2025

Modern Location Based Web Data Gathering

Astro Research

Snapshot of store-level data gathering / Internal research

explore more →

astro

Table of Contents

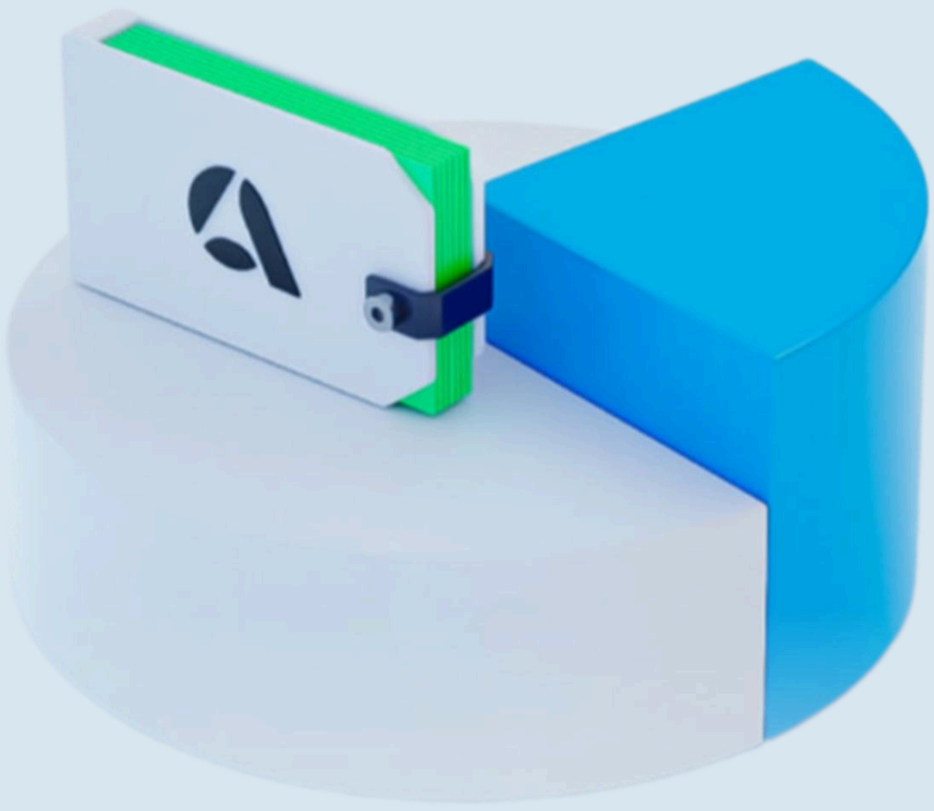
astro

01

This study is **Astro**'s practitioner-led deep dive into location-based web data gathering in 2025. Built on industry case work at massive scale and our own operational experience, it explains why per-store collection outperforms site-wide averages, how teams achieve complete coverage without overwhelming targets, and where AI-assisted monitoring and rigorous QA now sit in the pipeline. The goal is clear, source-anchored guidance you can apply to plan, run, and govern location-first programs with confidence — grounded in real systems, not theory.

What You'll Get

- **Fresh insights:** the state of the art in per-store collection (search vs. PDP specialization, store-switching flows), and today's reality of evolving access controls.
- **Practical patterns:** scheduling for freshness, geo-fidelity proxy strategy, modular configs, KPIs for monitoring / incident loops, and efficient pipelines (containers, incremental updates, columnar storage).
- **Ethics-first guidance:** strict KYC and AML policies to keep the network safe for all participants, with verified access, logged traffic, and continuous protection against misuse.



Astro Edge

Astro is an advanced ecosystem of proxies, we take an entire range of measures to provide for compliance and safety within our platform. We can identify instances of unethical usage through: We deliver ethically sourced, geo-targeted access (including **2M+ residential & mobile IPs across 100+ countries** with automatic rotation), KYC / AML controls, and AI / ML-based abuse monitoring — slotting into your schedulers and QA to raise success rates, preserve politeness, and keep regional compliance front and center.

What “Location Based” Means

Location–based web data gathering — collecting product content, price, and availability **per physical store** (e.g., switching site context to each of ~**4,500** local store pages), rather than treating a retailer as one monolithic source. This removes aggregation blur and reveals local realities that national averages hide.

