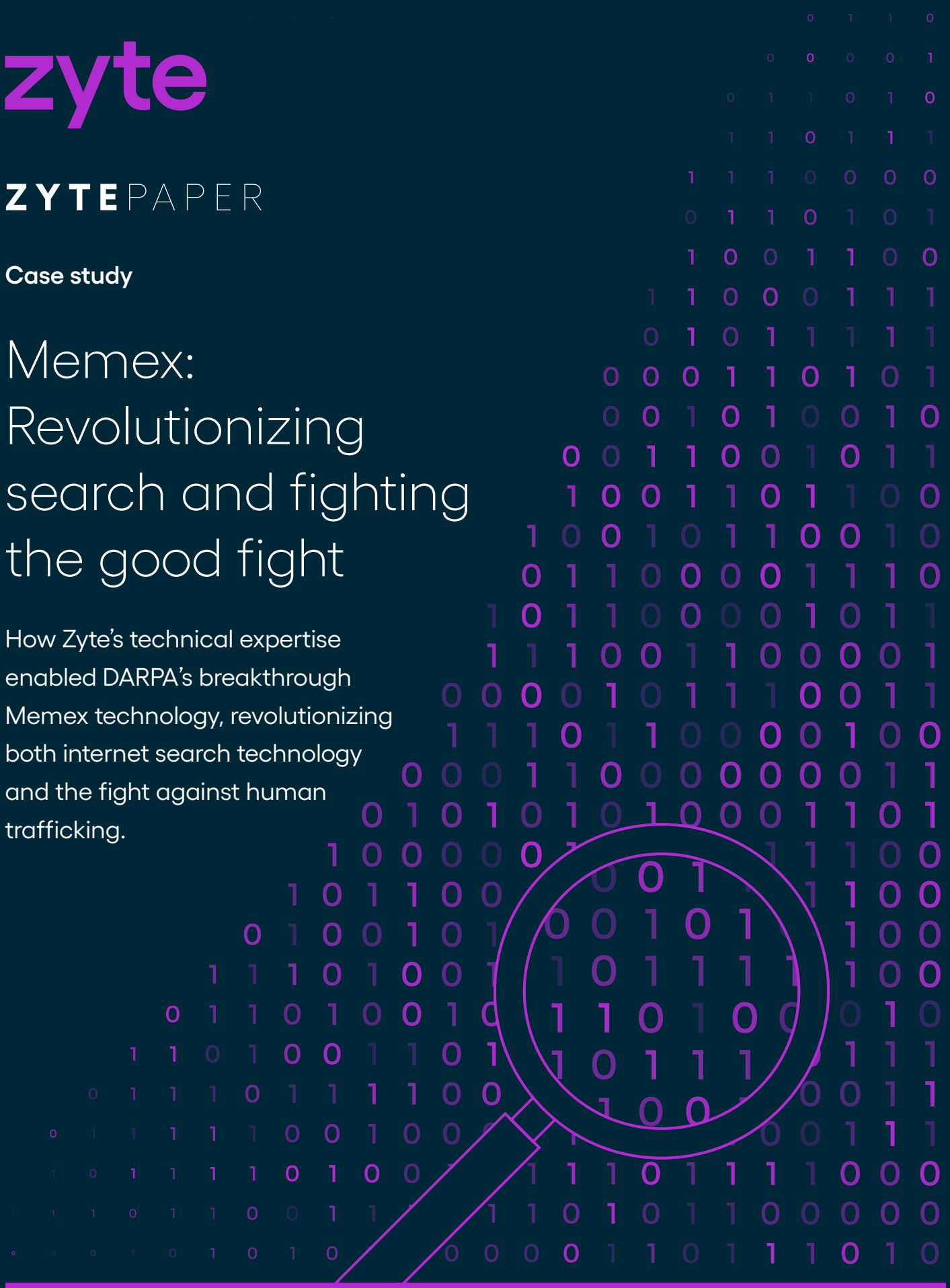
zyte

ZYTEPAPER

Case study

Memex: Revolutionizing search and fighting the good fight

How Zyte's technical expertise enabled DARPA's breakthrough Memex technology, revolutionizing both internet search technology and the fight against human trafficking.



What is Memex?

Imagine the internet as an ever-growing sea, ceaselessly expanding in all directions. The surface of this great sea is what most consider "the internet" - publicly indexed websites, accessible through mainstream commercial search engines.

