Michael Best®

AI & Data: What You Need to Know

Derek C. Stettner

November 20, 2025

Overview

- What will we cover (or at least touch on)
  - Intellectual property and data privacy issues you need to understand to implement your AI projects
  - Use Case – IBM's Slogan - Your Data Our AI
    - You are using someone else's model/AI with data that you "think" is yours
  - Privacy basics
  - Recent copyright fair use rulings
  - Terms and conditions
  - Data licensing
- What we won't cover
  - Topics from prior webinars

Perspectives on AI Progression and Data

- AI timeline
  - Predictive AI – old
  - Generative AI – ready for 4k
  - Agentic AI – toddler aged - two or more generative AI bots talking to one another – newer
  - Artificial General Intelligence – still not here
- Data -
  - What AI learns from or is trained on – data includes all sorts of things
  - Reinforcement learning – can influence the data that an agent or model learns from

Intellectual Property Refresher

- **Trade Secret** (confidential information) – any information that is confidential, subject to reasonable efforts to maintain its secrecy, and provides a competitive advance by not being generally known
  - If inadvertently made public, trade secret rights lost
- **Patent** - legal right granted to inventors to exclude others from making, using, selling, or importing their claimed invention for a limited period. Inventions includes processes, machines, articles of manufacture, compositions of matter, and improvements of these things
- **Copyright** – bundle of rights protecting original works of authorship fixed in a tangible medium of expression. Copyright grants the author exclusive rights to reproduce the work, prepare derivative works, distribute copies, and perform or display the work publicly
- **Trademark** – is any word, phrase, symbol, design, or combination thereof that identifies and distinguishes the **source of goods or services** from those of others. Trademarks can include:
  - Brand names (e.g., Nike)
  - Logos (e.g., the swoosh)
  - Taglines (e.g., “Just Do It”)
  - Product packaging and trade dress

Use Case – Your Data Our AI

- What is “your” data?
  - Anything that happens to have ended up on our information technology systems?
  - Anything that we collected from our customers or the public?
  - Anything that we are not blocked from accessing that is on the internet?
  - Anything that our machines generate?
  - **No, no, no, maybe/unclear particularly if you sold your machine to a customer**
- I/we want to own the data or at least have the rights to use it

