



How Glean Search Works

Providing better results through
innovative machine learning
and indexing

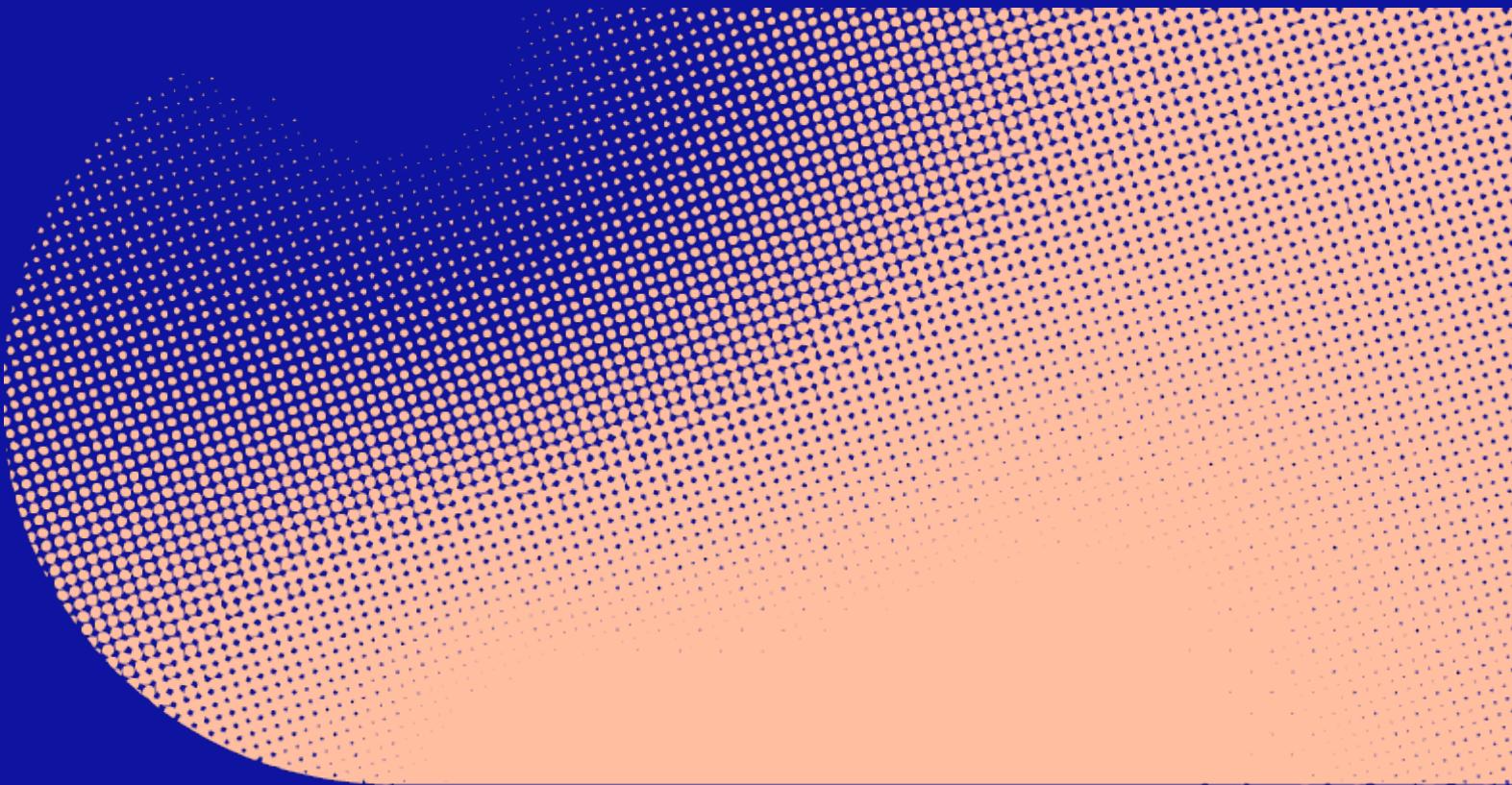
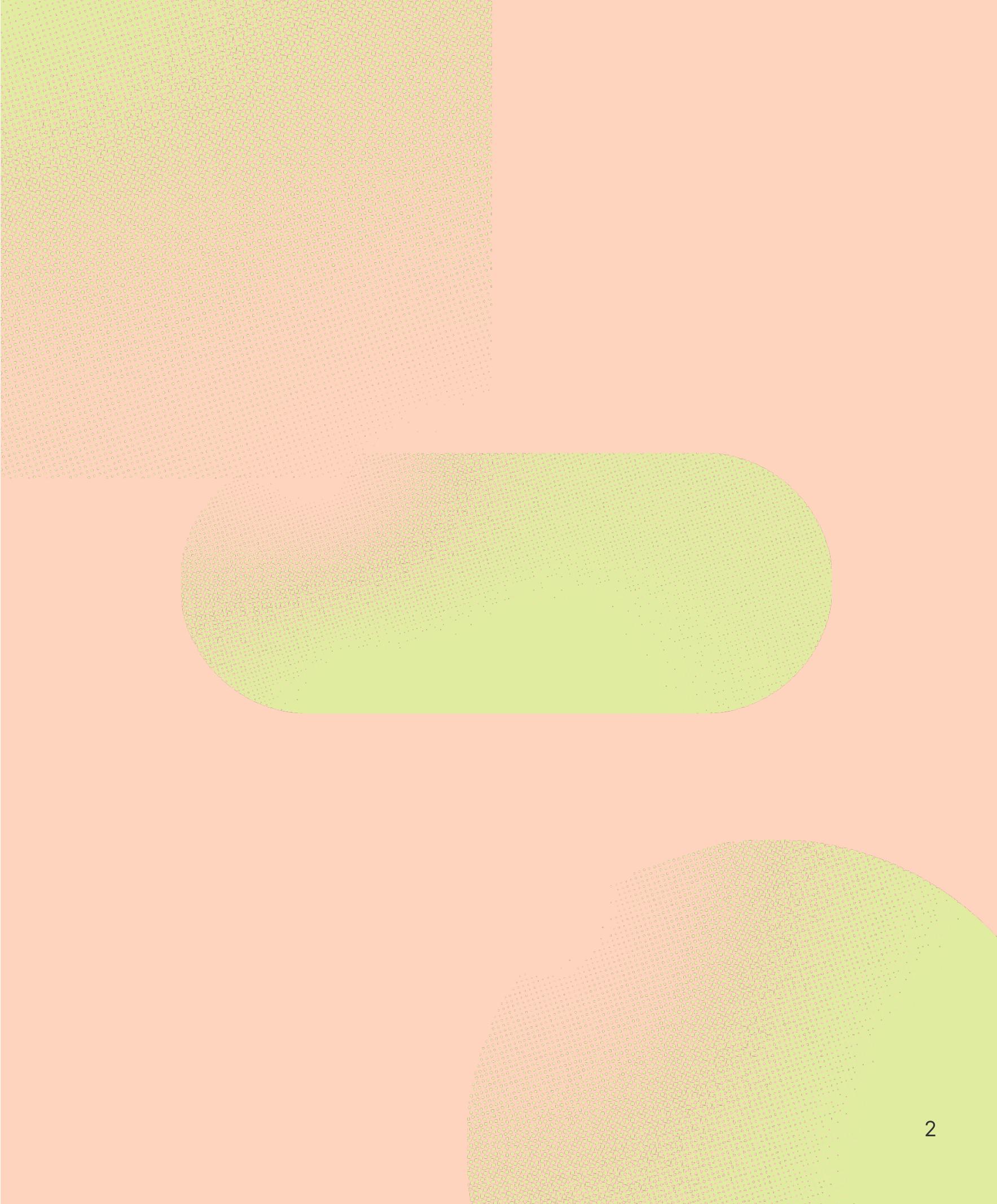


Table of contents

The state of search	3
Glean search architecture	4
Unboxing and setup	6
Building the knowledge graph	9
The search journey	14



The state of search

Most of us began our digital journey through a humble search bar. Thankfully, search engines are still around today, and are better than ever before. However, most search solutions for business and SaaS applications, especially native solutions, fall alarmingly behind in functionality compared to the consumer options we use on a daily basis outside of work.