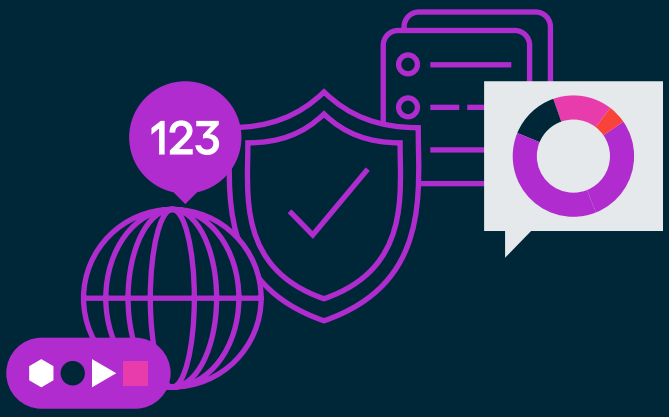
The surface of such a body alone is vast - but its depths are abyssal. Thus, today's internet, as we know it, represents only a sliver of a fraction of what's really out there - 400 to 500 times more content than that which is readily available.

The manner in which we search this vast body of information has, until very recently, been extremely manual, based on narrow queries, and tedious for most non-commercial use cases. For intelligence agencies and business research firms alike, data acquisition and analysis has been costly in terms of both resources and time spent. Further, the scope of such manual searches only touched on the very surface of the internet, wherein important information is often obfuscated.

Memex was designed to overcome this challenge: to make the invisible visible, and most importantly, to make it useful.

The project was initiated by DARPA, the Defense Advanced Research Projects Agency of the U.S. Military, which began the Memex project in 2014. In collaboration with a team of Zyte engineers and experts from around the globe, the Memex project led to advancements in state-of-the-art, paradigm-shifting technology for crawling the web - even its deeper and darker aspects

This new paradigm for search and indexing is radically different than the widespread activity undertaken by both commercial and state actors. Instead, Memex is a domain-specific search technology, one which intelligently discovers relevant content, including nontraditional content and that which is un-indexed, and collates the information in a manner truly useful to many use cases hitherto unsatisfied by previous means.



Memex, while extremely useful in research (both academic and commercial,) is critical for several Defense Department missions, specifically, the fight against human trafficking. Trafficking investigations utilize resources and personnel from across military, law enforcement, and intelligence agencies, and with Memex, previously hidden domains can be collected, curated, and made accessible across agencies. Today, this next-generation search technology has exponentially increased the efficacy and efficiency of a great deal of investigations.

We're proud to have contributed to the R&D and continuous maintenance of this project and believe it's a powerful example of web scraping as a force for good.

DARPA, Zyte and Hyperion Gray

2014

DARPA issues Broad Agency Announcement to solicit proposals for the Memexb program, and hosts Proposers Meeting in Arlington, Virginia

Zyte and Hyperion Gray join forces, and our joint proposal is accepted. A team of engineers from both parties join a broader collaboration with scientists from NASA's JPL, the Tor Project, and MIT's Lincoln Laboratory.

2017

Three year grant is now complete, and the project's requirements fulfilled

DARPA's new Memex technology is now enabling new use cases across law enforcement and research of every kind.

Today

Zyte engineers continue to work on the project

continually updating and ensuring a smooth crawling operation for the technology and enabling the production of further innovations. A huge wealth of technology created during the Memex project is now open sourced.

Projects requirements

Flexibility and expertise

To undergo such experimental, state-of-the-art projects, it's critical that the team involved is agile and capable of quickly changing course between complex topics.

A competitive spirit

Given the nature of the contract and the wide array of specialists and experts involved, an important aspect of the validation process were friendly, albeit intense competitions.

Technical area expertise

The project was split into three Technical Areas (TAs):

01. Domain-specific indexing

02. Domain-specific search

03. Applications

As we pursued the project jointly, Zyte engineers were focused on the first two Technical Areas, while our co-contractors developed the third.

TA1: Domain-specific indexing

This project called for the creation of a domain-specific indexing software, featuring a highly scalable web crawling infrastructure for content discovery and information extraction. It must be capable of handling a broad variety of data formats and adapting to an array of situational challenges which would prove challenging or impossible for existing technologies.

A requirement for this Technical Area is the capability to develop sophisticated automated and semi-automated crawlers, capable of uncovering obfuscated links, exploring deep web and "dark web" content as well as hidden services.

