24 October 2024

Data Scraping in a Data-Driven Landscape

Rohan Massey
Ropes & Gray

Ellen Gilley
Takeda Pharmaceuticals

Colin Rooney
Arthur Cox

Privacy+ Security Forum

Speakers

Rohan Massey

Partner
Ropes & Gray (London)

Ellen Gilley

R&D Legal – Data & Technology
Takeda Pharmaceuticals (Boston)

Colin Rooney

Partner
Arthur Cox (Dublin)

What Are We Talking About?

What is Data Scraping?

Definition: Automated extraction of data from websites and online sources.

Methods: Web scraping tools, APIs, and manual extraction.

Applications: Market research, competitive analysis, academic research, and more.

Why is Data Scraping Important?

Access to Large Datasets: Enables the collection of vast amounts of data quickly.

Real-Time Data: Provides up-to-date information for decision-making.

Cost-Effective: Reduces the need for manual data collection.

The Old – Internet Web Bots And Scrapers

- EU – *Ryanair Limited v PR Aviation BV (2015)*

  - European Court of Justice ("ECJ") considered whether website operators can use their service terms to prohibit the use of scraping tools.
  - ECJ ruled that website operators can impose such restrictions where they cannot otherwise rely on IP rights (i.e., copyright, database rights) to protect their data.
  - Ryanair has subsequently sued several price comparison sites for using scraping tools: In January 2019, a Romanian court upheld its breach of contract, copyright and database claims, but a Swiss court rejected similar claims in May 2019.

- U.S. – *hiQ Labs, Inc. v LinkedIn Corporation* (2022)

  - Ninth Circuit Court of Appeals upheld the district court's decision to allow HiQ to continue accessing LinkedIn's public profiles.
  - The Court reasoned that hiQ's scraping of publicly accessible data likely did not violate the CFAA and emphasized the public nature of the data and potential anti-competitive implications of LinkedIn's actions.

The New(ish) – Artificial Intelligence (AI)

- **Machine Learning (ML)**
  - Subfield of AI, which refers to computer systems that learn without explicit programming. Algorithms iteratively train computer models, which are then used to generate output (e.g., predictions) based on input data.

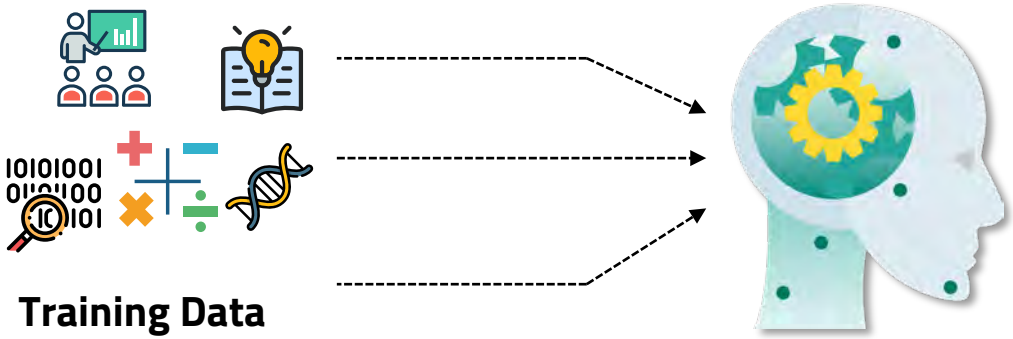
- **Generative AI**
  - Subfield of ML that goes beyond predictions to generate entirely new content, whether images, video or text – generative AI usually produces material in response to prompts.

- **LLMs:**
  - Large Language Models, a type of AI trained on vast amounts of text data to understand and generate human language.