Specific Use Cases

Price parity & compliance: Verify shelf–equivalent pricing and MAP adherence by store / region, not just at a national level.

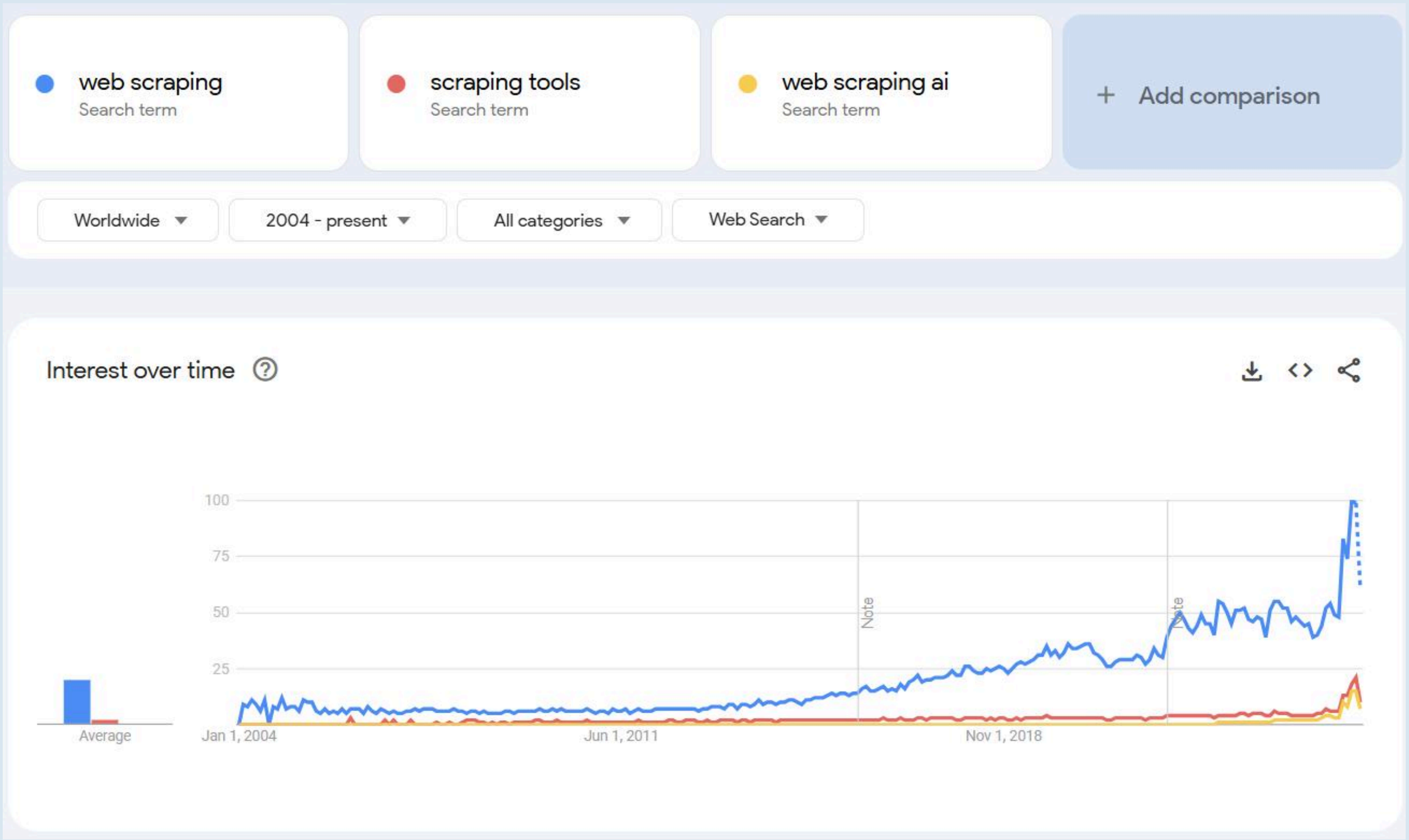
Availability & substitution: Detect OOS pockets, substitutions, and replenishment lag at specific stores to protect revenue and service levels.

Regional promo detection: Surface localized promos (bundles, BOGOs, seasonal campaigns) and measure competitor move market–by–market.