Furthermore, part of data extraction in this technical area requires expertise in developing systems for:

- The normalization and organization of heterogeneous data
- Natural language processing for translation and entity extraction (disambiguation and coreference resolution)
- Image analysis (object recognition)
- Relevance determinacy

TA2: Domain-specific search

Furthermore, the project required the creation of a configurable, domain-specific interface for web content. This interface would include conceptually aggregated results, linking, for example, an individual under investigation to a wide variety of related content. This content, within the interface, would be conceptually linked to other content and task relevant facets (such as entity movement, location, and content association.)

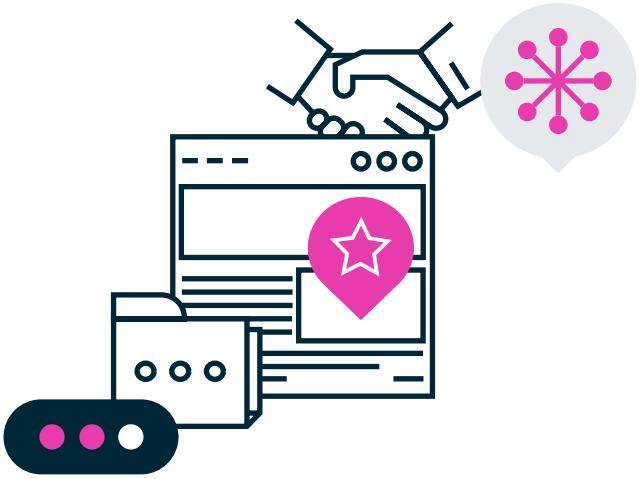
TA3: Applications

This technical area is key for supporting performers of TA1+2, through the development of system-level applications and through evaluation and cooperation with end-users.

Zyte's contribution

Our proposal chiefly stated our interest in creating a system capable of rendering a web browser within itself, emulating complex human behavior, and generating records of a page's changes over time - along with a clear record of how its data changes as well.

So we did just that. In addition to the wealth of open source technology produced by the team working on the Memex program, our scriptable, headless browser known as Scrapy was born.



The deep web is anything but shallow, and given the immense amount of complex, heterogeneous information it bears, keeping a historical record of data acquired - and how it changes - was critical. To do this, Zyte engineers collaborated with Hyperion Gray and others in the program to develop exciting new advancements in machine learning. "Eli5," for example, is a library for debugging and inspecting machine learning classifiers and for clarifying their predictions - an extremely useful tool when working with such a large amount of complex data.

Web scraping for good

While tech-savvy businesses use these technologies to extract profit-driving insights, we were thrilled that our web scraping expertise could be put towards technologies capable of aiding one chief purpose of the Memex program: fighting human trafficking. At any one time, 2-5 Zyte engineers developed the cutting-edge scraping and crawling capacities now used by law enforcement all across the U.S.

Sites bearing the most insightful data are prone to changing frequently and concealing their services, making tracking and generating data from these sources difficult. By developing Splash and working to provide our web crawling and scraping expertise, the Memex program is capable of tracking, mapping, and monitoring even the most mercurial of online data sources.

In written testimony to the U.S. House of Representatives, NYC District Attorney Cyrus Vance Jr. stated that Memex is used for screening over 6,000 arrests for signs of human trafficking - and that's only in Memex's "first year" - 2017. The Human Trafficking Response Unit has also reported that prostitution arrests screened with data generated by Memex tools has increased indicators for human trafficking from 5 to 62 percent - a huge step towards identifying victims of trafficking.

The technology and today

Working alongside engineers and scientists from Hyperion Gray, NASA's JPL, MIT's Lincoln Lab and others, Zyte is proud of our contribution to this important technology and what it represents. Today, we continue our work with the Memex program, and continue to maintain and update an undisclosed number of spiders for the project. If a technology was needed but didn't exist, we developed it for use.

From day one of the contract, we committed ourselves towards producing the highest quality software, to achieve real advancements in web scraping, and to open source as much of it possible. We're glad to see this goal become a reality, and such an open-ended yet impactful R&D project was a perfect fit for our flexible, international team.

Why did DARPA and Hyperion Gray choose to work with Zyte?

When Hyperion Gray reached out to our leadership to propose applying to the program, we were eager to co-draft an application.

Hyperion Gray is a renowned team of data scientists and software engineers, extremely capable in their own right. But why did they reach out to us, and why did DARPA choose us as finalists?