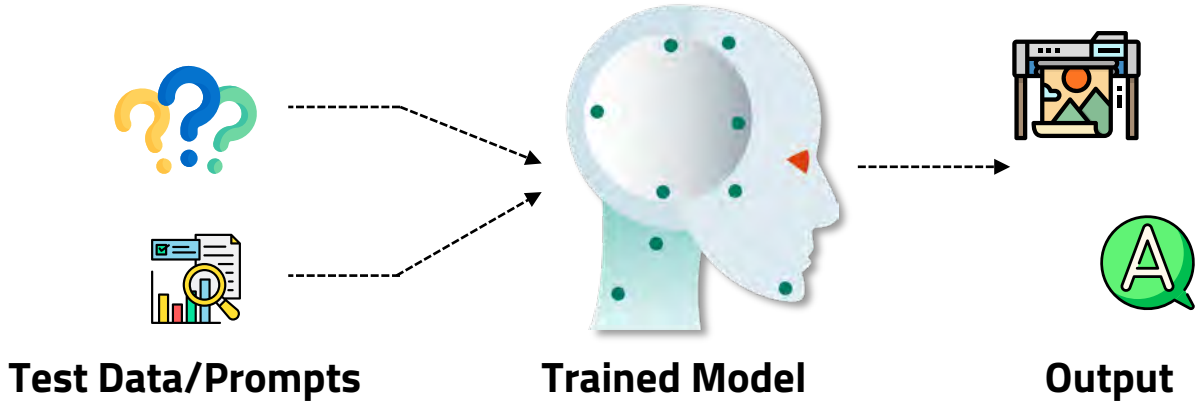
General Framework of AI and ML

Privacy+ Security Forum

Training the Model



Using the Model



Why Are Companies Interested In Data Scraping?

Data Scraping – the 'pros'

- **Data has been commoditized as an asset**
  - Look at all the "free" online services now available.
- **Organizations can currently set their own level of risk tolerance**
- **OpenAI GPT-3**
  - Approximately 175 Billion trainable parameters, Full version of OpenAI GPT-3 is the largest generative AI model trained to date…but rumors abound that GPT-4 has 1.76 trillion parameters, which will have a huge impact on the AI landscape.
  - GPT-3 model was trained on about **45 terabytes of text data** from multiple sources, including Wikipedia and books.
  - Some of the Datasets used to train the OpenAI model include:
    - **Common Crawl**: petabytes of data collected over eight years years of web crawling, containing raw web page data, metadata extracts and text extracts.
    - **WebText2**: text extracted from web pages showing all outbound Reddit links from posts with three+ upvotes.
    - **Wikipedia**: English language pages
- ***The greater the dataset, the greater the knowledge…***

Uses of Generative AI Tools In Life Sciences

- Drug discovery: AI models can generate new molecules with specific properties for drug development, reducing the time and cost of the discovery process.
- Image analysis: Generative AI can be used to generate synthetic medical images for training and validation of other AI models that assist in diagnosis and treatment planning.
- Personalized medicine: AI can analyze a patient's genomic and medical data to generate personalized treatment plans.
- Predictive modeling: AI can generate models that can predict disease progression, drug efficacy, and toxicity based on patient data.
- Synthetic biology: AI models can be used to design and optimize new biological systems, such as metabolic pathways or synthetic organisms.

Credit

The Challenges of Data Scraping

Privacy+ Security Forum

Data Scraping – the ‘cons’

- **Contamination**
  - Comparison to use of open-source software – what are future obligations?
  - Risk of poor data being ingrained and magnified in importance (accuracy, bias etc)
  - “Scope creep” – unforeseen future use may be in breach of EULA or informed consent
- **Regulatory and legal risks**
  - Privacy and data protection laws
  - IP infringement
  - Anti-trust claims
  - Breach of contract

The regulatory challenges

- **Established laws apply pre-internet, to internet and now to AI:**
  - Data protection
  - IP
  - Contract
  - AI-specific regulation

- **EU and U.S. are not aligned**
  - U.S. appears to be more tolerant of data scraping (for now)
  - EU looks to prohibit data scraping
    - Recent joint statement from DPAs
    - Comments from Dutch DPA and EDPB

Legal Frameworks for Data Scraping

- **US Legal Framework**
  - Computer Fraud and Abuse Act (CFAA): Prohibits unauthorized access to computer systems.
  - Copyright Law: Protects original works of authorship, including website content.#
  - Case Law: Notable cases like *LinkedIn vs. hiQ Labs*.
- **EU Legal Framework**
  - General Data Protection Regulation (GDPR)
  - ePrivacy Directive
  - Database Directive: Protects databases and the data they contain.
- **UK Legal Framework**
  - Data Protection Act 2018 (UK GDPR)
  - Copyright, Designs and Patents Act 1988: Protects original works, including digital content.