Astro is

1. **2 million+** ethically sourced residential and mobile IPs.
2. **100+ countries** with precise city and ISP targeting.
3. **99.9% uptime** across proxy pools.
4. Up to **250 concurrent connections per port** for high–throughput collection.
5. Automatic rotation modes (per request / per session / per time).
6. Built–in **AI / ML monitoring** and QA for geo–fidelity and access stability.

Rising Interest in Web Scraping Technologies



Global search trends confirm a steady growth of attention to web data collection — especially toward AI-enhanced and location-based scraping.

Astro continues to track how real-world data gathering evolves — and **builds the infrastructure behind it**.

1. **Sustained growth**: Searches for “web scraping” have increased over 4× since 2018 — signaling a mainstream shift from niche automation to strategic intelligence gathering.
2. **Emerging segments**: Terms like “web scraping AI” and “scraping tools” show the rise of automation frameworks powered by LLMs and smart proxies.
3. **Commercial adoption**: Companies increasingly rely on structured web data for price intelligence, product availability, and market mapping.
4. **Infrastructure maturity**: Modern proxy stacks and AI-driven orchestration make large-scale, compliant data gathering accessible to any team.

astro

02

Scope of Modern Operations

- **Enormous data volume**: Collecting every product from every store can yield billions of product entries per day. In practice, Fred de Villamil notes his team extracts data on billions of products daily while controlling costs and quality.
- **Goal: complete coverage**: The objective is near–100% of targeted SKUs in all target stores, refreshed frequently (daily or multiple times per day). This ensures analytics and decisions are based on up–to–date, all–included data without blind spots.

Specialized Crawlers for Different Pages

- **Divide and conquer**: Use separate crawlers for **search / category pages** versus **product detail pages (PDPs)**. Search crawlers issue keywords or browse categories across all stores (hundreds of queries per store, massive volume). Product crawlers simply fetch known item URLs.
- **Maintainability**: If the site changes (e.g., new search layout), only the search spider needs updates, leaving the product spider untouched. This modularity localizes impact and simplifies maintenance as sites evolve. It also prevents overloading a single spider with conflicting tasks.

astro

Web Scraping Innovations

Page Level

AI-based element detection, adaptive parsing, headless rendering, human-like interaction.

Session Level

Smart rotation, ban prediction.

Crawl & Dataset Level

Distributed orchestration, real-time monitoring, auto-scaling optimization, synthetic data enrichment.

Recent innovations in web scraping focus on three core layers of operation

At the **page level**, AI–driven parsing and rendering now allow crawlers to interpret modern, dynamic websites more like humans — understanding structure, detecting elements, and interacting with scripts in real time.

At the **session level**, smarter proxy and identity management help maintain stable access, reducing bans and improving reliability.

At the **crawl and dataset level**, orchestration frameworks coordinate thousands of jobs simultaneously, while built–in monitoring, scaling, and data enrichment ensure consistent output quality across global targets.

astro

03

Proxy Management & Compliance

Need for proxies
To simulate local users and avoid single–IP throttling, we use large proxy pools. Astro's network provides over 2 million IPs (residential, mobile, datacenter) in 100+ countries. Each request is routed through a rotating IP from **Astro's** geo-targeted pool, helping distribute traffic across regions and reduce repetitive patterns that websites could detect.

Proxy Types:

- **Datacenter proxies** — IPs from VDS / VPS providers and legal entities' IP addresses. High scores and high level of trust.
- **Residential proxies** — IPs from real home / office ISPs. They blend in better (as if real user traffic) but are slower / more costly.
- **Mobile proxies** — IPs from cellular networks. Highly trusted (sites often treat mobile traffic as distinct via carrier NAT), but expensive. We reserve these for the hardest targets.