With the volume of information, data, and applications we use each day only growing, having a readily available and personalized solution for discovering existing knowledge is no longer a luxury – it's essential to enabling your business to work optimally without interruption, as well as onboarding new employees and tools as seamlessly as possible.

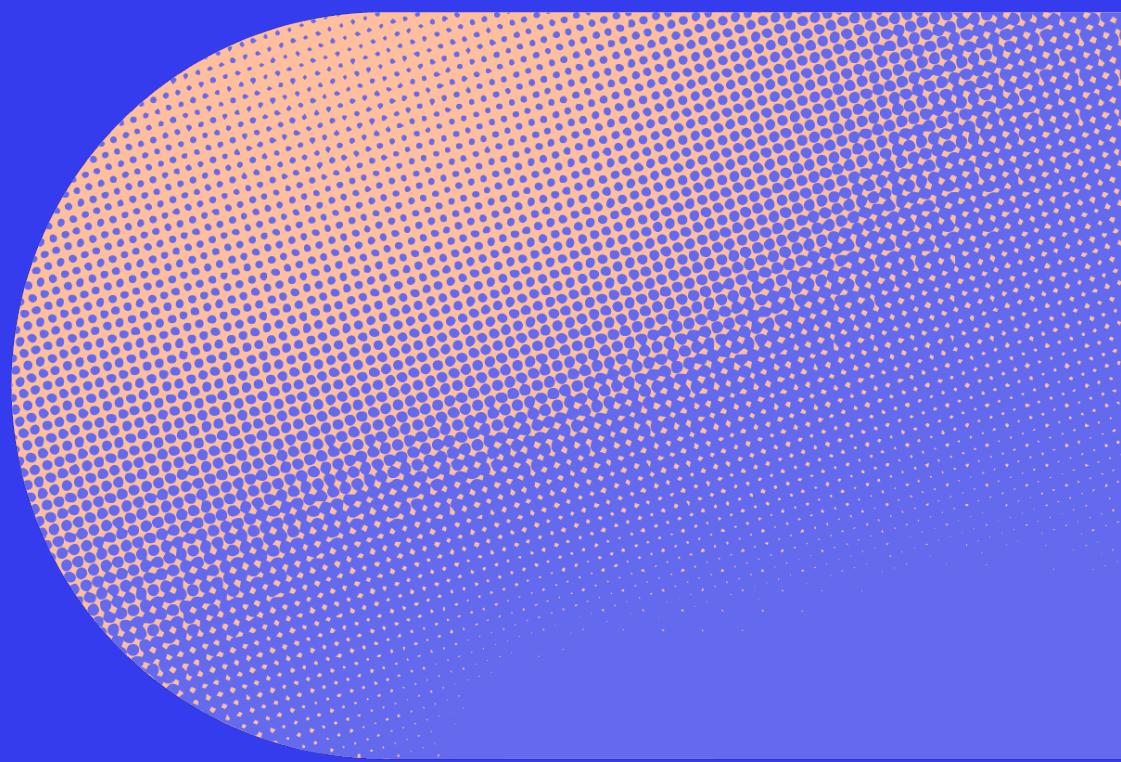
Glean is solving the problem of workplace search, once and for all. Glean is the enterprise search platform that helps every employee find the information that they need, instantly. It takes everyday queries at work and processes them in a way that produces accurate and personalized results according to the way you work, and the people you work with.

Our founders – including Arvind Jain, CEO of Glean, who came from over a decade at Google as a Distinguished Engineer – also recognized that a generic search tool would never be good enough for the job. It requires a uniquely personalized and constantly improving solution that helps search work the way you need it to. Our platform uses the latest advancements in machine learning to understand each business' unique way of working, and provide users with the most personalized results for their queries.

This eBook will guide you through that unique search process. It also includes details on how Glean builds its index on existing enterprise data, along with how it ensures the results you receive are exactly what you need to work without interruption or delay.

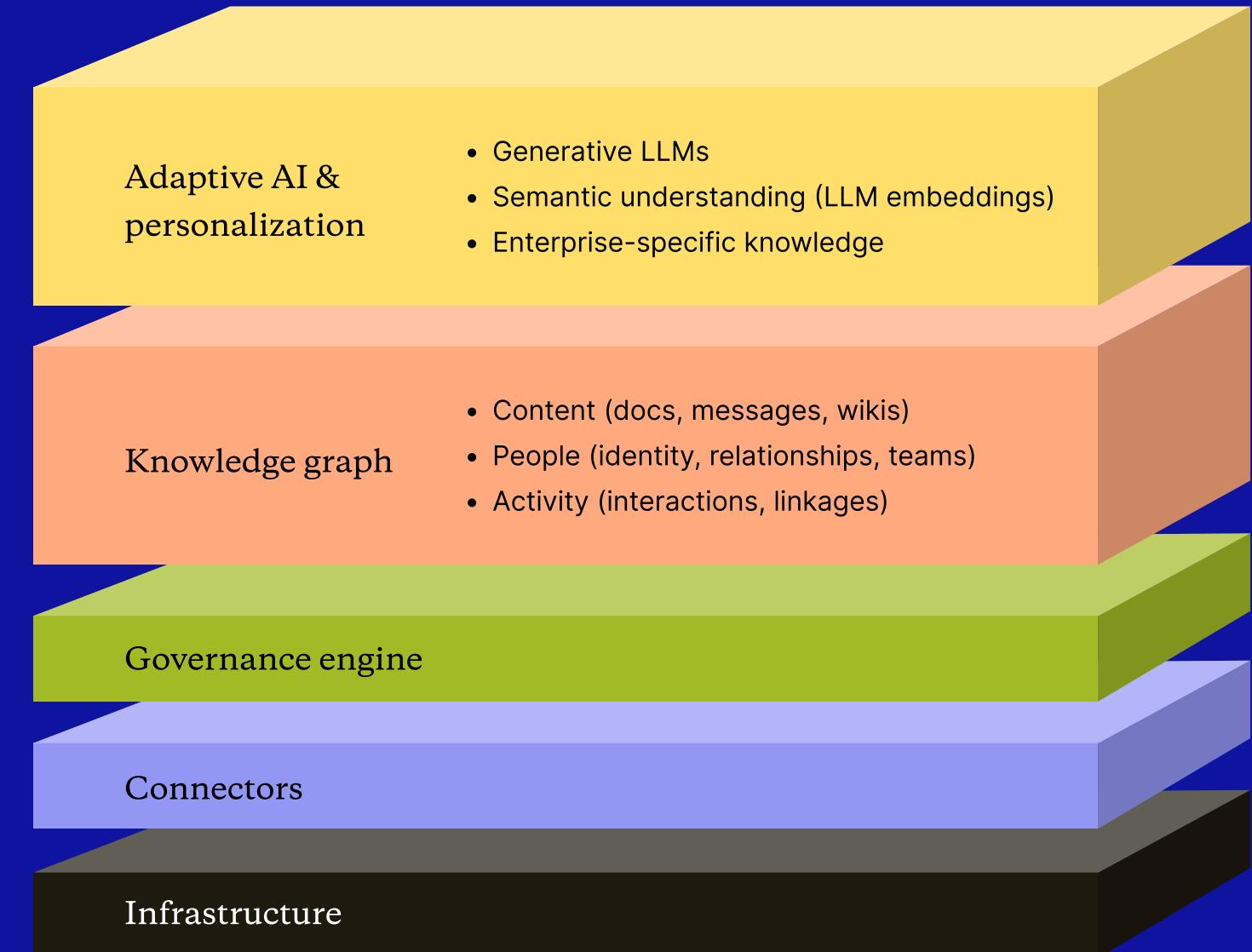
Let's get started!