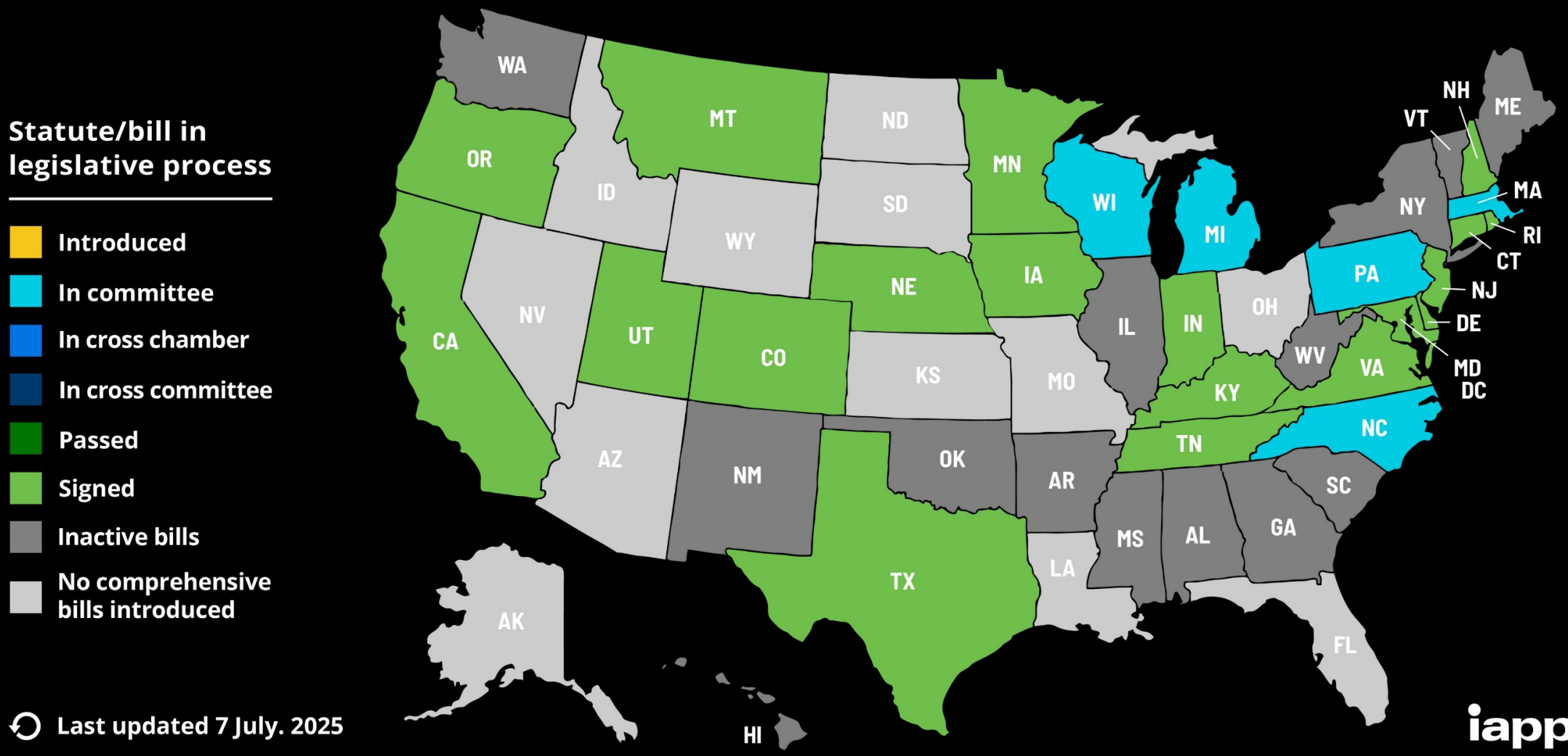
Before You Can Own - Understand Data Collection, Use, and Protection

- PII, PHI, and other sensitive data must be collected in accordance with applicable privacy and other laws (e.g., CCPA, GDPR, and HIPAA) (more later)
  - Should obtain broad consent – e.g., to de-identify, aggregate, create derivative data, and use for purposes of training AI
- Data subject has controlling rights over PII or PHI
  - Consent to use PII or PHI can be withdrawn
- To account for withdrawal of consent
  - Build AI solution so that it can be trained on de-identified data and does not require PII/PHI as input
  - Build technology so that the AI feature can be turned off
- However, properly collected PII or PHI and data that is not PII or PHI is potentially
  - Copyrightable, protectable as a trade secret
  - Just like open source, there is an open data movement

Owning Data

- Can you “own” data? Maybe, but can’t own raw facts
- Some data copyrightable (text, photographs, news articles, graphical works) (compilation copyright and/or trade secrets in some data)
- EU Database Directive
  - Copyright protection – protects structure not the raw data
  - Sui generis right – protects databases that lack originality but involve substantial investment in obtaining, verifying, and presenting data
- Contract Rights
  - Party B recognizes and agrees that:
    - a) The Data is Party A’s valuable property; b) The Data includes Party A’s trade secrets; c) The Data is an original compilation pursuant to the copyright law of the United States and other jurisdictions; and/or d) Party A has dedicated substantial resources to collecting, managing, and compiling the Data.

US State Privacy Legislation Tracker 2025



PII (Help Your AI Technologists Understand)

- What is PII
  - Direct identifiers (uniquely identify a person):
    - Full name; Social Security Number (SSN); Driver's license number; Passport number; Email address (personal); Phone number; Home address
  - Indirect identifiers (can identify someone when combined with other data):
    - Date of birth; Gender; ZIP code; IP address; Device ID; Employment details
- What is NOT PII
  - Information that cannot reasonably identify an individual, such as:
    - Aggregated or anonymized data (e.g., "average age of users is 35")
    - Generic demographic info without linkage (e.g., "20% of users are from Wisconsin")
    - Randomized identifiers that cannot be traced back to a person
- Data that may require special treatment
  - PHI (next slide), financial info, biometric data (fingerprints, facial scans).
- Context matters:
  - A ZIP code alone may not be PII, but combined with birth date and gender, it could identify someone
  - Possible to reidentify – take a deidentified data and combine with another source (e.g., public information)

