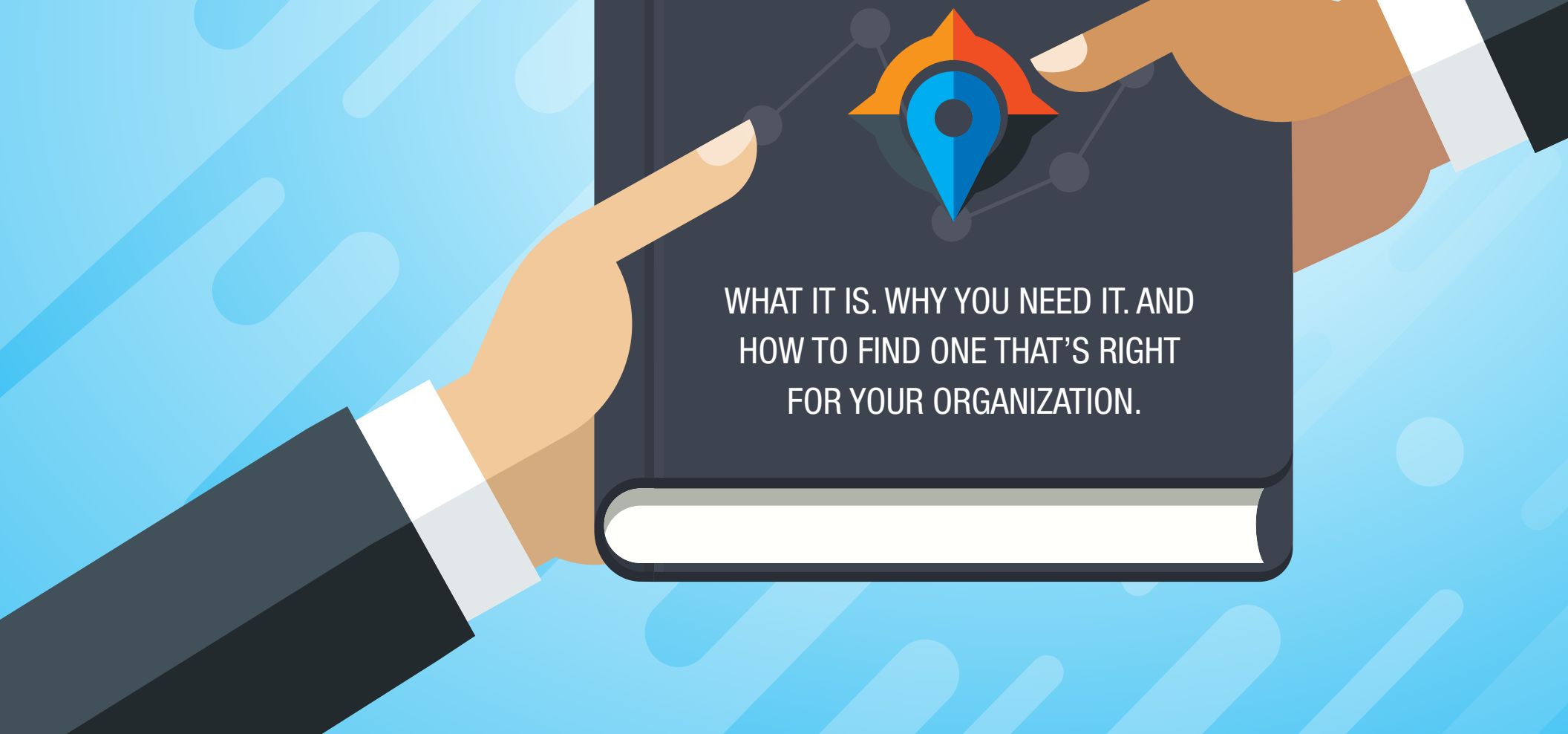
collibra®

A COMPREHENSIVE GUIDE TO THE DATA CATALOG



WHAT IT IS. WHY YOU NEED IT. AND HOW TO FIND ONE THAT'S RIGHT FOR YOUR ORGANIZATION.

By 2020, organizations that offer users access to a curated catalog of internal and external data will realize twice the business value from analytics investments than those that do not.

– Gartner Magic Quadrant for Business Intelligence and Analytics Platforms, February 2017

THE CASE FOR A DATA CATALOG IS GROWING.

There's no question about it. Data is a valuable business asset with the potential to transform almost every aspect of the enterprise. In fact, investments in BI and analytics software are predicted to grow to **$22.8 billion** by the end of 2020.

BUT DRIVING REAL VALUE FROM THOSE INVESTMENTS HAS NOT KEPT PACE.

About 60% of new business intelligence initiatives will **fail to get off the ground**.

Knowledge workers waste **50% of their time** searching for data, correcting mistakes, and looking for the right people to confirm the data they do find is trustworthy.

Data scientists spend **60% of their time** cleaning and labeling data, rather than using it to drive new insights.

With the shift to the cloud, data sources—and data silos—are proliferating, making it harder to find and share good data.*