Glean search architecture



At the highest level, Glean's **workplace search platform** delivers best-in-class relevance through a suite of powerful AI capabilities, indexing tools, and what we call the enterprise knowledge graph.

Highly customizable and compatible with today's most common workplace apps, Glean takes into account all the complexities and concerns of modern digital collaboration through powerful capabilities like generative AI, automation, and deep learning.



Unboxing and setup

Implementing a great search solution can be hard work. For most options out there, it involves:

- Manual search tuning
- Dedicated developer support to ensure proper ranking
- Laborious tweaking with permissions to ensure security

It's a lot of lift and commitment from the user end, both to get the system up and running as well as constantly iterating to get it to work the way you need it to.

Glean takes that hard work out of the equation. Glean's search solution is fully customizable, but is built to require minimal operational overhead to set up and needs no third-party engagements or professional services investments.

Other solutions



glean



- ✓ Quick setup
- ✓ No professional services
- ✓ No third party engagement
- ✓ Fully customizable

The deployment process begins through the creation of a new Project in Google Cloud Platform (GCP). Each organization retains their enterprise data within this exclusive cloud infrastructure. This single-tenant architecture model, which delegates a separate GCP project per customer, also guarantees complete and secure data isolation. Your enterprise information is exclusively used to train your own deployment's machine learning models – no one else's.

Glean also works with the principle of least privilege, and strictly enforces data access rules that exist at multiple levels.

Then, once Glean is given management access and is connected to specific data sources (Google Drive, Box, SFDC, etc.) at the customer's request, our software begins to take the necessary steps to construct the foundational enterprise knowledge graph that powers Glean's unique search system.

Levels of data access



Product

Users without access to a certain product (e.g. Salesforce, Google Drive, Figma) will not be shown any results from that data source.

Object

Users will not have access or visibility over individual files they don't have permissions for.

Record

Typically applicable for CRMs
Fine-grained permissions will be followed on specific objects with stricter rules than the entire object. Results will not be displayed to users without these fine-grained permissions.

Building the knowledge graph

Glean's unique search engine operates based on a real-time model of your enterprise's indexed information – **the enterprise knowledge graph**. This system involves three key pillars:

Content

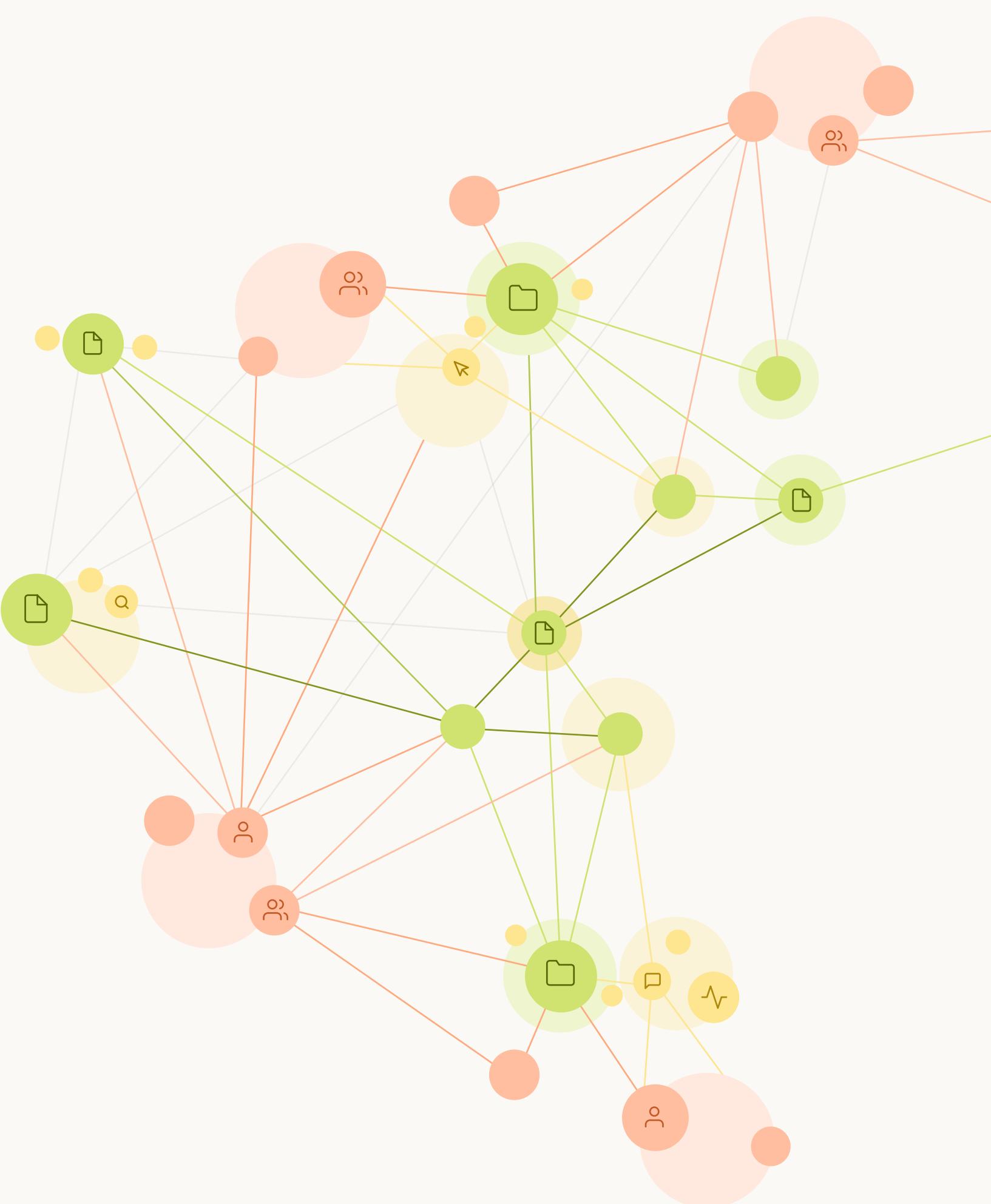
- Individual assets
- Documents
- Messages
- Tickets
- Entities