astro

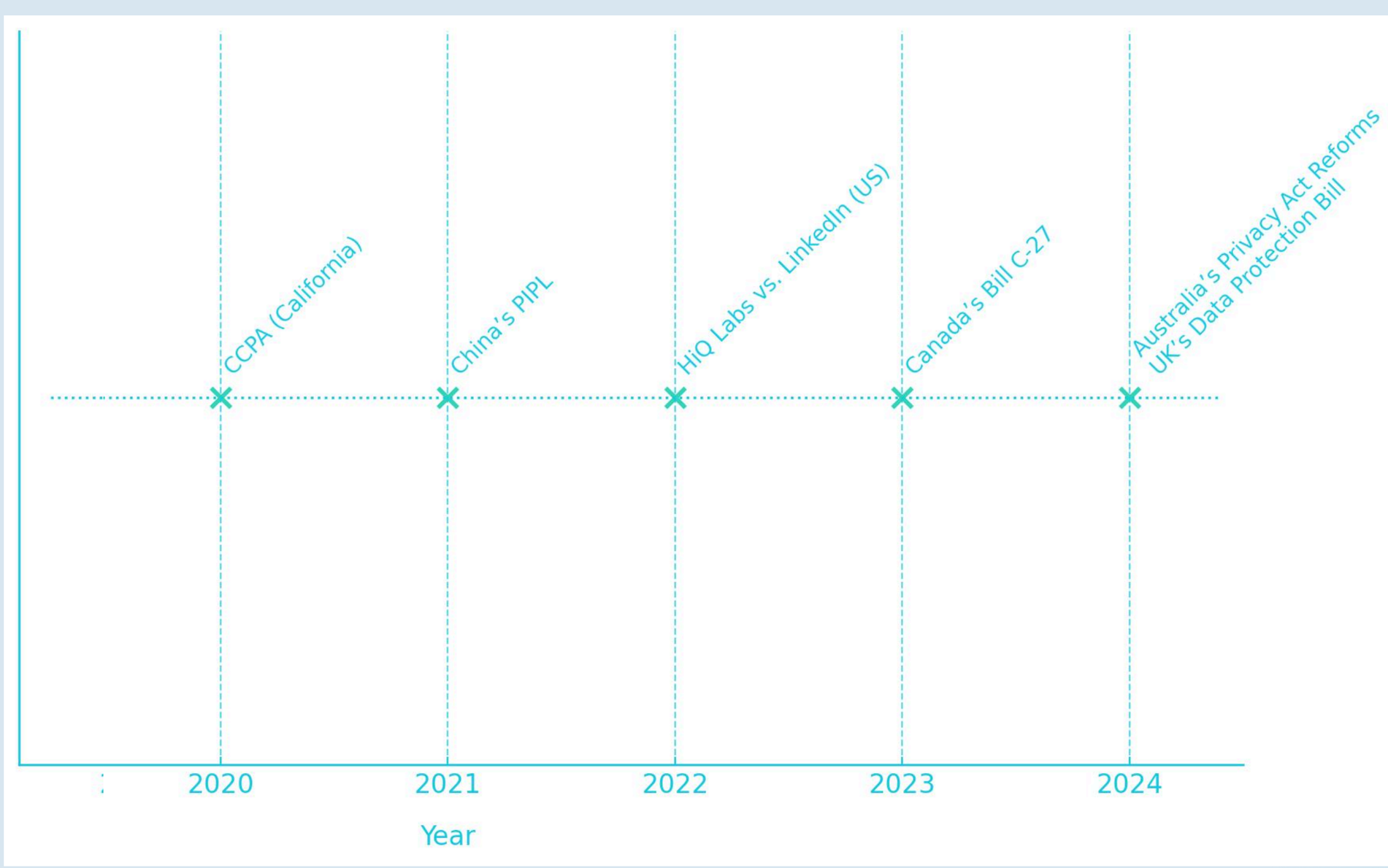
Regional & Ethical Complexities

- **Local laws**: Global web data collection must account for diverse privacy regulations. Frameworks like KYC and AML impose strict rules on handling personal information, so scrapers should be limited to publicly available product data. Any accidental exposure of PII (such as a user's name in a review) should be detected and removed. Systems should also support honoring data deletion requests if personal information is ever found.

- **Sensitive content**: Some product categories, such as alcohol or tobacco, are age-restricted. In these cases, a legitimate user would encounter an age-verification form. Ethical crawlers either simulate such verification with realistic inputs or skip restricted content altogether when compliance is uncertain.

- **Configuration by region**: Well-designed scraping frameworks allow rule adjustments per country or site type. For example, a configuration might specify: "if country=US and category=wine, simulate age verification; otherwise, continue normally." Schedulers can also shift activity to local off-peak hours (e.g., scraping APAC sites at night).