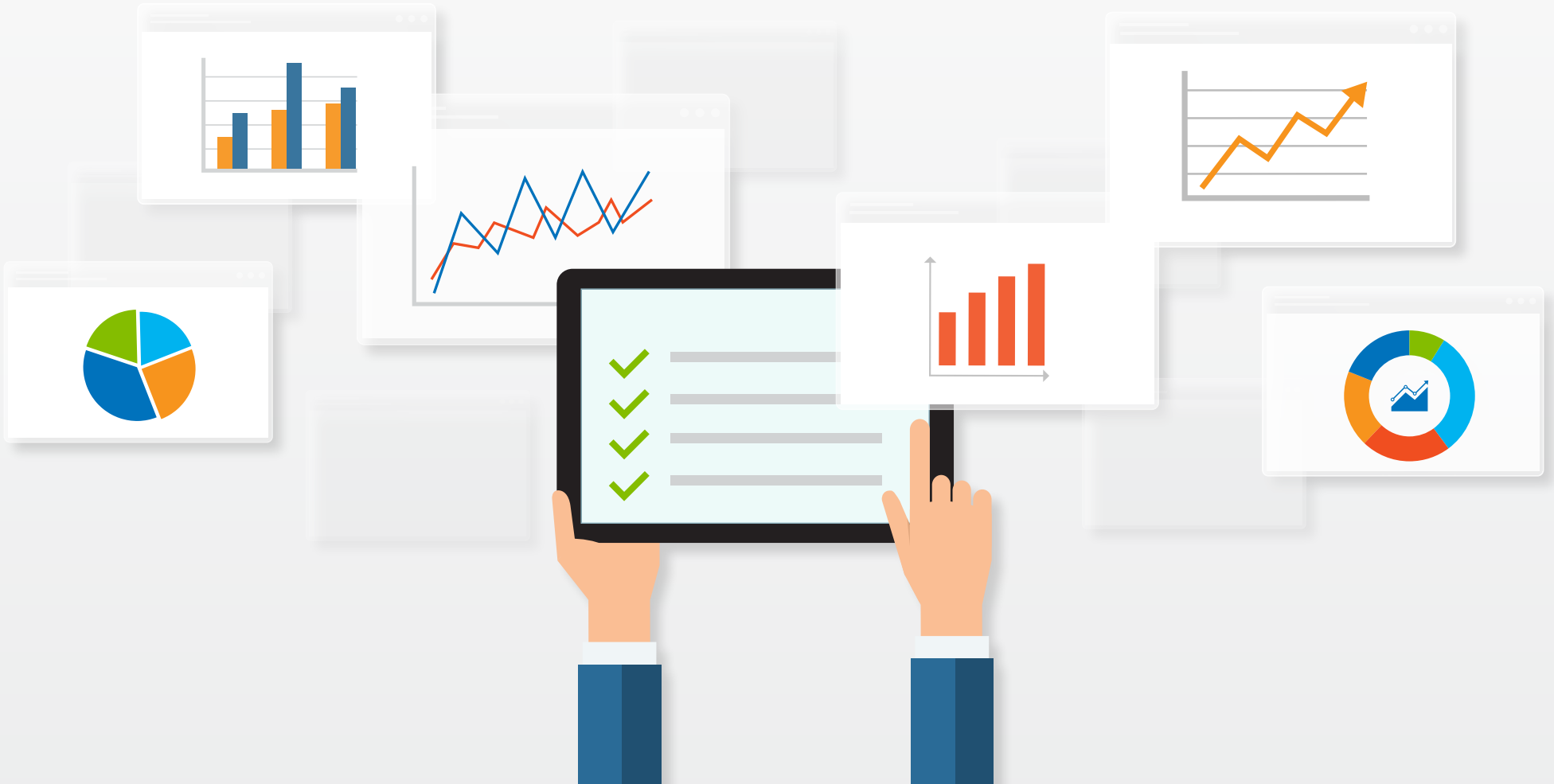
Organizations still struggle with data quality. In fact, a staggering **$3.1 *trillion* of U.S. GDP** is lost because of poor quality data.

* http://www.datacenterknowledge.com/archives/2017/06/15/time-downsize-integration-efforts/ | https://hbr.org/2016/12/breaking-down-data-silos
http://www.computerworld.com/article/2924739/data-analytics/attack-of-the-data-silos-2-coming-soon-to-an-enterprise-near-you.html

Simply put, as data becomes more critical to enterprise success, people are spending too much time hunting for meaningful, trustworthy data. To help data consumers at every level of the enterprise find, use, and understand their data, many organizations are considering how a data catalog can help.

DOES YOUR ORGANIZATION NEED A DATA CATALOG?

Likely, it does. According to a McKinsey Global Institute report, organizations across all industries and geographies have made uneven progress in capturing value from their data.



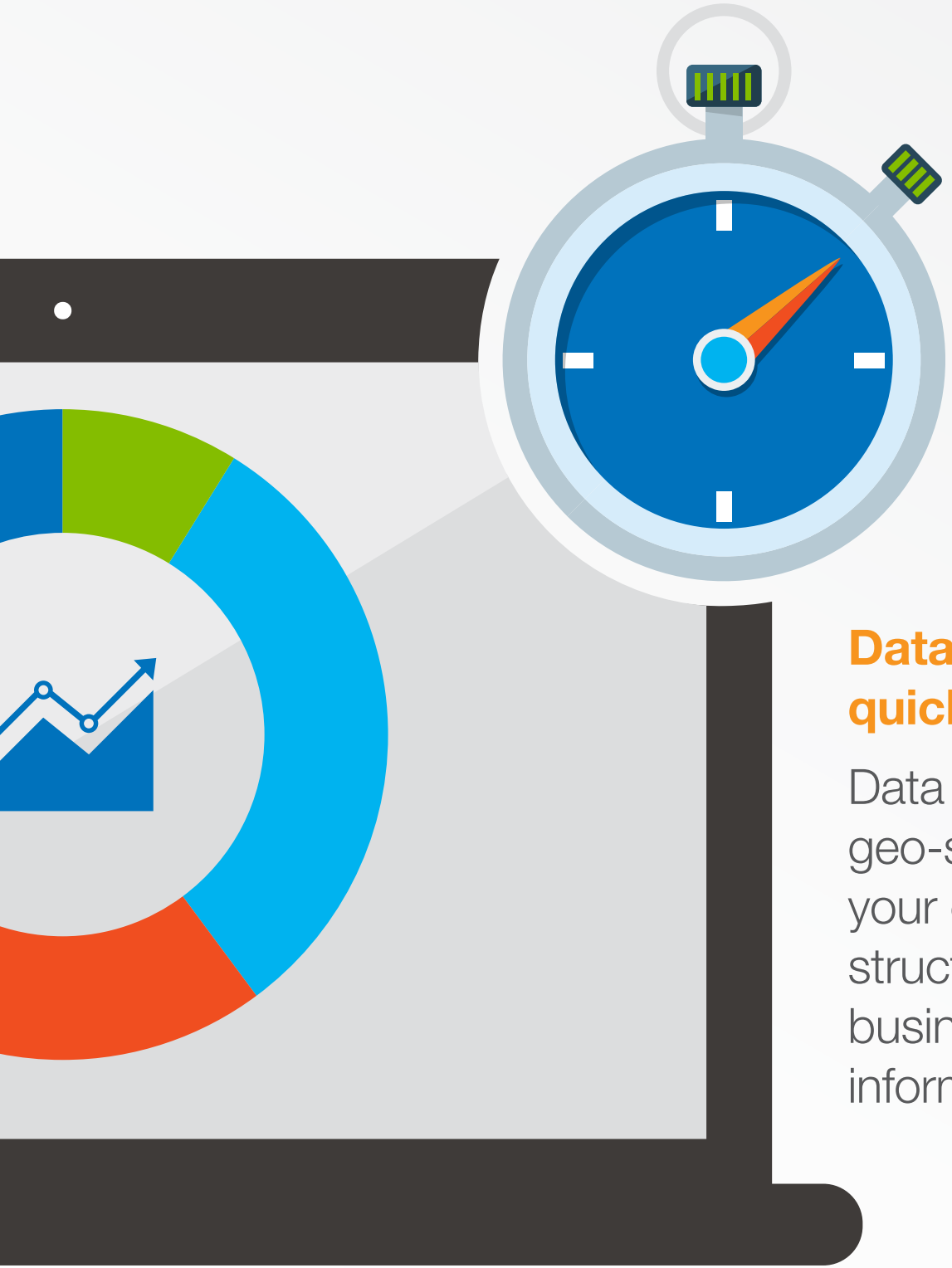
If you're trying to decide whether or not your organization needs a data catalog, consider some of these common indicators:

