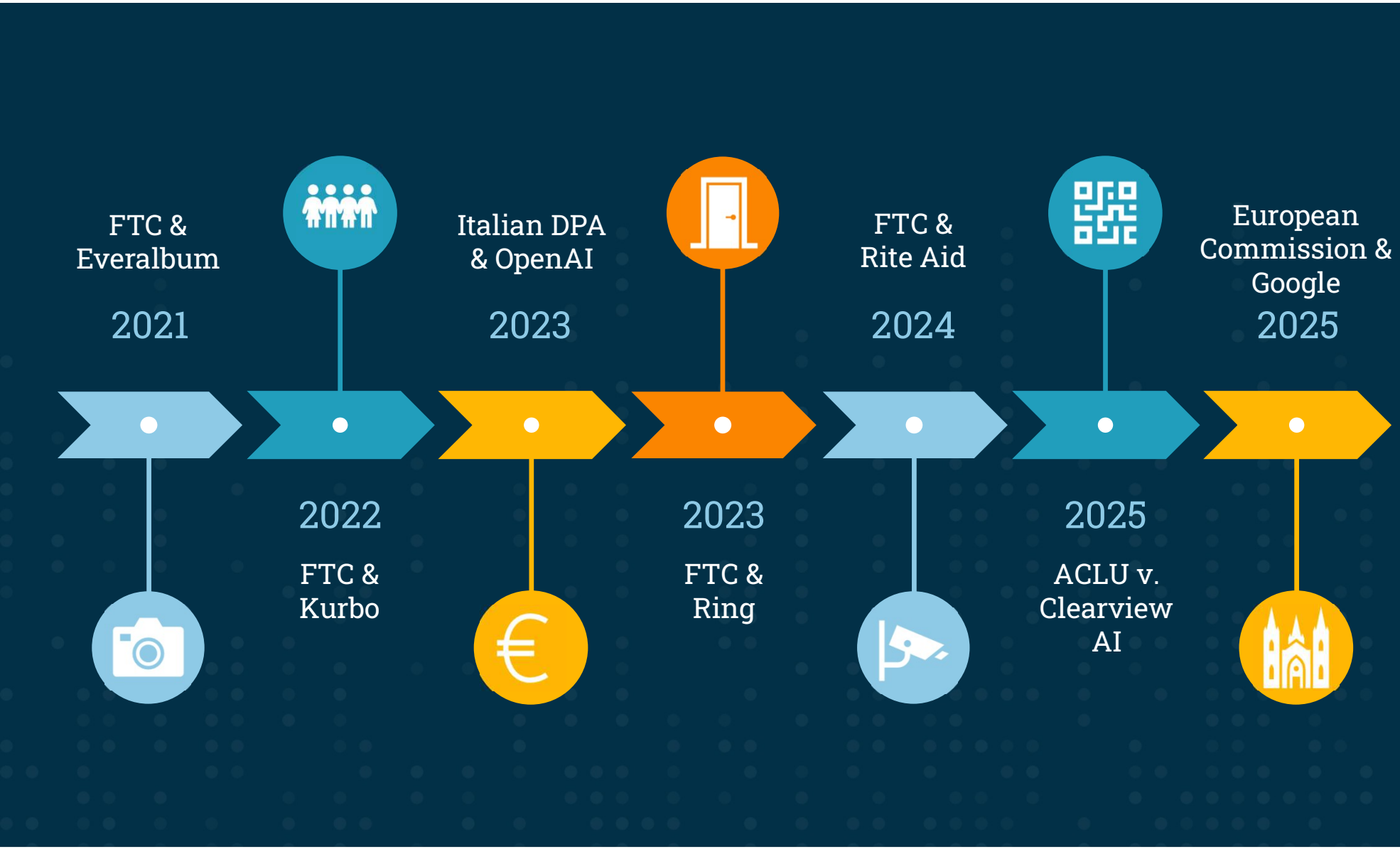
General description of other publicly available datasets not *For other publicly available datasets that are not listed above, provide a general description of their content. The description could include indication of: (i) the types of modality (e.g. text, images), (ii) nature of the content (e.g. personal data, copyright protected content,*