Industry-leading expertise

Since our formation in 2007 and the launch of Scrapy, the first open source Python web crawling and scraping framework, we have been providing the highest quality web data extraction services for some of the world's largest and most innovative companies.

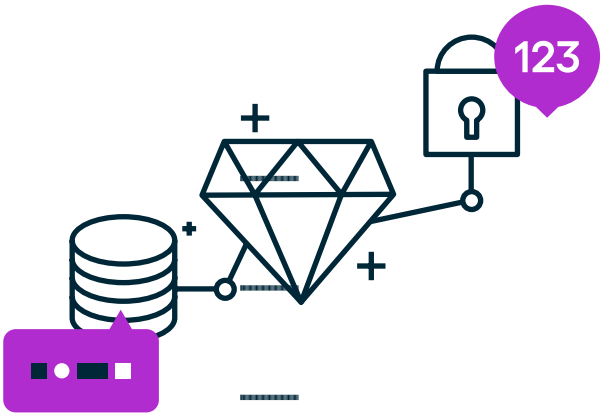
Data quality and complimentary skills

The Memex program is anything but simple, and some of the more challenging aspects of scraping the deep web and the broader web at large were unique hurdles requiring the most sophisticated scraping technology.

Hyperion Gray and Zyte's competencies in data analysis and in developing crawlers and scrapers enabled exciting inter-team insights into the biggest challenges facing web scraping.

No vendor lock-ins

At Zyte, we're phobic of the unfortunately popular practice among today's software companies of locking customers into dependency on proprietary services. Working with Zyte, no matter the size of the client, means working with highly portable, almost entirely open source technology that we've developed - freely available online.