- You've spent a lot of money on self-service BI, but finding the data to populate those tools is still difficult.
- Your organization has numerous data sources, but data consumers have no easy way to identify those sources in one central location.
- Your "let's store everything here" data lake has turned into a murky data swamp that makes finding meaningful, trustworthy data close to impossible.
- There is no process in place for data consumers to request the data they need.
- Even when data consumers can access data, they don't have information about what the data means or how it should be used.
- Data consumers don't know the source of the data they find and so cannot confirm the data's trustworthiness.
- Data consumers don't know who 'owns' the data and therefore have no way to contact subject matter experts.
- Data consumers don't know what data sets already exist across the enterprise or who has used similar data to explore similar problems.

WHAT CAN A DATA CATALOG DO FOR MY ORGANIZATION?

Understanding the real value that a data catalog can bring to your organization can help you articulate real business benefits to gain buy-in from your executives as well as technical and business users, garner cross-organizational support, and promote adoption.

Here are some ways a data catalog can help.

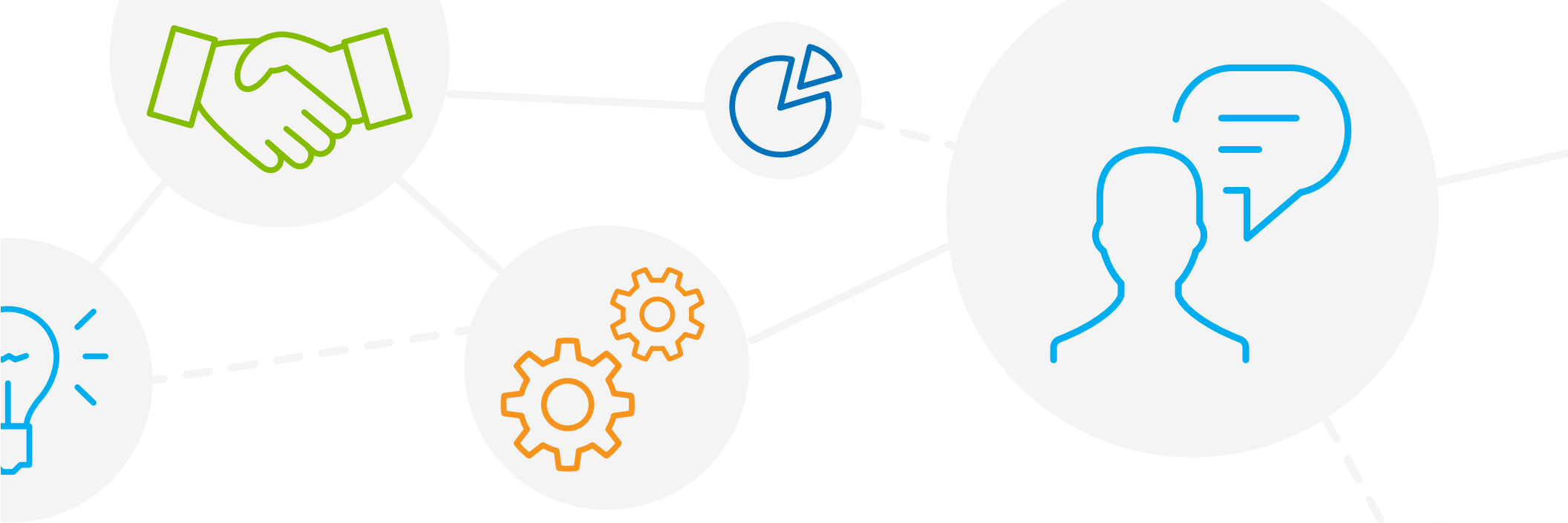


Data catalogs help data consumers quickly find the data they need.

Data catalogs serve as a kind of geo-spatial guide to the data across your organization, providing a sensible structure to help data experts and business users alike find relevant information more quickly.

Data catalogs can make it easier for everyone to understand the data they're using.

Good data catalogs typically contain descriptive information about the data in easy-to-understand business terms. Because data consumers don't have to interpret technical jargon or trudge through hundreds of rows, columns, and tables, they're more likely to understand and trust the data sets they find—and use them to power the business.



Data catalogs can promote sharing, collaboration, and, ultimately, innovation.

Data catalogs allow data consumers to tag, document, or annotate data sets to help other data users across the organization understand how that data is being used and how helpful it's been. And the best data catalogs provide a clear snapshot of the data's lineage and profile, so that data consumers can combine data in new ways to drive new insights.