People

- Identities and roles
- Teams
- Departments
- Groups

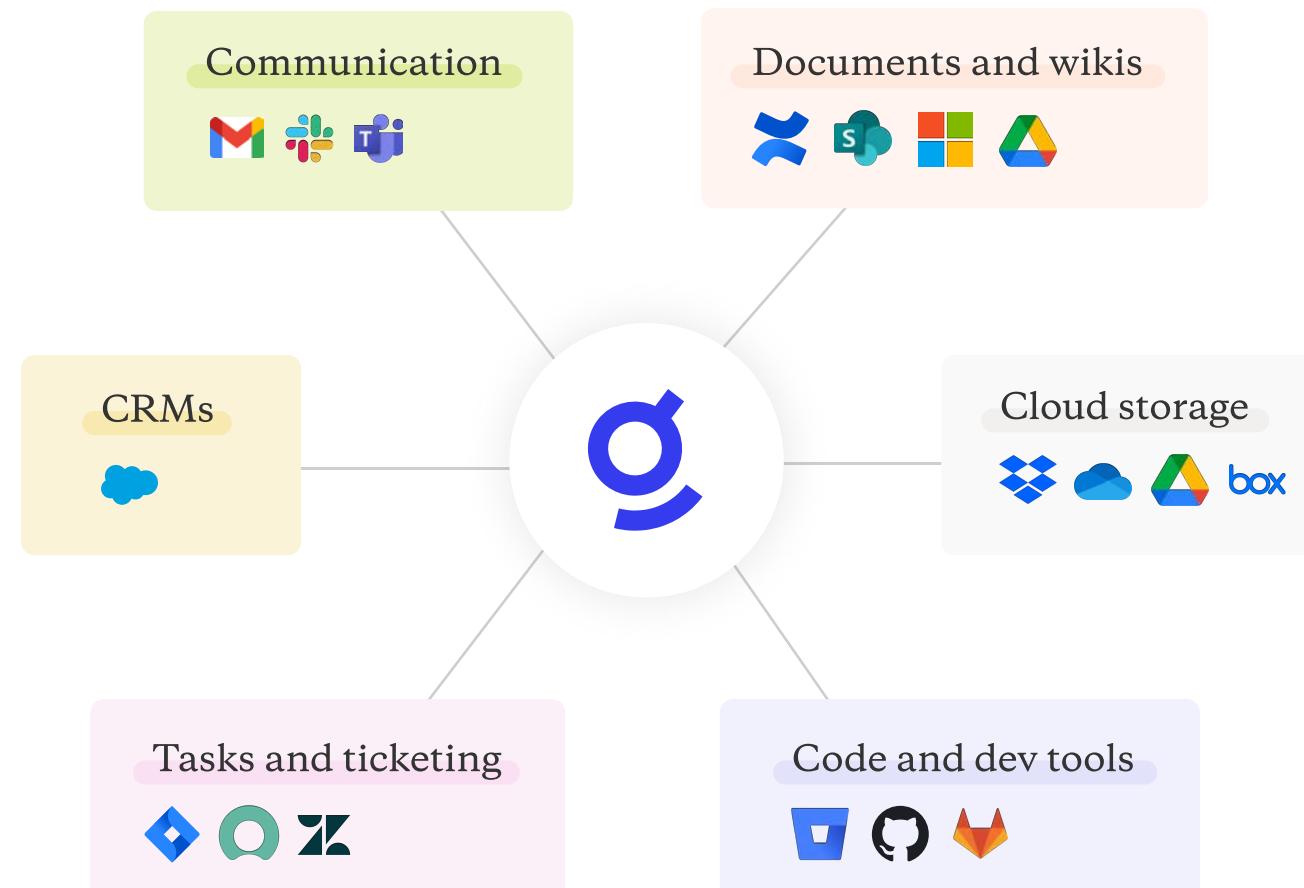
Activity

- Content creation
- Editing history
- Comments
- Searches
- Clicks



Content

Filling the knowledge graph with content begins with our easy-to-use 100+ connectors. Each connector is tailor-made for each application's unique data model and API endpoints, requires no additional professional services to use, and is fully permissions-aware – ensuring all accessibility and sharing protocols are strictly followed for each source.



Through these connectors, Glean's content crawler searches over every part of a piece of content, not just the title. This involves relevant item content (titles, body copy, comments, media, etc.), as well as metadata (file creator, time of creation, update history, file type, folder structure, etc). All of these elements, stored within the knowledge graph's index, become readily searchable. Customizable weights can also be set for specific categories to influence search results. Facet search results can be tuned based on metadata fields specific to each app.

Permissions can also be individually set for specific items. Files can be selectively included or excluded from Glean's crawl system by specifying asset IDs or by a broader categorization, such as item containers (folders or drives).

Content crawl strategy is also completely tunable according to your organization's preference. Adjust crawl cadence, assign blackout hours to avoid peak work hours, and shift between different crawl methodologies to ensure your knowledge graph contains the most relevant and up-to-date information throughout the day.

People

One out of ten enterprise searches are about people. It makes good sense – workers want to know exactly who they're working with, what their role is, and what they've most recently worked on. Glean facilitates this by aggregating data across multiple tools, providing a comprehensive and information-rich view of anyone in the company.

Our engine is also capable of providing deeply personalized and permissions-aware results, especially as it better understands each individual's role within an organization.

Glean builds each enterprise's knowledge graph with a deep understanding of the people within, such as what their role is, which team they're on, their tenure, and location. Our system is then capable of constructing a unified identity for each person across all apps, along with a holistic organization structure that understands everything from each person's closest collaborators to what projects they've most recently worked on.

The underlying data model and individual sources of information used to build these profiles can be customized according to preference.

Christine Martinez · She/Her
Principal Software Engineer, Engineering
San Francisco
Joined 3 yrs 9 mos ago · Tenure > 90%
christine@acme.com