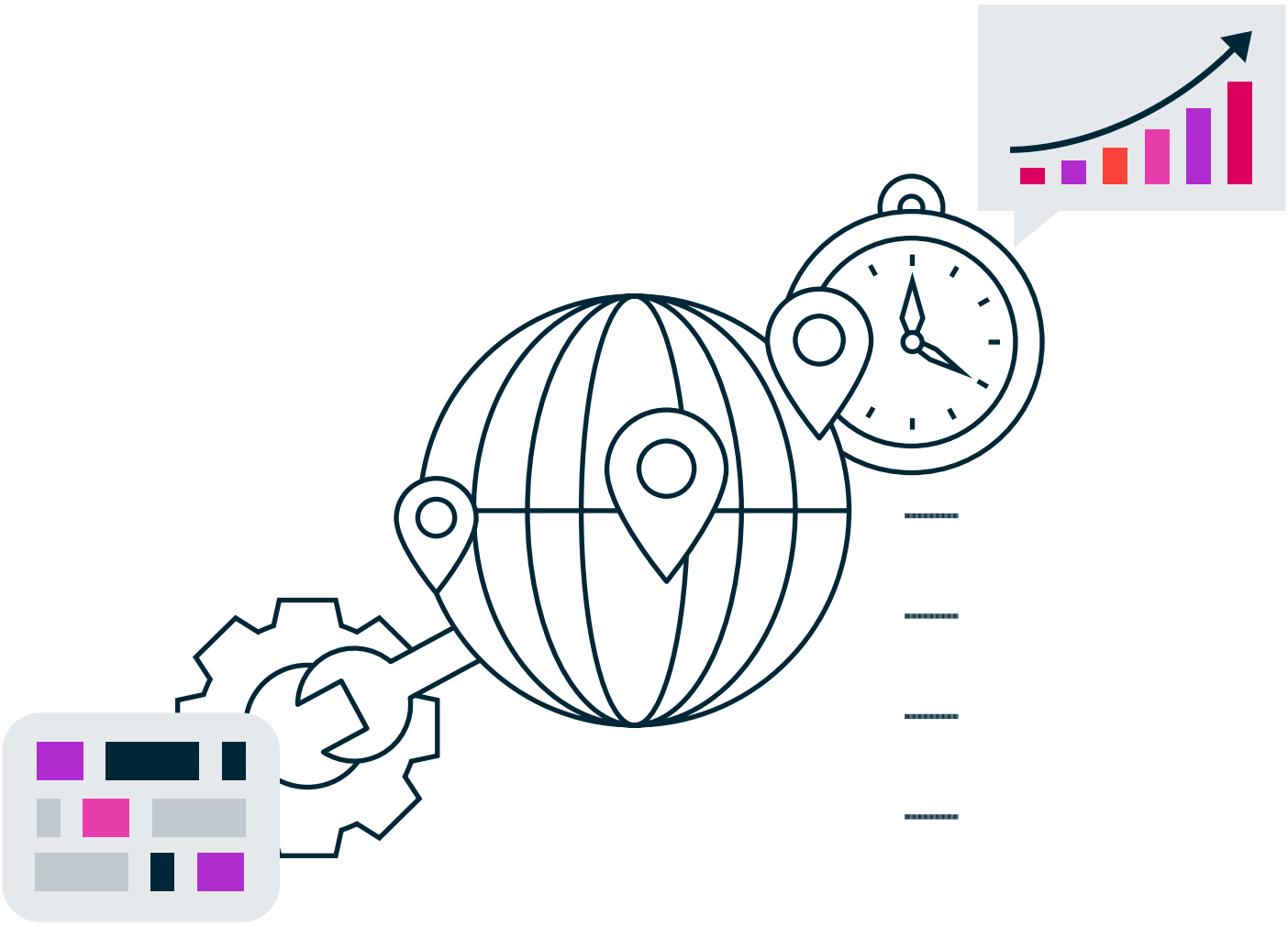


When providing the infrastructure for commercial (or, in this case, military-grade) scraping projects, we're proud to work with clients in a professional, unobtrusive manner, free of obnoxious lock-ins and claustrophobic contracts.

Anytime, anywhere

While we're based out of Ireland, our company truly lives in the hearts of our skilled engineers, working out of locations all across the world. No matter the time of day in Arlington, Virginia, somewhere on Earth, a talented Zyte engineer is available to answer a last-minute question or solve for an unanticipated change of plans.

As Zyte and Hyperion Gray worked together towards the completion of DARPA's Memex program, this flexibility enabled us to participate in the program's competitions and exciting, fast-paced innovation sprints. This geographic non-dependence is also why we're the world's #1 web scraping company, as our turnaround time on spider updates and fixes is second-to-none.



How Memex is revolutionizing law enforcement and ‚search' iitself

Human and labor trafficking

Instead of a Google or Bing search, which yield only a concrete amount of sites with textually-relevant information, Memex can search and aggregate useful information within any particular domain. In this case, within the conceptual domain of human trafficking, prostitution advertisements and escort service listings can be intelligently and automatically searched at once, generating a huge amount of data.

Most importantly, Memex explores hyperlinks and collects useful information. Does a perpetrator hide important information in a photo, to hide it from text-based search? Memex handles that. By linking records and producing concrete intelligence, then displaying both the geographical, historical, and conceptual insights, the work of law enforcement specialists is enhanced to an extraordinary extent.

With this automated suite of analytical tools, each analyst can have a vastly expanded caseload and potentially save many more lives.

Arms trafficking

The illicit arms trade is worth billions online, and the means through which Memex can enable law enforcement to shut these rings down is extremely effective. The historical and geographical data of listings found online, collected and aggregated along with conceptually linked attribute data such as phone numbers and sales trends enable deep analysis and empowered, well-equipped investigations.

Child exploitation

Memex is effectively a suite of super-search engine tools, and given the extremely dark nature of this kind of crime, its tools can be customized to spare law enforcement of extremely harrowing, tedious manual investigation.