CA AB 2013

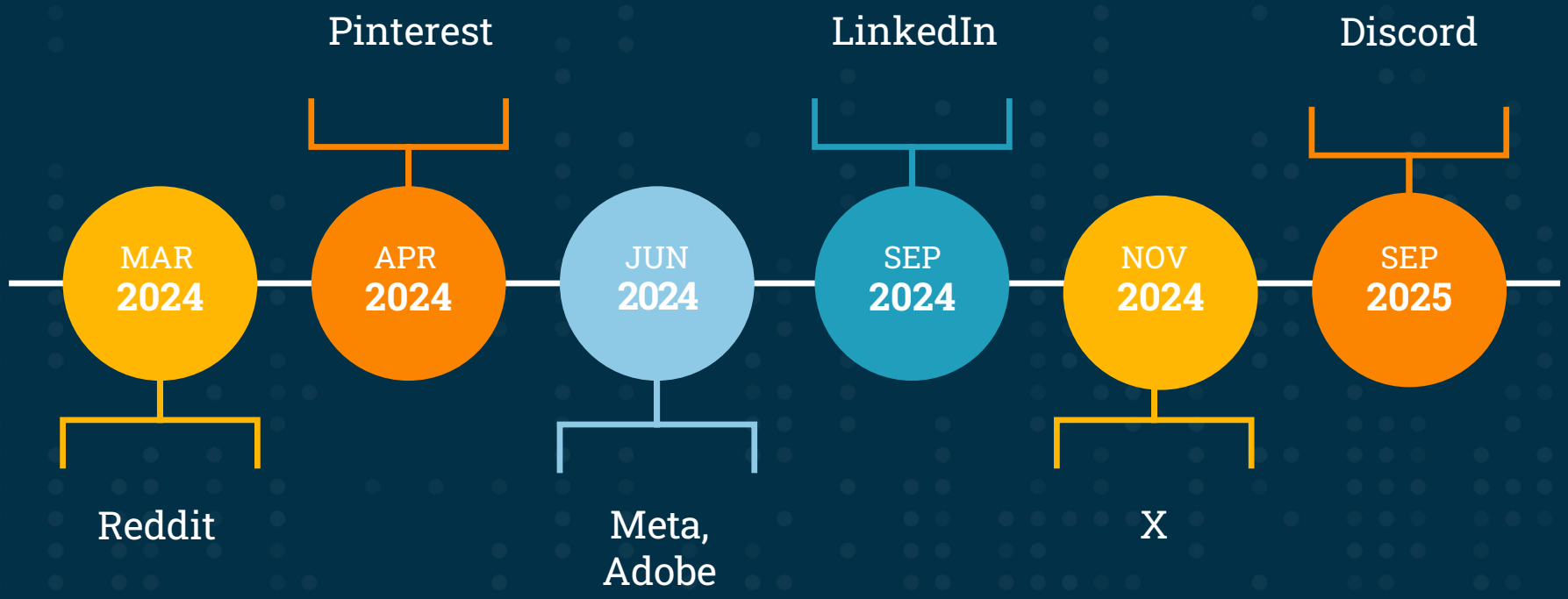
- Generative AI Training Data Transparency Act
  - Sources and ownership
  - Data characteristics
  - IP and copyright status
  - Personal information
  - Data provenance
  - Processing methods

Penalties

- Fines for non-compliance
  - GDPR: up to €20 million or 4% of global annual turnover
  - EU AI Act: up to €15 million or 3% of global annual turnover
- Algorithmic disgorgement, dataset deletion
- Sanctions, technical blocking
- Legal obsolescence (i.e., model functional, but unlawful)


FTC & Everalbum
2021
2022
FTC & Kurbo
Italian DPA & OpenAI
2023
2023
FTC & Ring
FTC & Rite Aid
2024
2025
ACLU v. Clearview AI
European Commission & Google
2025

Licensing & ToS Expansion



What can we do?

- In-house development
- Third-party risk management

In-house Development (1)

- Internal, enterprise data
- Public domain, openly licensed datasets
  - Common Pile
  - Common Corpus
- Commercially licensed datasets
  - Reddit, Toll-Wall news content
  - Bloomberg
  - Reuters
  - Data annotation firms (e.g., Scale AI)

In-house Development (2)

- Privacy-enhancing technologies
  - E.g., anonymization, de-identification, differential privacy
- Synthetic data
  - E.g., Gretel.ai, Mostly.ai
- Inventory your "scrapers"
  - Ask data science team: using Common Crawl? Or web-scraping scripts? If yes, audit.
- On lawful basis: EDPB Opinion 28/2024

TPRM

- Request documentation
  - AI Bill of Materials (AIBOM)
  - Data provenance
  - Model, system cards
  - Privacy / Data Protection Impact Assessments
- Data Processing Agreements
- Request broad legal indemnification

What haven’t we discussed?

- Shadow AI
- Security safeguards throughout data pipeline, AIDLC
- Retention and disposal
- Input, output validation, handling
- Adversarial ML attacks
- Transparency, explainability requirements

Questions


Kyle David PhD
3x Bestselling AI & Privacy Author | CIPP/US/E, CIPM, AIGP, FIP, CISSP, AAISM