You worked together on

- Project A
- Project B
- Project C
- Project D
- Project E

Activity

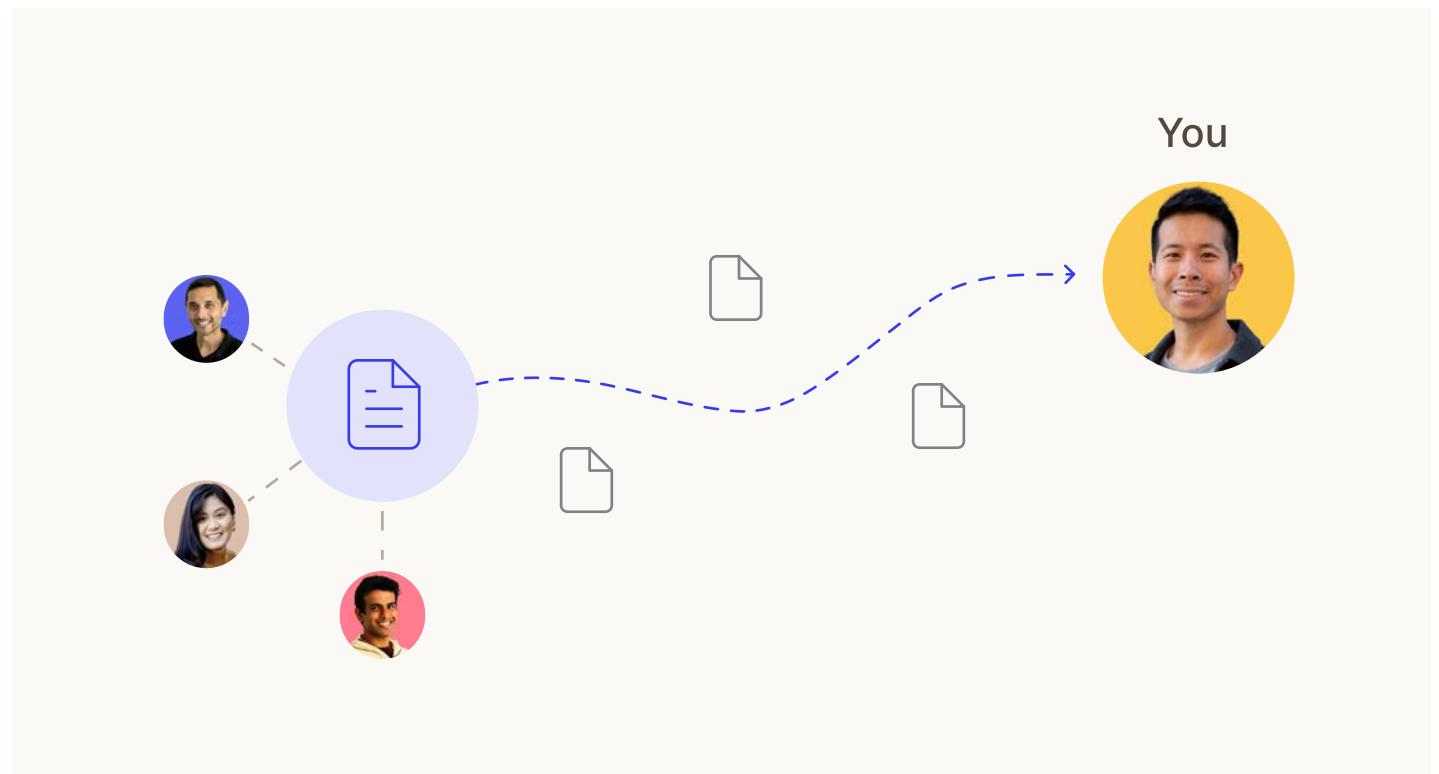
Glean collects activity data from several apps (Teams, Slack, email, plugins, Chrome extension, etc.) to index critical signals required for better search personalization and relevance. We only collect activity on sources connected to the product. As mentioned, none of this activity information ever leaves your exclusive GCP project, and follows strict data protection rules to ensure privacy.

The activity information is used in two ways:

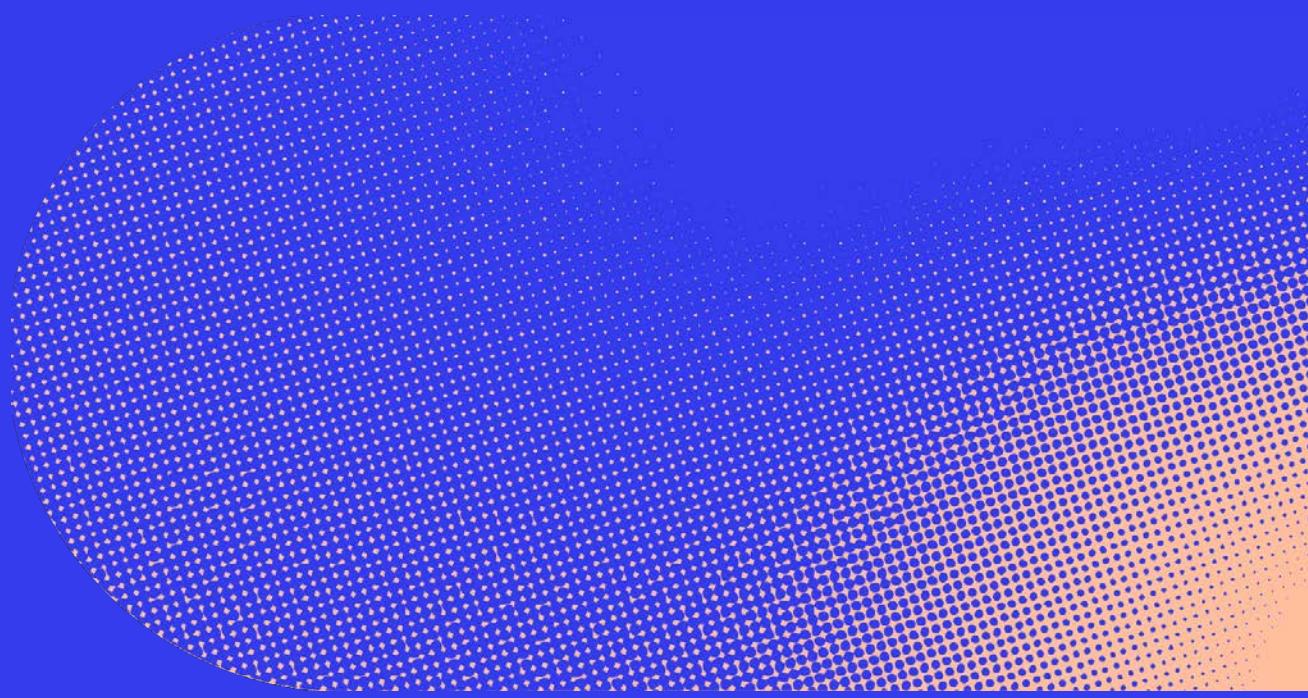
- **Learning what information matters most to better personalize results for users** – individual user data does not leak over to any other user.
- **Improving personalization for a collection of users** – privacy thresholds ensure data is only collected when we see a common datapoint across multiple users.

Understanding activity enables Glean to better detect sharing and usage patterns, and provide more personalized search results by adjusting ranking based on your activity and that of your teammates.

Is there an exceptional piece of enablement content that your team's been circulating often lately? New checklist your IT team's been referring to during audit season? Latest corporate branding guidelines your design team's been working on? Glean already knows they're in high-demand, and that you'll want to see it as well.