Data catalogs can help uncover “dark data.”

A widely-adopted data catalog makes it easier for data consumers to uncover previously “hidden” data by recommending approved and trusted datasets and making them easily discoverable through semantic search.

WHERE SHOULD I BEGIN?

Before you begin to evaluate a data catalog for your organization, make sure you understand why you're doing it and what you hope to achieve.

Organizations have different reasons for purchasing a data catalog solution. What are your goals and priorities?

Are you trying to drive value from your data assets?

Then you will want to focus your evaluation on how easily the data catalog will be adopted by data consumers across the enterprise, how well it will break down data silos, and whether or not it will improve and accelerate decision making.

Are you trying to help your data scientists spend less time wrangling data and more time analyzing it?

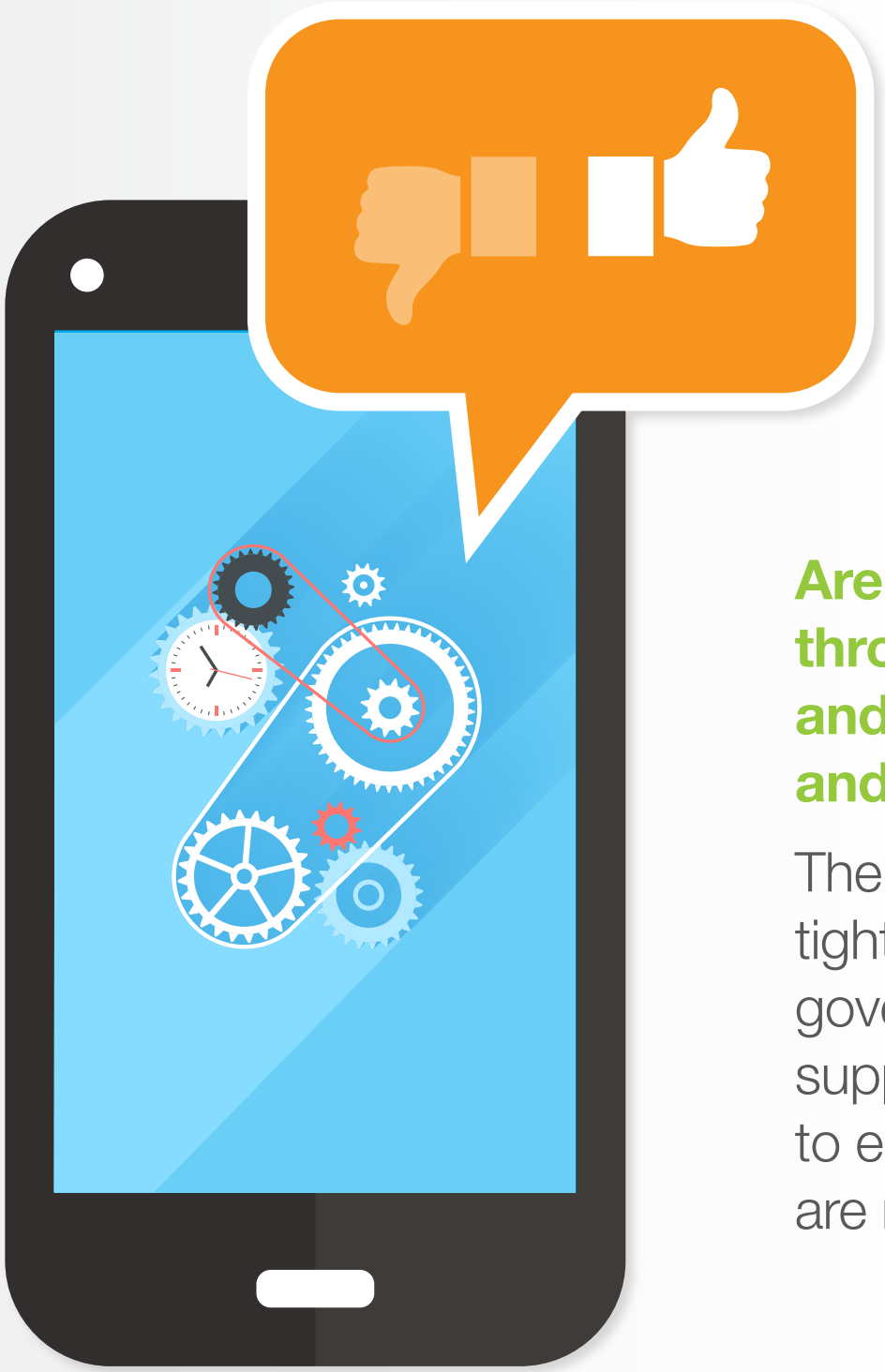
Then you will want to focus your evaluation on the things your data scientists will care about: how well does the data catalog help them find the data sets they need? How well does data profiling help data scientists reveal the shortcomings or validate the usability of the data sets?

Are you trying to help business users find information faster?

Then you will want to focus your evaluation on how well the data catalog helps business users quickly and easily find and understand the data they need by presenting data in business-friendly language they will recognize.