Similarities & Differences

Common Elements

- **Data Protection:** Emphasis on protecting personal data and privacy.
- **Copyright Protection:** Safeguarding original content from unauthorized use.
- **Unauthorized Access:** Legal restrictions on accessing computer systems without permission.

Key Differences

- **US vs. EU/UK:** The US focuses more on unauthorized access, while the EU/UK emphasize data protection and privacy.
- **Enforcement:** Variations in how laws are enforced and penalties imposed.
- **Case Law:** Different legal precedents and interpretations in each jurisdiction.

Data Protection Laws – How do they apply?

- **The General Data Protection Regulation ("GDPR")** applies to organisations which are (i): based in the EU and process personal data in the context of the EU establishment; or (ii) based outside the EU but offer goods or services to, or monitor the behaviour of, individuals located in the EU.
- Does the GDPR apply to bots and scraping technologies?
  - Where the processing is undertaken by its establishments in the EU → Yes.
  - Where it monitors the behaviour of individuals in the EU → Yes (scraping = monitoring).
- What is "processing"?
  - Any operation performed on personal data (e.g., collecting, organising, storing, deleting, etc.).
  - Includes the initial use of bots / scraping and any subsequent use of the collected personal data.
- What is "personal data"?
  - Any information relating to an identified or identifiable natural person.
  - Identifiable person → one who can be identified, directly or indirectly, including by reference.
  - Identifier → e.g., name, photograph, location data, IP address, opinion, health information, etc.

What does GDPR require?

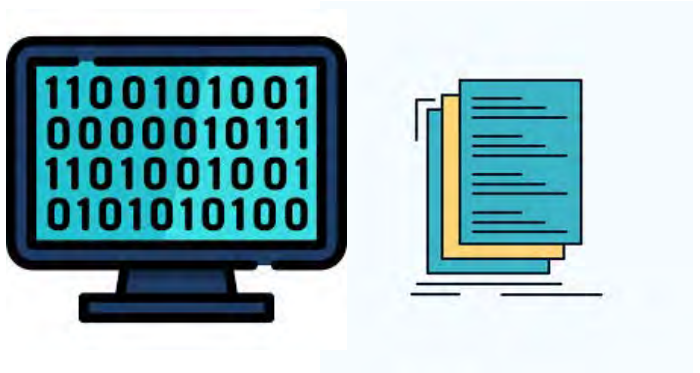
- **The GDPR does not explicitly prohibit the use of bots or other scraping technologies.**
- **But where the use of both involves processing of personal data (i.e., almost always), one must comply with the obligations the GDPR imposes on data controllers. These include:**
  - Providing fair processing notices to data subjects;
  - Putting in place appropriate policies and procedures to demonstrate compliance;
  - Entering into agreements with processors which process the data;
  - Implementing technical and organisational measures to ensure security of the data;
  - Reporting certain personal data breaches to regulators and affected individuals;
  - Responding to data subject rights requests (e.g., to provide, delete and restrict data); and
  - Protecting the data when they are transferred outside the European Economic Area.

GDPR – What are the challenges?

- **The service terms of most social media and similar platforms expressly prohibit scraping.**
- **This arguably means that if one uses non-sensitive personal data collected from platforms in breach of their terms, it is processing the data unlawfully and also in breach of the GDPR.**
  - "Arguably" → This point has not been tested in the privacy context, but the requirement to process data lawfully may be impinged where the collection of data was itself unlawful.
- **The privacy notice individuals must be given must provide information about (i) the sources from which the personal data originate, and (ii) whether the data came from publically available sources.**
  - It must also state that it is permitted to process this data may violate its transparency obligations; and
  - Give individuals a misleading impression about the collection and use of their personal data.