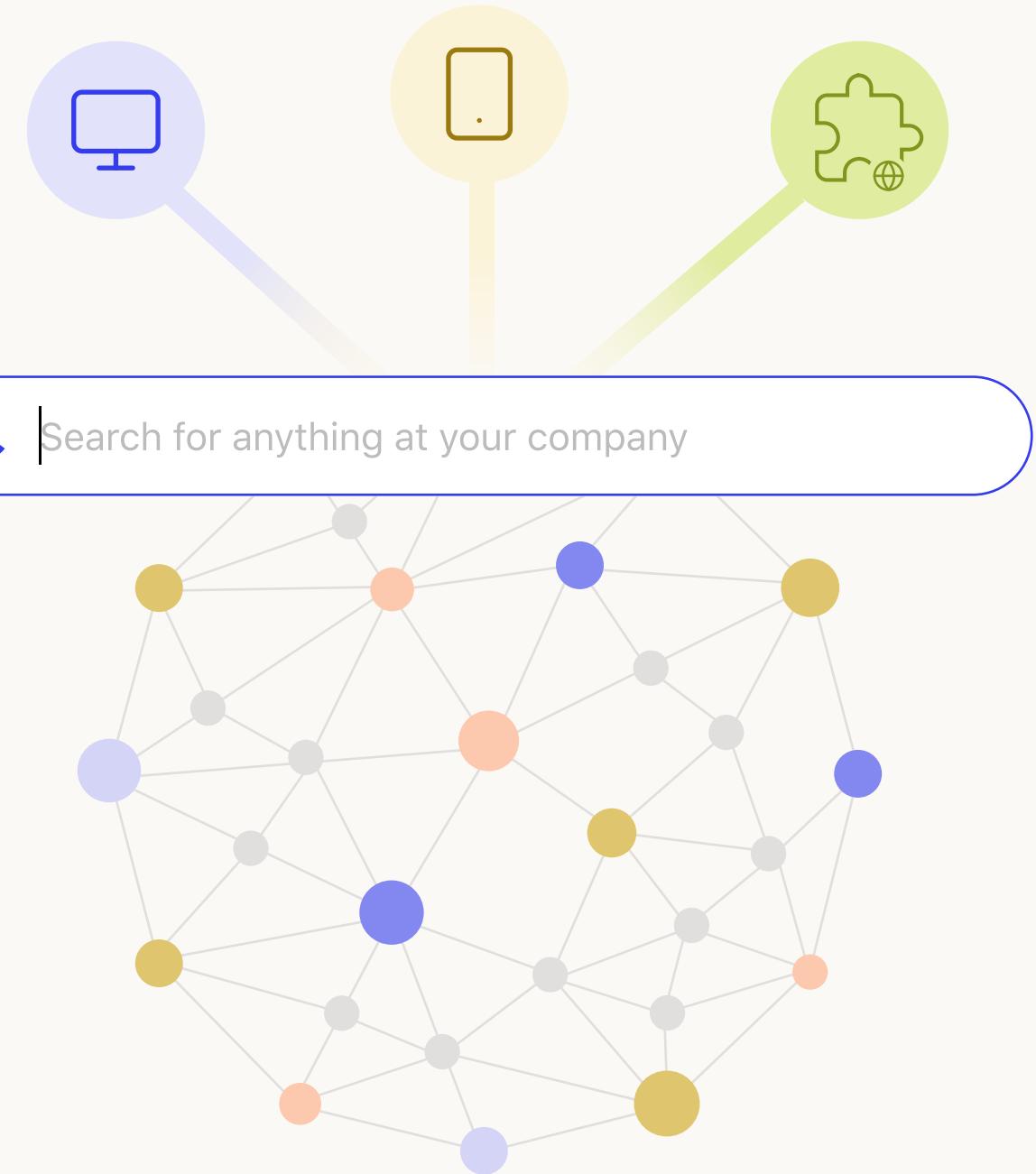


The search journey



Once all this information is collected, Glean then compiles it into a **central search index**.

Users can then send in queries through either the web app, mobile app, or browser extension to begin discovering what they need to know through the compiled search index.



Let's take a trip through the search process using a common example.

There's a curious employee looking for a recap of last quarter's goals and achievements, and decides to search through Glean.

However, they accidentally input "Quorter 4 OKRs" – and that's okay!

Typos happen. Glean offers spelling correction suggestions beneath the search bar, so it's easy to get back on course.

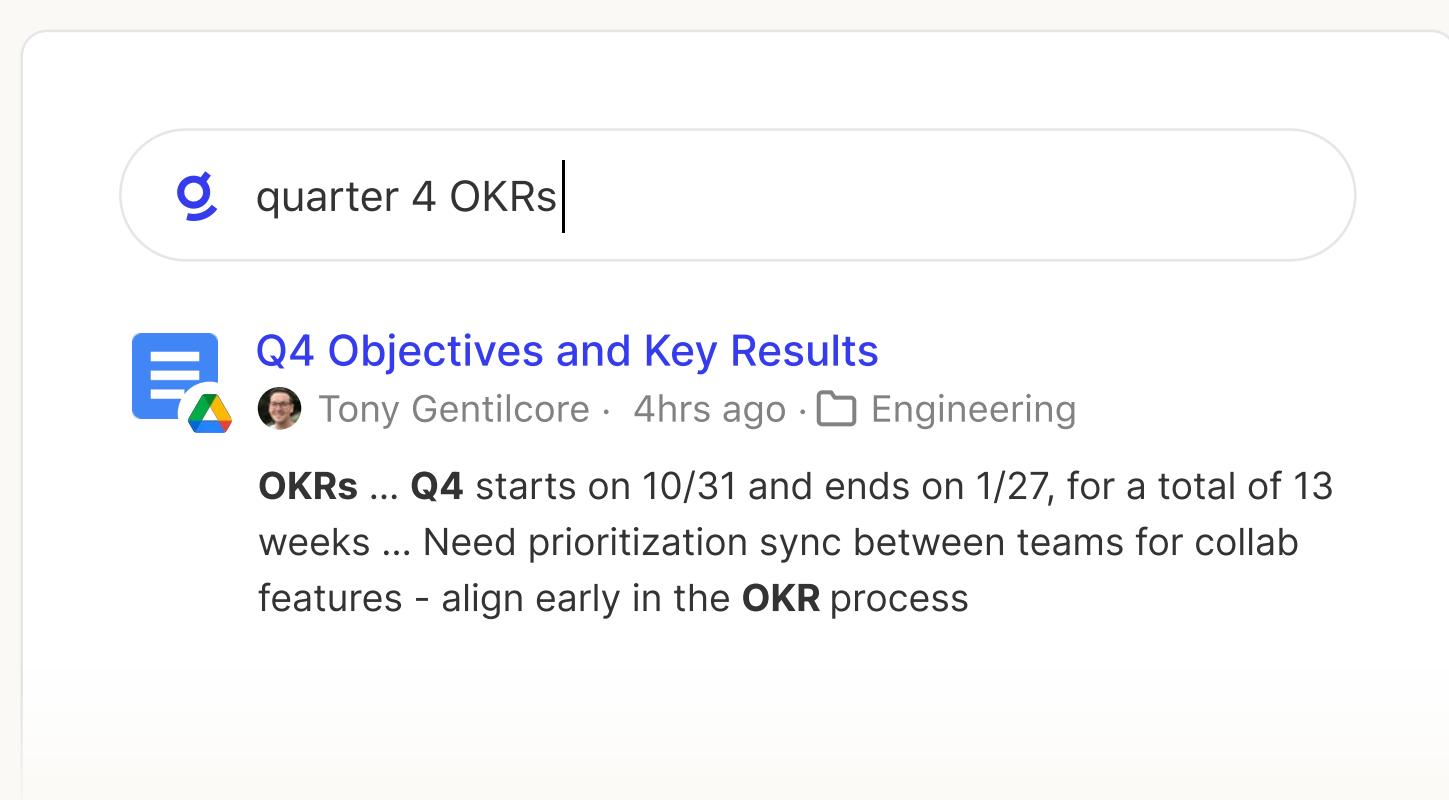
g quorter 4 OKRs |

Did you mean: *quarter 4 okrs* ?

Unfortunately, the name of the actual file is a little different from what they searched for. The document they need is titled “**Q4 Objectives and Key Results**.”

This isn’t a problem, either.