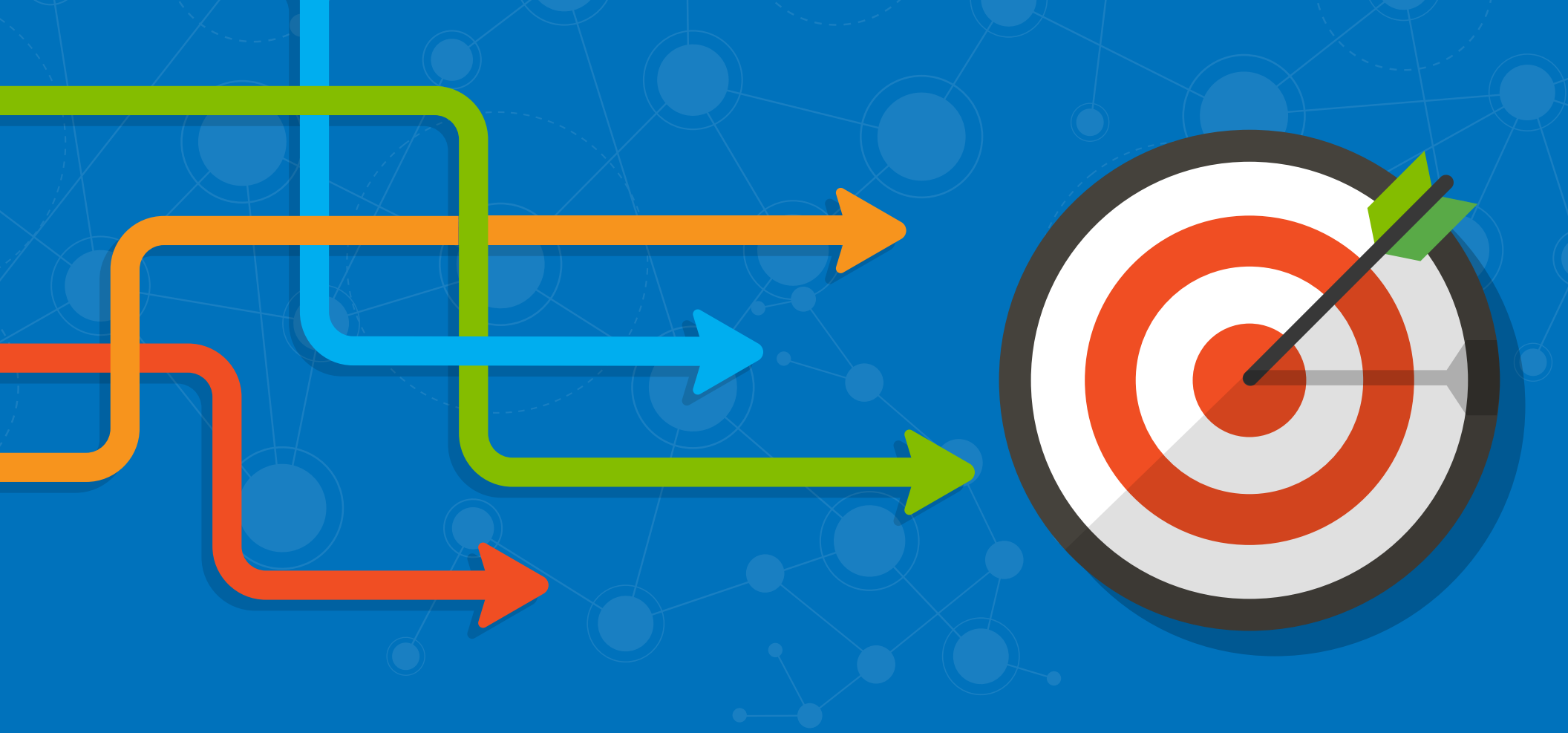
Are you trying to build a more collaborative context for decision making?

Then you will want to focus your evaluation on how the data catalog can help data consumers understand the context of their data—where it has been, who owns it, who can access it, and the rules governing its use.



Are you trying to promote trust through better data policies and improved data quality and consistency?

Then you will want a data catalog tightly integrated with data governance capabilities which support data rules and allow you to easily configure workflows that are right for your organization.



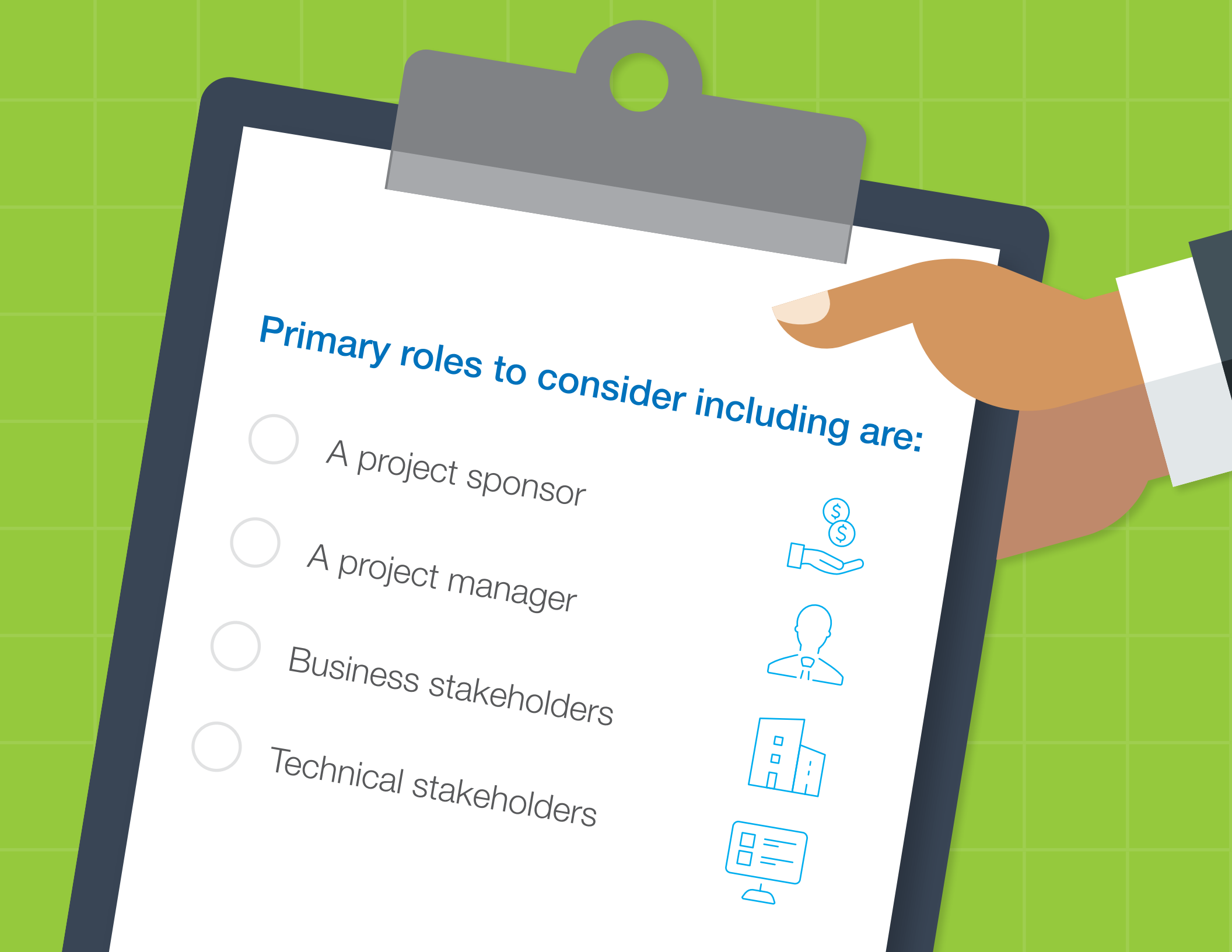
You may have more than one goal for adopting a data catalog—most organizations do. Understanding them and documenting the business problems you hope to solve will help you build the framework and metrics you need to properly evaluate proposed solutions.