AI & IP Infringement Liability

- **AI models are typically trained on large data sets.**
- **If the AI platform uses data inputs to train models without obtaining sufficient rights, generated output may be subject to multiple sources of liability.**
  - A lawsuit filed in California on June 28, 2023, alleged that OpenAI amassed its training data improperly by scraping social media sites and using personal data without permission, and argued that restrictions on outputs generated with that data should be imposed.
- **Fair use: Courts have yet to decide if generative AI's use of protected code to generate output is transformative fair use. (*Google LLC v. Oracle America, Inc., 141 S. Ct. 1183* (2021)), but this may be a defense to copied code snippets.**



Risk Mitigation

Data Protection Laws – Addressing the challenges

- **GDPR provides a legal basis for processing sensitive personal data which are "manifestly made public by the data subject" (i.e., including data posted on social media platforms).**
- **If an individual has taken deliberate steps to make the data public, the platform's prohibition on scraping may be unenforceable….**
  - This argument does not appear to have been tested in the privacy context and does not apply to non-sensitive personal data.
    - Limited EU regulatory guidance on scraping does not provide clarity as to its legality under data protection law → WP29 Opinion on Facial Recognition states that "…images may also be unlawfully obtained by scraping other public sites such as search engines".
    - Prior to the AI language models and the Dutch DPA announcing that data scraping is usually illegal (1 May 2024) the use of scraping tools has not to date been an enforcement priority for EU regulators, it will be interesting to see what action is now taken.

AI & Anti-Hacking Restrictions

- **Scraping data from servers to gather input to use in training data models may be subject to potential statutory restrictions:**
  - Protecting computer systems against unlawful access
  - Establishing separate liability in addition to contract and IP
  - Federal Computer Fraud and Abuse Act (CFAA)
  - State Law Statutes
    - California Computer Data Access and Fraud Act
    - Virginia Computer Crimes Act

Risk Mitigation Strategies –Appropriate Use of AI

- **Policy compliance:** Ensure that generative AI use is consistent with company policies, standards and values. Consider creating policy terms specific to the use of generative AI tools, or legal review processes/checklists for the use of third-party AI tools.
- **Protection of Inputs:** Consider trade secret risks regarding all inputs.
- **Security:** Independently verify AI-generated results and actively consider how to use outputs.
- **Discrimination/Bias:** Use multiple frameworks to address potential bias/discrimination through the use of generative AI tools, whether resulting from training data set, training algorithm or output algorithm. Consider adopting standardized ethical principles/norms.

Data Scraping in a Data-Driven Landscape: The Takeaways

Privacy+ Security Forum

What and where now?

- **Data scraping is a powerful tool with significant benefits and challenges.**

- **Digital world is increasingly regulated and complex**
  - The EU relies on 100+ pieces of legislation.
  - Is the “Brussels Effect” working / likely to work?
  - Rise of extra-territorial effects and proportionality
  - Legal frameworks vary across jurisdictions, emphasizing the need for compliance.

..... Is The Stable Door Just Slamming In The Wind?

- AI and LLMs face unique challenges when using scraped data, requiring careful consideration of legal, ethical, and technical issue
- Is it too late to regulate what has already happened?
- Can AI unlearn – do we need it to?
- Do we need globally aligned regulation?
- Will greater regulation or clarification hamper or promote “progress”?

Reference Materials

- **EU Position**
  - EU Ai Act - https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689
  - EU and UK Regulators Statement - https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf
  - https://www.edpb.europa.eu/system/files/2024-10/edpb_guidelines_202401_legitimateinterest_en.pdf
- **US Position**
  - hiQ - https://casetext.com/case/hiq-labs-inc-v-linkedin-corp-5
  - Ryanair US Complaint - https://drive.google.com/file/d/13QWzOI_P8XpKGGkgzfJcAuYEIRpN2Xob/view
  - https://www.justice.gov/jm/jm-9-48000-computer-fraud
  - Jury verdict - https://drive.google.com/file/d/1Rq0KHoZ-FXfnSrNFQn45UGV-LhwaZ1j-/view
  - US Copyright class action against OpenAI - Tremblay P. and Awad M. v. OpenAI INC. et al, No. 3:23-cv-03223
  - https://www.ropesgray.com/en/sites/artificial-intelligence-court-order-tracker

Questions & Contacts

Rohan Massey

Partner
Ropes & Gray
London
Rohan.massey@ropesgray.com

Ellen Gilley

R&D Legal – Data & Technology
Takeda Pharmaceuticals
Boston
ellen.gilley@takeda.com

Colin Rooney

Partner
Arthur Cox
Dublin
colin.rooney@arthurcox.com