Glean is capable of utilizing deep learning techniques to understand company-specific language, common industry synonyms, and natural human language to compensate for imprecise search queries. Glean also uses semantic search techniques, which analyzes relationships between words in a query and understands the deeper context behind a searcher’s input. So, in this case, it’ll still provide results most relevant to what’s being searched for. Although there’s no file in the search index titled “**Quarter 4 OKRs**”, it’ll still be able to provide our searcher with results titled “**Q4 Objectives and Key Results**”.



quarter 4 OKRs

 [Q4 Objectives and Key Results](#)
Tony Gentilcore · 4hrs ago · Engineering

OKRs ... Q4 starts on 10/31 and ends on 1/27, for a total of 13 weeks ... Need prioritization sync between teams for collab features - align early in the **OKR** process

However, it appears that there are **multiple files** in the search index titled **“Q4 Objectives and Key Results”** – files made in different years, as well as files made by different teams.

This isn't a problem.

Glean is capable of surfacing all topical files relevant to the importance of their query. Furthermore, as long as Glean has access to appropriate activity data, it's capable of providing exactly what the searcher needs.

How exactly does this happen?

quarter 4 OKRs



[Q4 Objectives and Key Results](#)

 Tony Gentilcore · 4 hrs ago ·  Engineering

OKRs ... **Q4** starts on 10/31 and ends on 1/27, for a total of 13 weeks ... Need prioritization sync between teams for collab features - align early in the **OKR** process



[Q4 Objectives and Key Results](#)

 AJ Tennant · 1 yr ago ·  Sales

Enterprise Sales **Q4 2022** OKRs ... 1 **Quarterly** pillar of enablement ... Reps submit peer shout-outs every month/quarter to Josh and AJ ... Highlight culture awards at QBR...



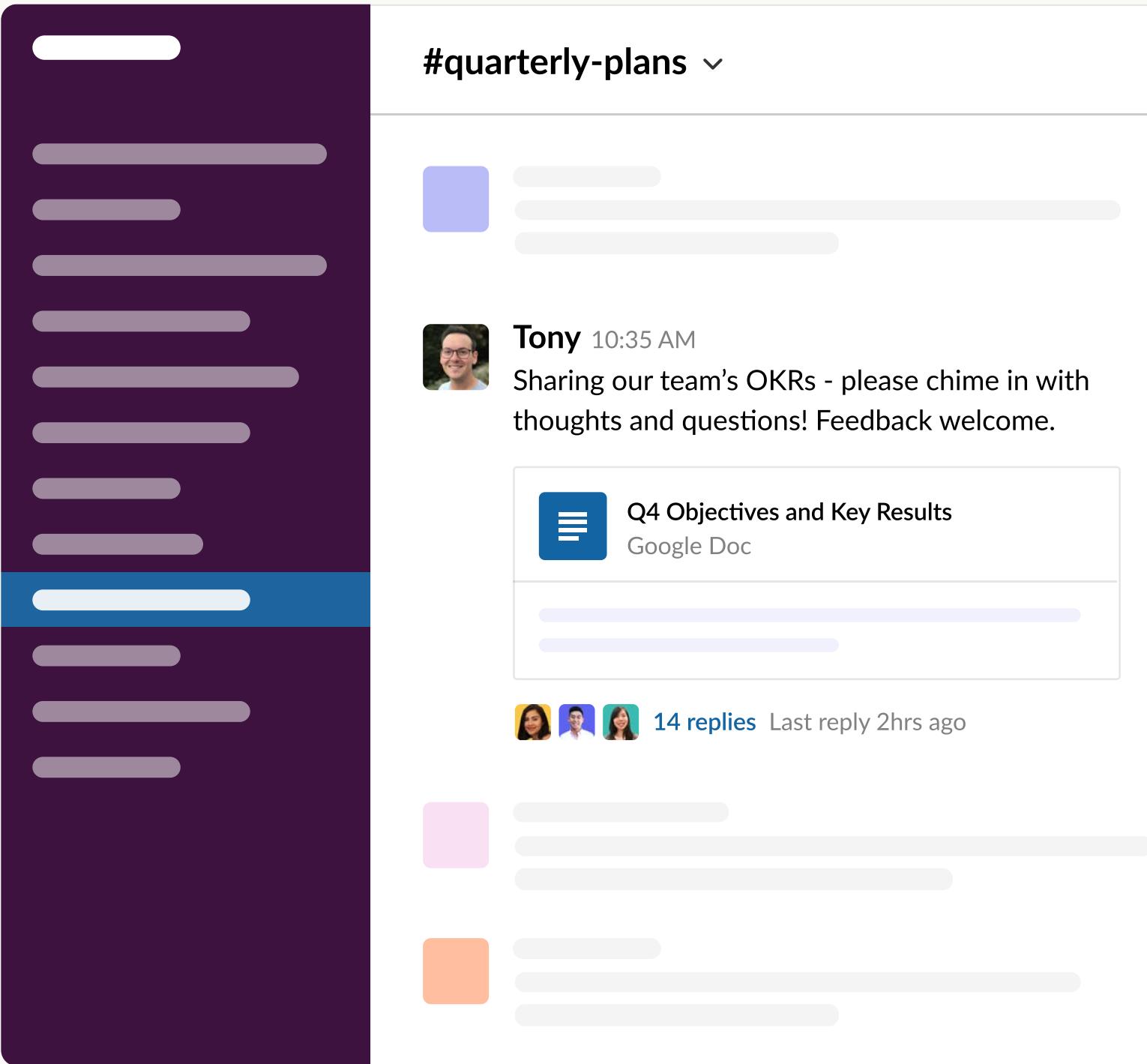
[Q4 Objectives and Key Results](#)

 Ankit Jain · 2 yrs ago ·  Platform

Platform India Excellence **Q4 2021** OKRs ... Moved to a team's OKRs ... **Q4** starts on 10/31 and ends on 1/27, for a total of 13 weeks ... we'll get a **quarter's** heads up to localize as needed

Glean takes into account the activity of the searcher and their teammates across multiple apps while providing results.

Although there are multiple results with the same title, Glean understands that there's only one file that the searcher's collaborators have been recently viewing and/or regularly sharing. Metadata from the file also informs Glean that this file is currently extremely popular and most relevant for the searcher compared to other possible results.



#quarterly-plans ▾

 **Tony** 10:35 AM
Sharing our team's OKRs - please chime in with thoughts and questions! Feedback welcome.

 **Q4 Objectives and Key Results**
Google Doc

 14 replies Last reply 2hrs ago

It also understands that the searcher is on a specific team – let's say for example, **the engineering team**.

Although the sales team may also have a file with the same name stored in the search index, Glean knows the right file to share with the searcher based on the fact they're on the engineering team.

Resultantly, **search results will often vary based on the searcher's role and the team they're on**.



For an engineer...