PHI (Help Your AI Technologists Understand)

- PHI (Protected Health Information) refers to individually identifiable health information that meets these conditions:
- Relates to health
  - Past, present, or future physical or mental health condition
  - Provision of healthcare
  - Payment for healthcare
- Identifies the individual (directly or indirectly)
  - Includes any of the 18 HIPAA identifiers, such as:
    - Name; Address (smaller than state); Dates (birth, admission, discharge, death); Phone numbers; Fax numbers; email; Social Security number; Medical record number; Health plan beneficiary number; Account numbers; Certificate/license numbers; Vehicle identifiers; Device identifiers; Web URLs, IP addresses; Biometric identifiers (fingerprints, retinal scans); Full-face photos; Any unique identifying code
- Created or maintained by a HIPAA-covered entity or business associate
  - Healthcare providers, health plans, clearinghouses, or their vendors

What is NOT PHI (Help Your AI Technologists Understand)

- Information that does not meet all three conditions above is not PHI under HIPAA. Common exclusions:
  - De-identified health data (all 18 identifiers removed)
  - Aggregated statistics (e.g., “20% of patients have diabetes”)
  - Employment records (even if health-related, when held by employer for HR purposes)
  - Student health records (covered by FERPA, not HIPAA)
  - Data collected by non-covered entities (e.g., fitness app not acting as a business associate)
  - Research health information kept outside medical records and without identifiers

Reducing Your Risk

- Educate your team to be privacy aware –
  - Do they know what PII and PHI are so that they ask for help when they encounter it?
- Build with privacy in mind?
  - Do they really need PII and PHI to achieve our objective?
  - Can we get permission?
  - Can we use aggregated/deidentified data?
  - Can we use synthetic data?
  - Can we license data from a third party that comes with warranties?

What About Copyright Protection?

- Does the use of data to train AI infringe copyrights?
- Is it fair use? - Kadrey v. Meta Platforms and Bartz v. Anthropic
- Defendants Meta Platforms and Anthropic were sued in California
  - Among other issues in the cases, the defendants argued that their use of copyrighted material (books) for training AI was fair use under the copyright statute
  - Under 17 U.S.C. § 107, you must evaluate
    - (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes
    - (2) the nature of the copyrighted work
    - (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole, and
    - (4) the effect of the use upon the potential market for or value of the copyrighted work

Key Fair Use Factor - Transformative

- Both Judges ruled that LLM training was transformative (purpose and character),
  - Fair use is fact intensive - another court looking at different LLMs might think differently
  - A generative pretrained transformer (GPT) must be transformative because transform is in the name.
    - Name reflects its ability to transform sequences of data (like text) using a mechanism called attention
    - A transformative use does not merely supersede the original work but instead provides new insights, commentary, or utility, thereby serving the broader goals of copyright law to promote creativity and public knowledge. Campbell v. Acuff-Rose Music, Inc. 510 U.S. 569

Outcome of Meta and Anthropic Cases

- Meta showed that the use was transformative and the plaintiff failed to demonstrate an impact on the market for the copyright work. The case demonstrates that a plaintiff may lose when they fail to prove adequate evidence on market value. Better evidence on market value in a different case might change the fair use analysis.
- Meta case still on going and probably will be appealed
- In Anthropic, the plaintiff was able to demonstrate that Anthropic never lawfully obtained the copyright works (e.g., by purchasing them). So, there is no fair use if you pirate the book first. The case demonstrates that pirating source material can override a fair use defense.
- Anthropic's alleged use of pirated books to train its AI model was set to go to trial, but the case preliminarily settled for $1.5 billion, roughly $3,000 each for the nearly 500,000 registered books that were allegedly pirated.

The Fair Use Case that Gets Less Attention

- Thomson Reuters v. Ross Intelligence, 765 F. Supp. 3d 382 (D. Del. 2025), the court rejected a fair use defense (currently on appeal before the Third Circuit)

What about Terms and Conditions?