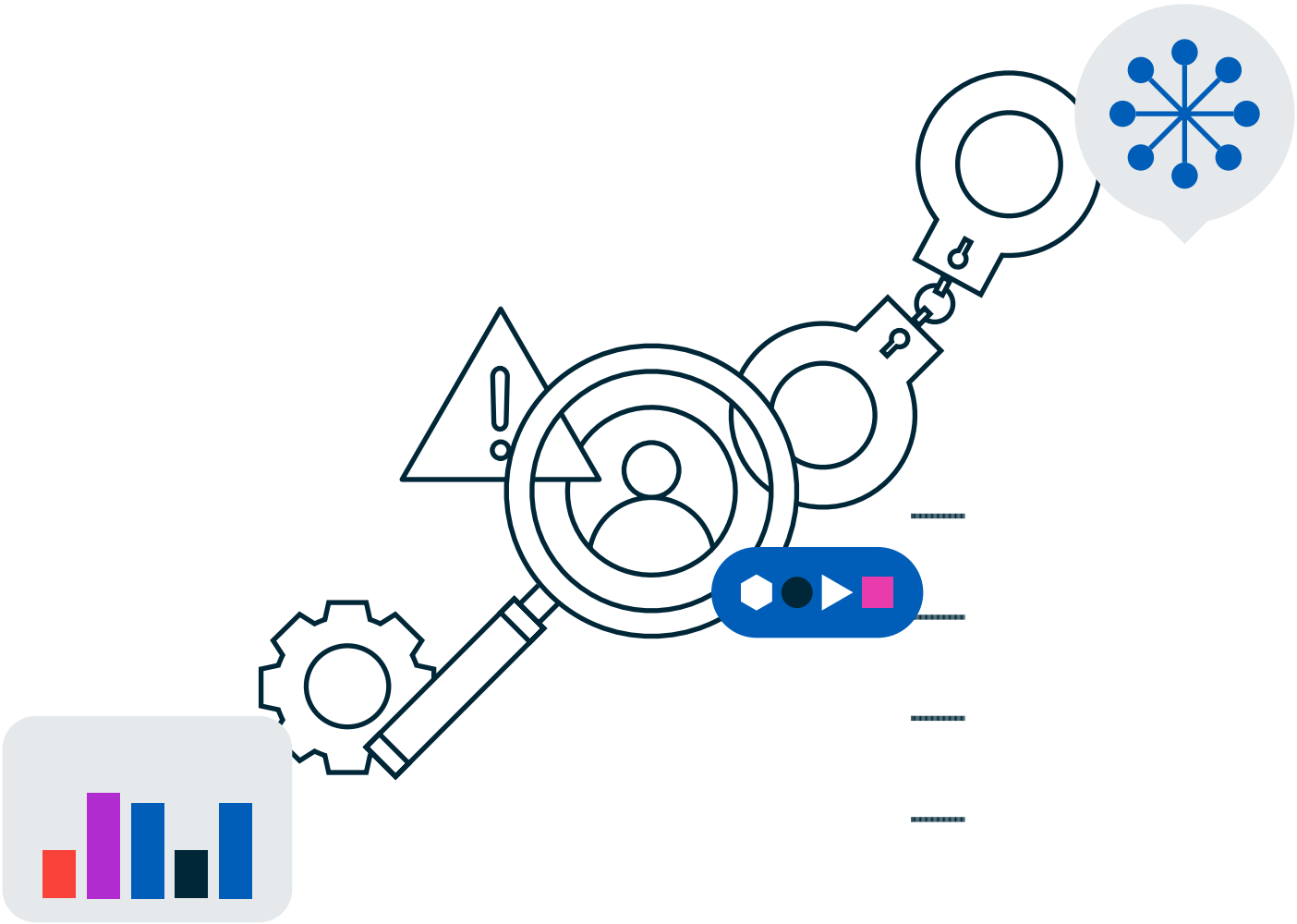
Memex can automatically and safely scrape data from sites in the dangerous sphere of child exploitation, then store and analyze the intelligence in an effective manner according to pertinent laws.

Illegal pharmaceuticals

Pharma trafficking, like counterfeit electronics and sites that peddle stolen financial information, often involves the obfuscation of normal text and the hiding of services. Memex is capable of rapidly detecting and overcoming such obstacles in a manner outright impossible to previous technology - and at scale.

Search itself

Let's say you're a material scientist interested in the development of a specific application of graphene. With Memex, you can track and analyze the progress of such technology through accessing a totalizing record of related data - geographic, scientific, textual, and nontraditional in scope. The commercial and scientific applications are virtually endless.