- **Respecting site policies**: Responsible data gathering involves considering website owners' preferences. If a site explicitly prohibits scraping — whether in its terms of service or via robots.txt — operators should assess the legal and reputational risk before proceeding. Many choose to focus on open sites or obtain explicit permission for sensitive targets.



Key global milestones shaping the legal landscape of web scraping.

astro

04

Key Scaling Challenges

1 **Complete coverage**: To truly hit every store / SKU, crawlers must adapt to constant site changes. Layout or URL shifts can break parsers, so configuration management (versioned rules, quick patches) is vital. The aim is never to miss a store or product. Even a small parser bug in one region could hide a significant issue. Continuous parser QA and fast updates are mandatory.

2 **High–volume polite crawling**: Hundreds of thousands of requests in short windows is routine. We must throttle carefully — e.g., strict concurrency limits, randomized delays — so your traffic looks human–like. Avoiding bans means crawling slowly enough to be respectful. (Unethical "spammy" crawling risks getting IPs harming the target site.)

3 **Resource & cost efficiency**: Thousands of processes and large proxy pools are expensive. Teams optimize at every level: using efficient async crawlers, consolidating incremental updates (rather than full refreshes when possible), and smart scheduling to eliminate idle effort. Cloud and proxy costs scale linearly, so wasteful retries or duplicate fetches must be minimized.

4 **Rapid incident response:** When failures happen (e.g., CAPTCHA, zero results), a well–built system immediately pauses the affected jobs to avoid further penalties. Engineers then diagnose: maybe a layout changed or proxy pool got tagged. Automated alerts (e.g., "30% drop in products scraped") trigger high–priority tickets. Fast triage and fixes (often within hours) are crucial to avoid extended data gaps.