WHO SHOULD BE INVOLVED IN THE EVALUATION PROCESS?

Determining who should be involved in the data catalog evaluation process isn't simply a matter of technical expertise. A data catalog is also—even primarily—a business asset, and so, technical *and* business users must be involved in the selection. And remember, for many team members, the evaluation process is time spent away from their core responsibilities. Before assembling a team, consider the contributions individuals will make to the project and any constraints on their time.

Primary roles to consider including are:

- A project sponsor
- A project manager
- Business stakeholders
- Technical stakeholders



A project sponsor.

Sponsors provide strategic vision, helping stakeholders understand how the project is connected to larger business goals.

Sponsors translate that vision into clear deliverables, helping to define the scope of the project.

Sponsors secure funding and champion the project at the executive level.

A project manager.

Project managers execute the software evaluation process.

Project managers typically lead the planning and budgeting, requirements analysis, vendor research, structured demos, contract negotiation, and often, implementation.

Business stakeholders.

Involve representatives from each business area that will benefit from, or be impacted by, the implementation of a data catalog.

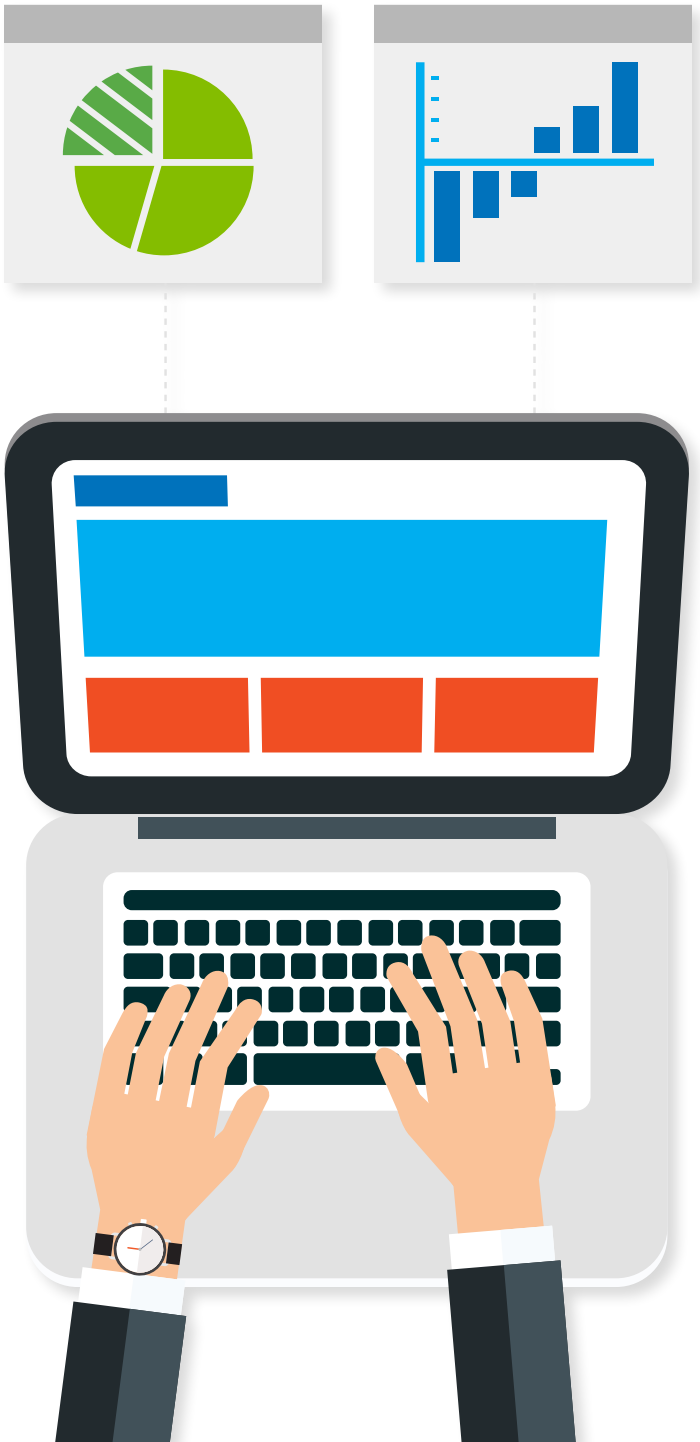
Involve those business users who can best articulate business requirements for a data catalog that meets their needs.