Q4 Objectives and Key Results



Tony Gentilcore · 4 hrs ago ·  Engineering



Q4 Objectives and Key Results



John Chang · 1 yr ago ·  Sales



Q4 Objectives and Key Results



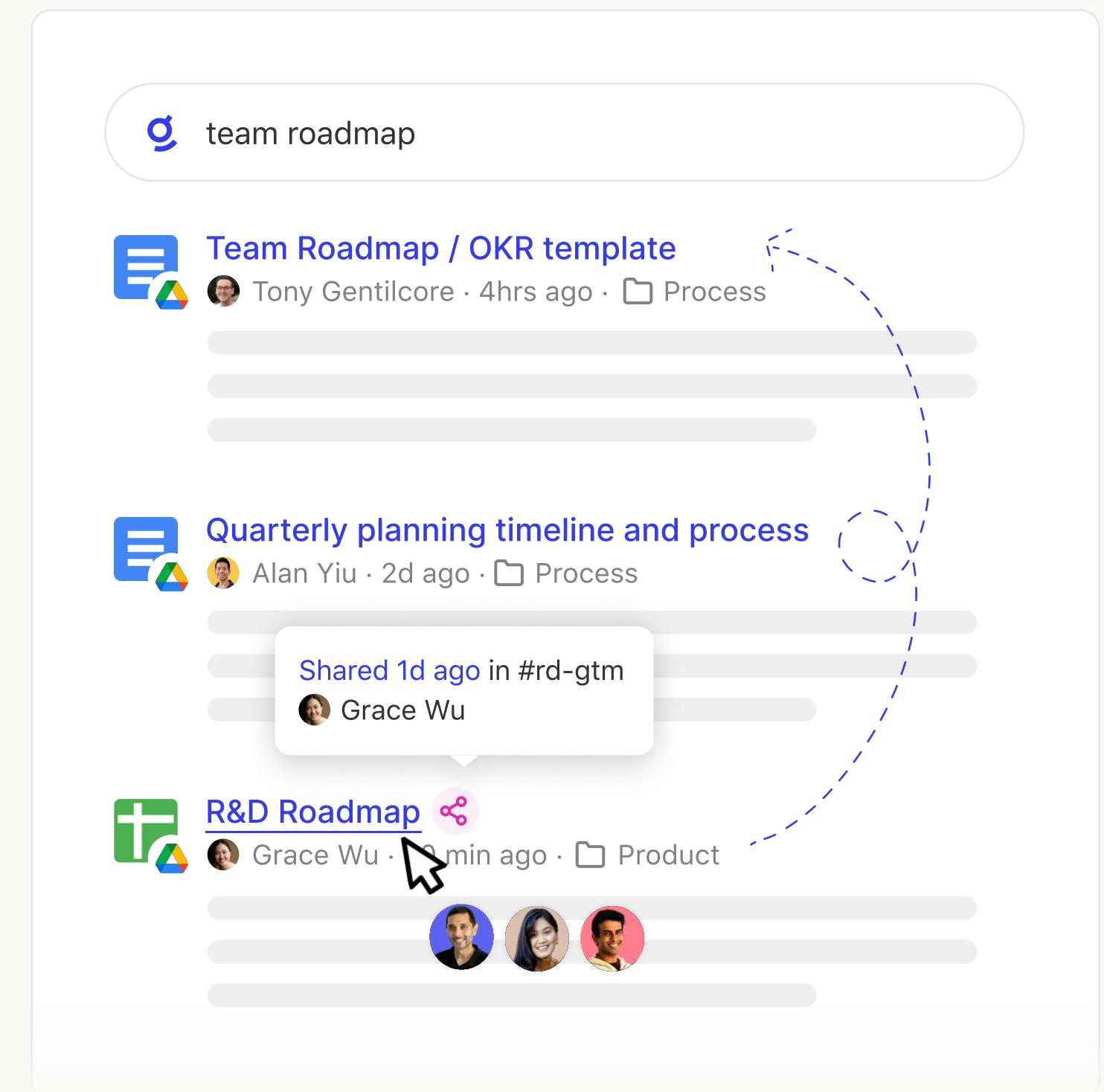
Ankit Jain · 2 yrs ago ·  Platform

On the other hand, if the search result somehow still isn't perfect, the ranking algorithm consistently learns and improves.

If multiple searchers from a team with the same query click on the link in the third result, it'll adjust to boost that result to a higher position for future searchers from that team.

Looking to double-check the relevance of a file? Glean is capable of showing the most recent activity regarding a file through a popup – where it was shared, and how recently.

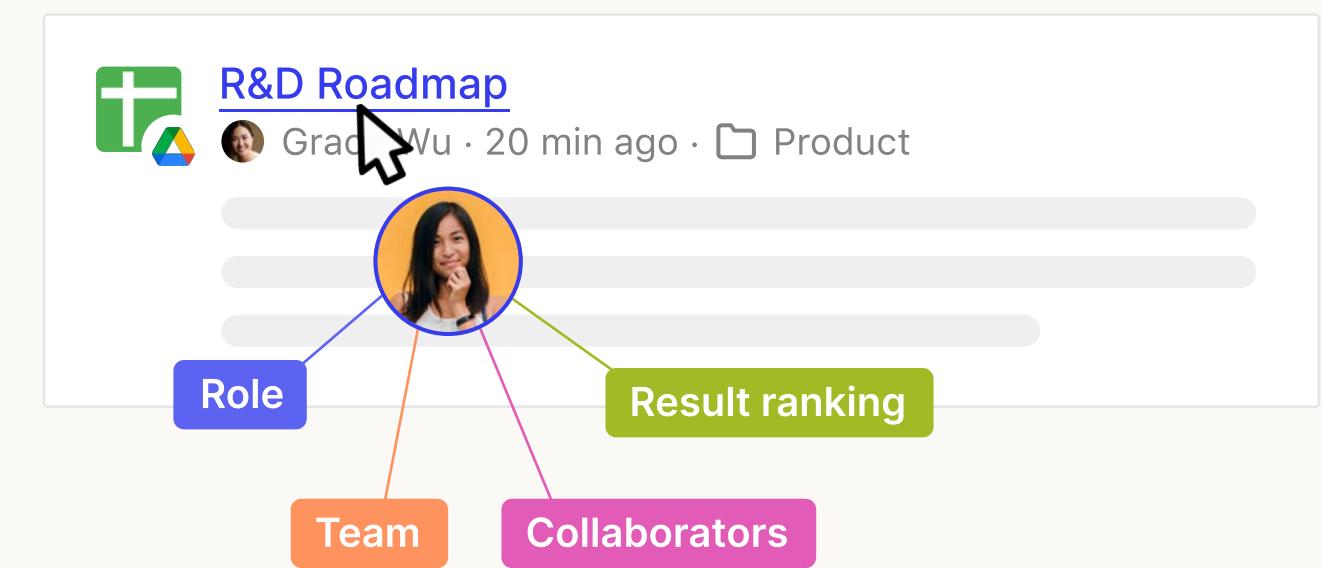
Power users looking for specific files can also specifically search for what they're looking for by utilizing search filters, separating results based on filters like source, file type, and update recency.



Once the searcher locates what they need and completes the search, Glean stores any information relevant to this search session to improve future searches. This includes data such as the resulting file they clicked on, which team they're on, their role, and the ranking of the result.

Best of all, this entire process takes the searcher just a few seconds from start to finish. Compared to search solutions that rely on federated search – which runs every query through multiple search APIs that need to be individually authorized for search – Glean's indexing system keeps company data frequently updated, and results consistently fresh, accurate, and quick to load.

Query submitted, query answered!





Search across all your company's apps.

Excited about Glean search, and ready to learn more? Discover the world's most comprehensive and powerful enterprise search tool for yourself. Get a demo of Glean today, and join other leaders that have chosen Glean as their champion of information accessibility.

[GET A DEMO](#)

Canva



okta



Vanta



greenhouse