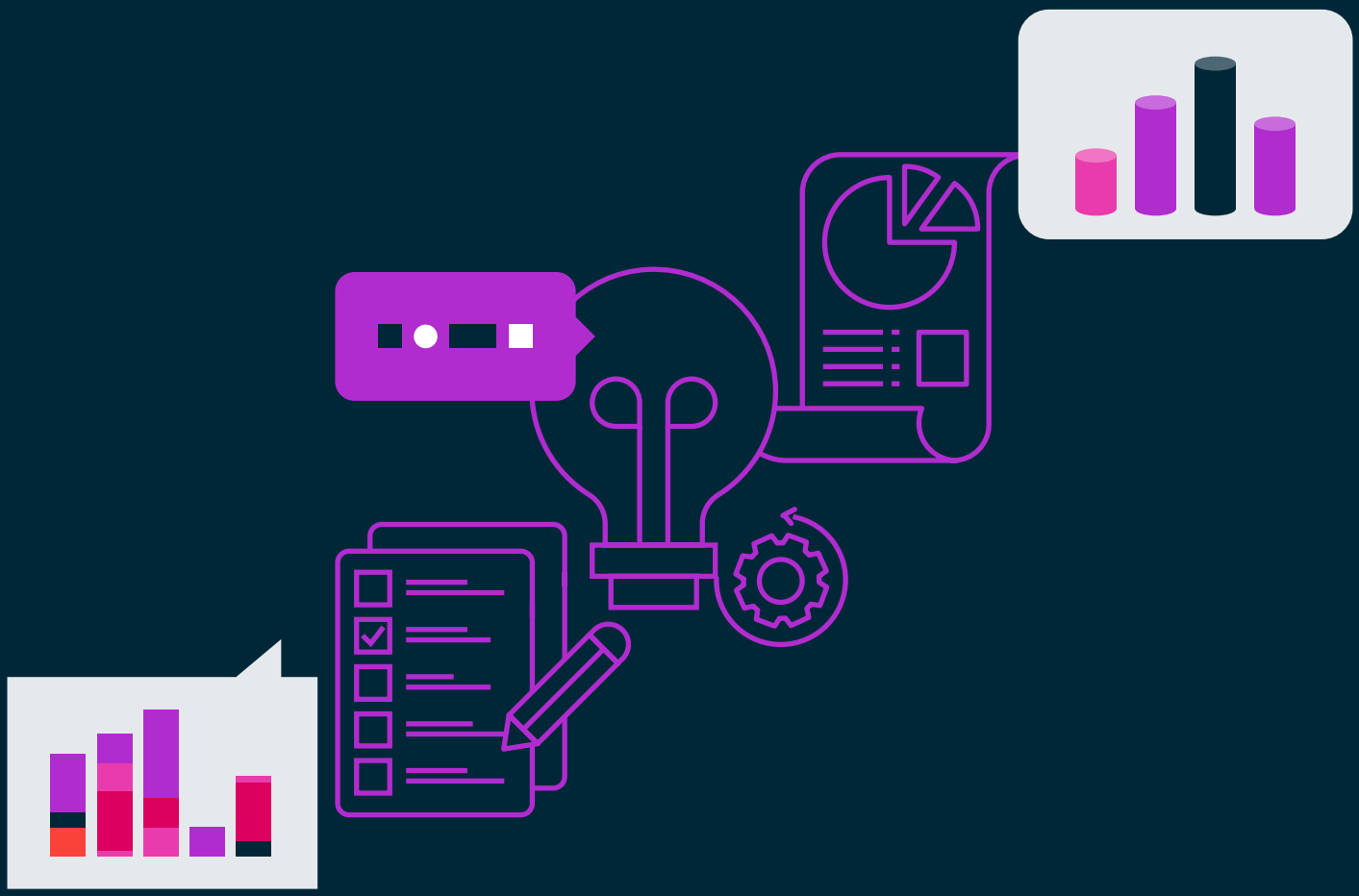
Conclusion

The technologies discussed in this document are truly shifting today's search and discovery paradigm, and are only born of long-term, nontraditional research and development cycles.

Projects such as the Memex program are an exciting opportunity for us to continue our work into advancing and overcoming what was previously thought possible when it comes to web data extraction.

R&D is in our blood, and the technologies we continue to develop serve as the backbone for a majority of today's scraping efforts across the globe. By partnering with companies and organizations interested in creating effective new technologies, we aim to build a better world.

If this case study inspired you or you're curious about how our experts might contribute to your project, request a free consultation today.



At Zyte we turn websites into data with industry leading technology and services.

Our solutions include:

- **Data Extraction Service**
  Let our web scraping experts build and manage the bespoke data extraction solution for your business needs.
- **Smart Proxy Manager (formerly Crawlera)**
  Forget about proxy lists. We manage hundreds of thousands of proxies, so you don't have to.
- **Automatic Extraction powered by AI**
  Instantly access accurate web data through our user-friendly interface or various Extraction APIs and save time getting the data you need.
- **Data extraction platform**
  Access developer tools, data extraction APIs and documentation, built and maintained by our world-leading team of over 100 extraction experts.

zyte

It's yours. The web data you need.

Access clean, valuable data with web scraping services that drive your business forward.

Talk to us

www.zyte.com

Cuil Greine House, Ballincollig Commercial Park Link Road, Ballincollig / Co. Cork, Ireland