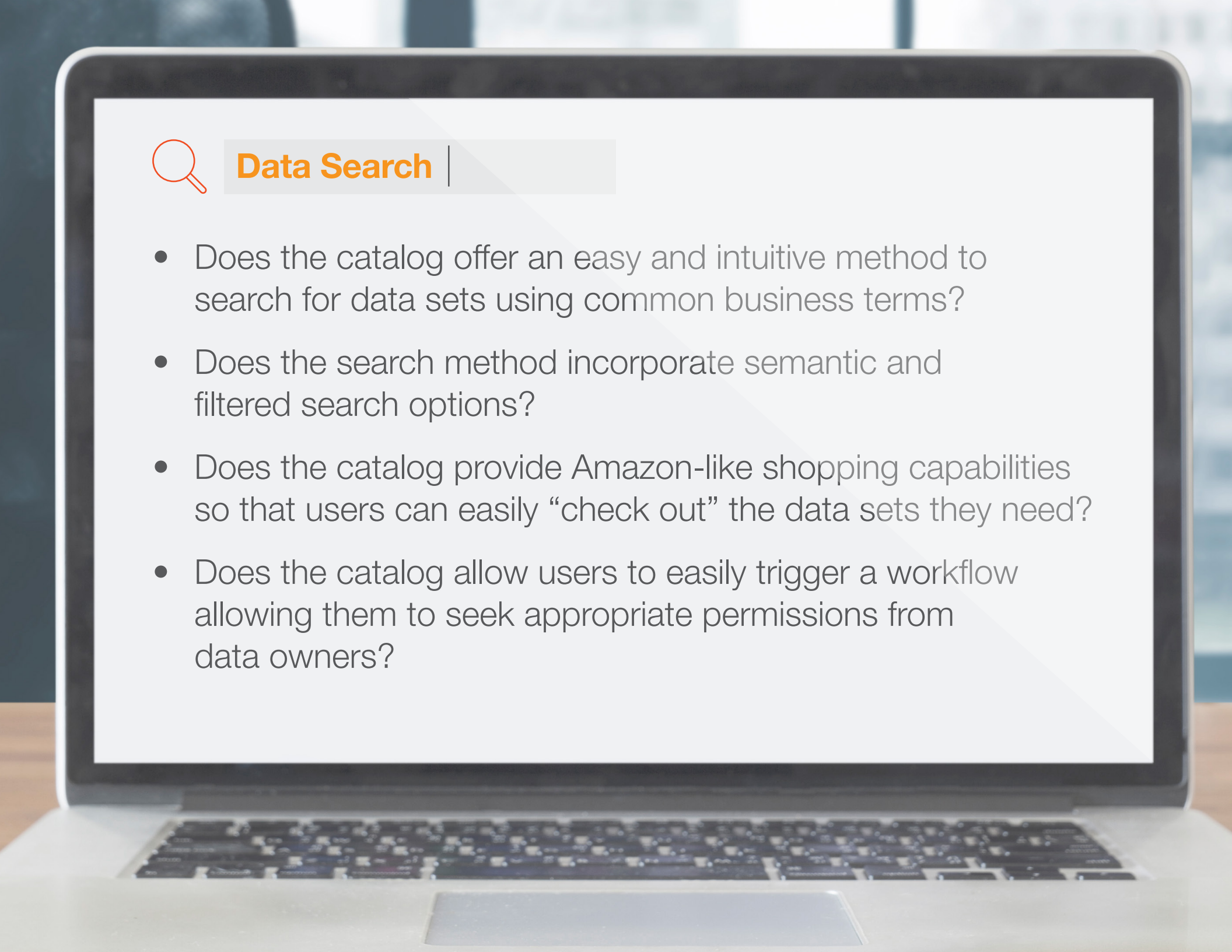
Technical stakeholders.

Involve technical users who can describe and document current data sources, data discovery and BI tools, and technical dependencies.

WHAT FEATURES SHOULD I BE LOOKING FOR?

For a catalog to be useful to business users, it needs to connect to data that's meaningful to them and then organize and present that data in business-friendly terms. While not a comprehensive feature list, the functionality on the coming pages will provide your business users the edge they need to more easily find, understand, and trust the data they need to do their jobs.





Data Search

- Does the catalog offer an easy and intuitive method to search for data sets using common business terms?
- Does the search method incorporate semantic and filtered search options?
- Does the catalog provide Amazon-like shopping capabilities so that users can easily “check out” the data sets they need?
- Does the catalog allow users to easily trigger a workflow allowing them to seek appropriate permissions from data owners?

Data Recommendations

- Does the catalog use machine learning to recommend additional data sets (based on what others users have accessed, similar data sets, matching data sets, or pattern matches)?

Data Assembly

- Does the catalog allow users to easily combine data sets from disparate sources to form new data sets?
- Does the catalog automatically suggest business terms related to the data set?

Data Sampling

- Does the catalog provide an actual sample view of the data?

Data Profiling

- Does the catalog provide summary information about data selected so that business users can easily assess its quality?
- Does the catalog use the profiling information to identify the meaning of the data?

Data Lineage and Data Usage

- Does the catalog provide a clear understanding of where data is coming from, who has used the data, and when?
- Does the catalog provide a granular view of data lineage—from columns to tables to data sets?
- Does the catalog help users understand allowed values (reference data)?
- Does the catalog present this information in an easy-to-understand, business-friendly dashboard?