Reddit Inc. v. SerpApi LLC, 25-cv-08736, U.S. District Court, Southern District of New York (Manhattan). Reddit Inc. sued multiple companies over alleged data scraping without permission, a sign of the growing demand and value of original data in the burgeoning AI industry.

Several companies have been illegally collecting Reddit data via Google search results for the purpose of reselling it, according to the complaint filed in federal court in Manhattan. Perplexity has been buying that data from at least one of the companies, the suit alleges.

Reddit is seeking monetary damages and a court order to stop the alleged scraping and use of its data in violation of federal copyright law. Reddit's growing repository of data has become a valuable commodity given the rise in AI models that rely on massive troves of information for training and surfacing relevant results.

Reddit has already inked deals with OpenAI and Alphabet Inc.'s Google to license its data for training purposes, but has taken legal action against others it believes to be using the data without a formal agreement.

Reddit sued AI firm Anthropic PBC earlier this year in San Francisco court over similar data scraping allegations.

Reddit v. SerpApi

“AI companies are locked in an arms race for quality human content — and that pressure has fueled an industrial-scale ‘data laundering’ economy,” **Ben Lee**, Reddit’s chief legal officer, said in a statement shared with Bloomberg News. “Reddit is a prime target because it’s one of the largest and most dynamic collections of human conversation ever created.”

Perplexity spokesperson Beejoli Shah said the firm “will always fight vigorously for users’ rights to freely and fairly access public knowledge.”

“Our approach remains principled and responsible as we provide factual answers with accurate AI, and we will not tolerate threats against openness and the public interest,” Shah said in an emailed statement. Perplexity called the lawsuit a “show of force” in Reddit’s data negotiations with OpenAI and Google in a statement posted on Reddit.

“Oxylabs has always been and will continue to be a pioneer and an industry leader in public data collection, and it will not hesitate to defend itself against these allegations,” Grybauskas said. “Oxylabs’ position is that no company should claim ownership of public data that does not belong to them.”

How Does Reddit Do IT?

Terms of Use - The Services may contain **information**, text, links, graphics, photos, videos, audio, streams, software, tools, or other materials ("Content"), including Content created with or submitted to the Services by you or through your Account ("Your Content"). We take no responsibility for and we do not expressly or implicitly endorse, support, or guarantee the completeness, truthfulness, accuracy, or reliability of any of Your Content.

By submitting Your Content to the Services, you represent and warrant that you have all rights, power, and authority necessary to grant the rights to Your Content contained within these Terms. Because you alone are responsible for Your Content, you may expose yourself to liability if you post or share Content without all necessary rights.

You retain any ownership rights you have in Your Content, **but** you grant Reddit the following **license** to use that Content:

When Your Content is created with or submitted to the Services, you grant us a **worldwide, royalty-free, perpetual, irrevocable, non-exclusive, transferable, and sublicensable license** to use, copy, modify, adapt, prepare derivative works of, distribute, store, perform, and display Your Content and any name, username, voice, or likeness provided in connection with Your Content in all media formats and channels now known or later developed anywhere in the world.

This license includes the right for us to make Your Content available for **syndication, broadcast, distribution, or publication by other companies, organizations, or individuals who partner with Reddit**. You also agree that we may remove metadata associated with Your Content, and you irrevocably waive any claims and assertions of moral rights or attribution with respect to Your Content.

Reddit's Privacy Policy

Reddit Is a Public Platform

Much of the information on the Services is public and accessible to everyone, even without an account. By using the Services, you are directing us to share this information publicly and freely.

Reddit allows moderators to access Reddit content using moderator bots and tools. Reddit also allows other third parties to access public Reddit content using Reddit's developer services, including Reddit Embeds, our APIs, Developer Platform, and similar technologies. We limit third-party access to this content and require third parties to pay licensing fees for access to larger quantities of content. Reddit's Developer Terms are our standard terms governing how these services are used by third parties. Please review our public content policy here for more information about how your public content is publicly available and accessible to anyone with access to the internet.

How We Share Information

. . .

Aggregated or de-identified information. We may share information about you that has been aggregated or anonymized such that it cannot reasonably be used to identify you. For example, we may show the total number of times a post has been upvoted without identifying who the visitors were, or we may tell an advertiser how many people saw their ad.