astro

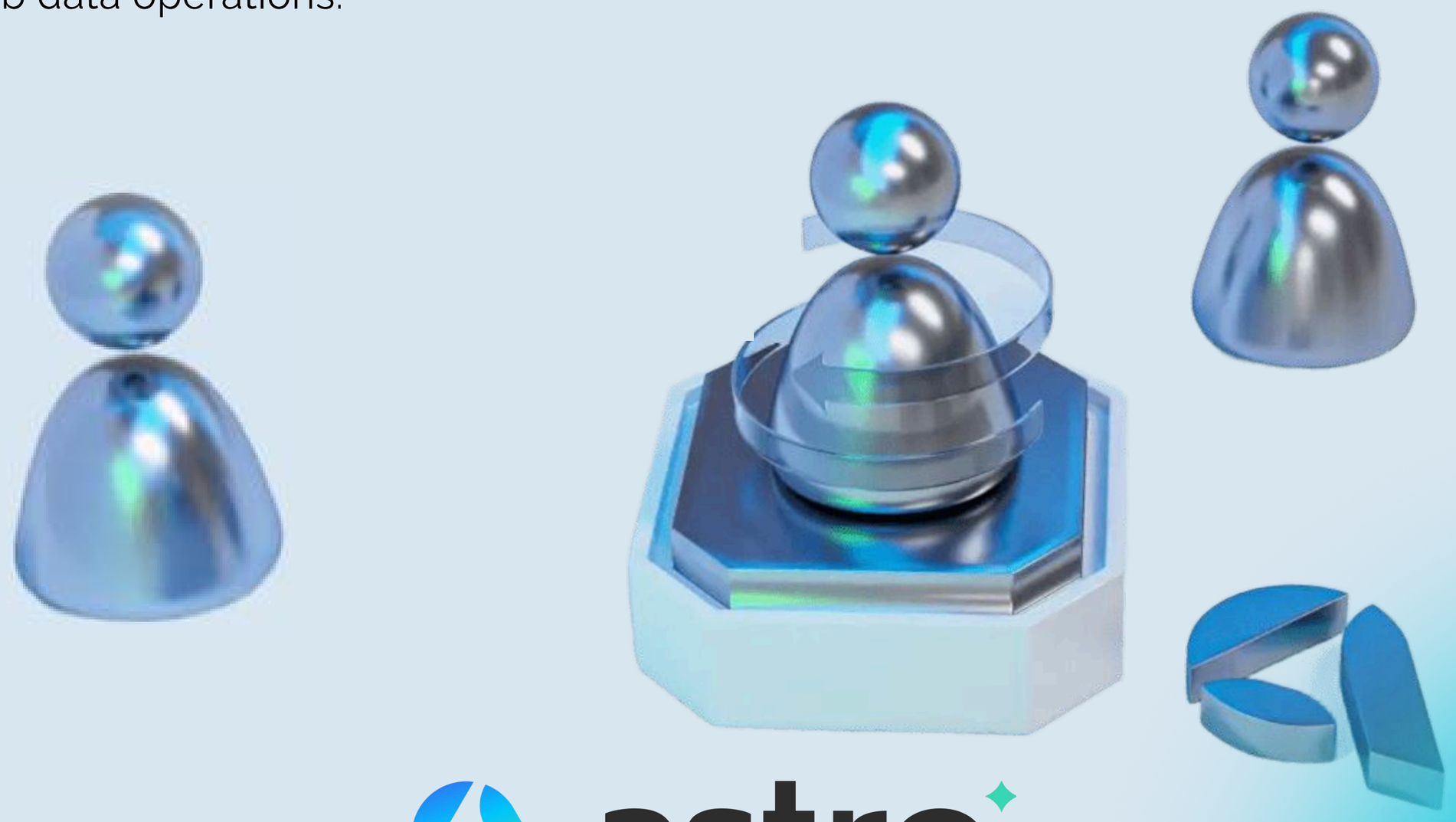
05

Anti–Scraping Evasion Techniques

Modern scraping frameworks are designed to adapt under restrictive environments. Instead of relying on static requests, they employ dynamic, browser–based systems that mirror real user behavior — rendering pages, executing scripts, and adjusting headers on the fly.

Advanced setups simulate human–like sessions through scrolling, timing, and locale matching, while centralized libraries distribute new bypass methods across multiple crawlers.

Continuous R&D keeps these systems resilient against changing detection patterns, making innovation and experimentation a permanent part of large–scale web data operations.



astro

KPI	Definition	Target (typical)	How to Measure	Notes
Store coverage	Share of target stores with fresh data in the last refresh window.	≥ 98% of target stores (daily); ≥ 95% hourly for priority categories.	Distinct stores scraped / target stores within SLA window.	Key success metric for location–based programs.
SKU coverage per store	Share of targeted SKUs captured per store.	≥ 95–99% per store.	Match scraped SKUs to master catalog per store.	Use per–store SKU lists to avoid false negatives.
Refresh latency	Time since last successful scrape for each store / SKU.	Median ≤ 24h; ≤ 4–6h for fast–moving categories.	Compute median & p95 latency per store / category.	Tie SLA to business use (pricing vs assortment).
Geo-fidelity	Requests originate from the intended geography (country / ISP).	100% country; ≥ 95% city where required.	IP geo / ASN checks per request or per session.	Beware CDN edge geo–mismatch; sample–verify.
Price accuracy	Match of captured price to ground truth.	≥ 98% on audit samples.	Manual spot checks or API receipts vs scraped price.	Store screenshots / HTML for evidence on disputes.
Stock accuracy	Correct detection of in–stock, OOS, or substituted.	≥ 97–99% on samples.	Cross-check with site / app at audit time.	Track false OOS and late replenishment labeling.
Promo capture rate	Share of localized promos / bundles captured.	≥ 95% of active promos.	Compare to weekly circulars or promo feeds.	Critical for regional marketing analysis.
Substitution detection lead time	Time from site update to detection of a substitution.	≤ 6h (priority); ≤ 24h (standard).	Measure time delta between ground truth and scrape.	Impacts shelf availability KPIs.

astro

References

“The 2025 Web Scraping Industry Report - For Industry Players ”, Zyte (2025).

“Location-Based Data”, ScrapeHero (2025).

“State of web scraping report 2025 ”, Apify Blog (2025).

“What Is a Geo Targeted Proxy?”, Astro Blog (2022).

“AI and data collection predictions to watch for in 2025”, Astro Blog (2025).

“How to Scrape Store Locations & Inventory Data from Retailers”, Unwrangle Blog (2025).

“Building a Web Scraping Solution to Track US Retail Market Trends”, Actowiz Solutions (2024).