Collaboration and Crowdsourcing

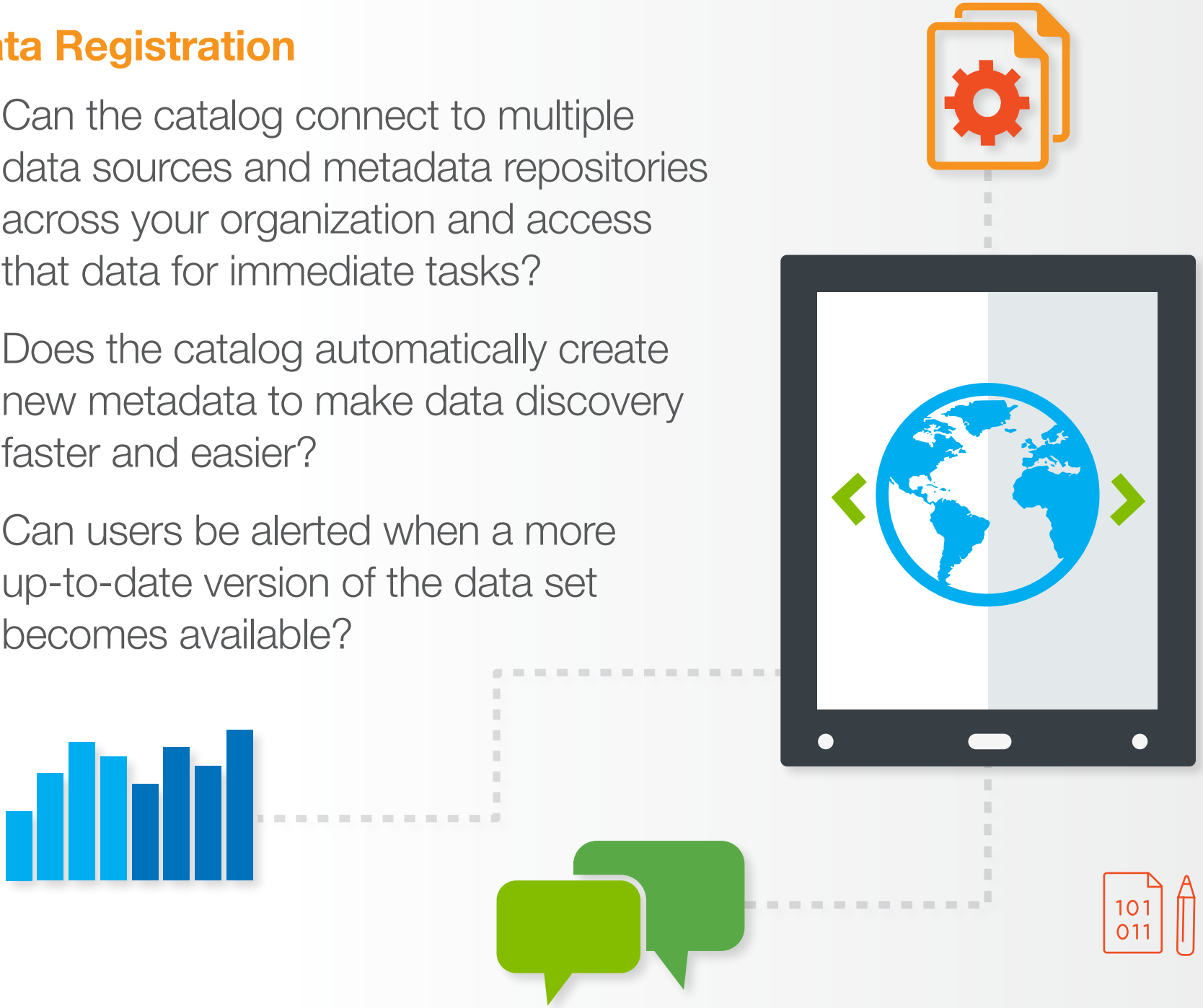
- Does the catalog support free-form tagging and commenting to allow business users to annotate data sets in a way that makes sense to them?
- Can new, better data sets be “crowdsourced,” i.e., shared, annotated, enriched, and improved—while still maintaining the integrity of the data?

Governance

- Does the catalog make it easy for data users to find data owners, data stewards, and other people who can provide context, answer questions, and correct bad data?
- Does the catalog link the data source to your business glossary so that business users can easily understand what the data means?
- Does the catalog provide insight into the quality of the data so that data users feel confident using it in their reports and analytics?
- Does the catalog provide a secure access process?
- Does the catalog provide insight into your data's lineage?
- Does the catalog provide a way to resolve issues easily and according to established data policies?
- Can the catalog connect to your existing BI ecosystem and serve as a hub for stewardship activities?
- Does the catalog provide a way for you to certify data sets according to data policies set by your organization?

Data Registration

- Can the catalog connect to multiple data sources and metadata repositories across your organization and access that data for immediate tasks?
- Does the catalog automatically create new metadata to make data discovery faster and easier?
- Can users be alerted when a more up-to-date version of the data set becomes available?



WHAT SHOULD I ASK DURING A DEMONSTRATION OF THE PRODUCT?

Comment*

If you've provided scripts to each of your vendors, they should be able to design a product demonstration that answers most of your questions. Still, when purchasing any technology, the *how* is as important as the *what.*

SUBMIT

During a demonstration of the solution or subsequent conversations, be sure to ask each vendor the following questions:

How do you incorporate data profiling and sampling to create metadata?

How do you link catalog data sources to business terms already established by the organization?

How do you approach collaboration – and what capabilities do you provide to business users to enable them to work together better across the organization?

What is your process for syncing and alerting users to changes in data sets?

How do you incorporate data governance principles and processes from your data governance platform into the data catalog?

HOW DO I EVALUATE POTENTIAL VENDORS?

Customers are the best resource for finding out the real skinny about a vendor, so make sure you talk to them. Trustworthy solution providers will be happy to provide customer contact information or invite you to participate in a customer event—typically an annual user group.

Here are some questions you might want to ask.

- If you had to do it again, what would you do differently?
- Were there any surprises you weren't prepared for?
- Were you satisfied with the service provider's support and/or training?
- How did you involve your business users in defining requirements for the catalog?
- Are your business users taking advantage of the catalog? Why or why not?
- According to business users, what are the catalog's best features?
- Does the catalog provide easy connectivity to workflows, collaborative platforms, and governed data?
- What support, training, and regular updates on best practices does the vendor provide?
- Have you attended a user conference?
- What doesn't it do that you would like it to do?

DO YOUR RESEARCH

Examples of useful research and resources include:

- Recent analyst research and reports
- Buyer's guides
- White papers, case studies, and market reports
- Customer feedback
- Online discussion forums
- Webinars or vendor comparison forums
- Recent blog posts about enterprise software releases or trends

Finally, as you evaluate your options, remember that a data catalog is ultimately a tool for helping business users—analysts, data scientists, and decision makers—quickly find trustworthy data to help them do their jobs. The catalog should be simple to use and easy to navigate. It should help users find not just *any* data but the *right* data—by presenting prioritized data and recommending useful data sets with a transparent and traceable history, identifiable owners, and clear policies. And it should make that data easy to find, easy to access, and easy to understand.

Learn more at
collibra.com/data-catalog

collibra™

collibra.com

info@collibra.com

Follow Us:
twitter.com/collibra