X Corp v. Bright Data LTD. (May 2024)

- X Corp (f/k/a Twitter) complaint includes various claims/causes of action.
- Citing to *hiQ Labs, Inc. v. LinkedIn Corp.* (9th Cir. 2022), the District Court in the N.D. of California dismissed complaint that alleged, among other things, violation of terms of use by Bright Data for scraping public "data."
- Browser wrap terms of use found enforceable, but X Corp. failed to demonstrate injury to its IT systems and federal copyright law pre-empts state law contract claim. Among other things, terms of use can not supplant fair use nor provide rights over non-copyrightable content.
- (Gosh, are all licenses of non-copyrightable data pre-empted?)
- Presumably X Corp does not take ownership of "data" to maintain liability shields provided under Section 230 of the CDA and Section 512 (a) of the DCMA.
- Do copyright owners of the data have a claim against Bright Data? Maybe, the license in the terms of use is between X Corp and the users; not the users and Bright Data.

Privacy Risks

- Privacy laws (like GDPR, CCPA) protect personal data, regardless of whether it's publicly available
- If you collect, store, or process data that identifies individuals (names, emails, photos, IP addresses), you may need consent or a lawful basis
- Vermont: On April 25, 2025, Attorney General Charity Clark refiled a lawsuit against Clearview AI, alleging violations of the state's Consumer Protection Act. The complaint centers on Clearview's purported scraping of facial recognition data without consent
  - Example of how states may pursue scraping even absent AI-tailored statutes.
- Connecticut: Connecticut amended its Data Privacy Act, Conn. Gen. Stat. §§ 42-515 et seq., (effective July 1, 2026), to require that privacy policies explicitly disclose whether personal data is collected, used, or sold for training large language models.
  - First law of its kind in the U.S. and includes a 60-day cure period before the AG can initiate enforcement.
  - Reflects concern over the use of consumer data in AI training and sets a precedent for other states considering similar transparency mandates.

Copyright Risks

- Fair Use
  - Cases split 2 to 1, highly fact intensive
  - Must have purchased the works to begin with
- It is possible that AI output could infringe the rights of others?
  - Yes
  - GitHub Duplicate Detection Filter
  - Disney Enterprises Inc. et al. v. Midjourney Inc., Case No. 2:25-cv-05275 (C.D. Cal., filed June 11, 2025)
    - Disney and Universal filed suit against Midjourney
    - Studios allege that Midjourney trained its AI models on copyrighted material from Disney and Universal without permission
    - Allowed users to generate unauthorized images of iconic characters, including Darth Vader, Elsa, Minions, and others
    - Ignored requests to stop and continued monetizing the service.

Terms and Conditions Risk

- Terms and conditions are generally enforceable
- Some limited rights to scrap public data
- But
  - Copyright owners may still have a cause of action if the data is copyrightable
  - Might still violate privacy laws (see above)
- Less Risk
  - Scraping fully anonymized or aggregated data (no identifiers)
  - Using data under open licenses or from official open data portals
  - Complying with robots.txt and site terms

Licensing Data In

- Perhaps you need data to implement your AI strategy?
  - PII, PHI, or IP rights involved? If so, was technology built with privacy by design?
  - Acquired respecting copyrights (IP), privacy, and contractual rights (terms of use and licenses)?
  - Can the licensor provide warranties on these points?
  - Is the data curated? (selected, organized, and reviewed by human experts) (more expensive)
  - Additional warranties regarding curation?
  - Can provider/vendor keep prompts and customer training data confidential?
  - Data could be unusually sensitive and require extra confidentiality/privacy precautions

Takeaways

- You don’t have the right to use the data just because you have the data in your possession or no one stopped you from obtaining it
- If the data qualifies as PII or PHI
  - Get permission to use it
  - Deidentify (less risk, but still risk)
  - Don’t use
- If the data is copyrightable
  - Get permission to use it
  - Rely on fair use (less risk, but still risk)
  - Don’t use
- Law still in flux

Questions?

Presenter

Derek C. Stettner
Partner
dcstettner@michaelbest.com
414.225.4947

Thank you!

Michael Best