Element 84
element84.com
info@element84.com

The Case for a Centralized Earth Observation Vector Embeddings Catalog

A catalog of this nature would search the Earth, understand crucial changes, and facilitate democratization.

Jason Gilman, Adeel Hassan, Nathan Zimmerman

Abstract

Massive Earth Observation (EO) archives containing hundreds of petabytes of data offer a huge potential for understanding the Earth's resources and potential hazards to security while also addressing societal challenges. Even so, extracting actionable insights remains a challenge. This data requires specialized knowledge in machine learning and geospatial expertise as well as significant computational resources. While the size and value of these archives is constantly increasing, we are not making meaningful progress fast enough to handle the challenges of our changing world.

We propose creating a centralized vector embeddings catalog that enables users to extract insights from EO data. EO Foundation Models, having been trained on millions of geographically diverse data points, have the ability to convert raw EO data into vector embeddings, a compressed semantic representation of the data. By indexing these vector embeddings into a catalog, we enable users to directly query complex phenomena like agricultural areas showing drought stress, illegal mining operations, or construction projects in flood zones without specialized technical knowledge. Additionally, users can use a vector embeddings catalog to quickly search for relevant changes over time, such as an increase in conditions leading to potential wildfires.

We argue how this effort aligns well with NASA's declared goals of accelerating data discovery and democratizing data access and why NASA is uniquely positioned to undertake it. We further lay out a phased program for building this catalog and discuss important challenges and risks that will need to be addressed along the way.

Introduction

Providing fast and easy access to EO data has long been a goal of data distributors like NASA, NOAA, and commercial data providers. For example, the vision statement of NASA's Earth Science Data Systems (ESDS) Program is "to make NASA's free and open Earth science data interactive, interoperable, and accessible for research and societal benefit both today and tomorrow. This is behind much of the work put towards building EO catalogs, establishing metadata standards and specifications like ISO 19115-2 or SpatioTemporal Asset Catalogs (STAC), and moving data into cloud-optimized formats. While we normally think about searching the data through indexed metadata fields we don't often think about searching the data by its contents. But it is the contents that we really care about: the natural features of the Earth, human created infrastructure, and how it's changing. Thanks to advances in Artificial Intelligence (AI) and Machine Learning (ML), such as the creation of EO Foundation Models (EO FMs), there's an opportunity for data distributors to create systems that can directly answer our questions about the Earth such as the following:

- Which agricultural areas in California's Central Valley are showing early drought stress?
- Where are unmaintained quarries in the Balkans?

- Where are unreported construction projects expanding into flood zones in Bangladesh's Ganges Delta?
- Which areas in the Southwestern United States are experiencing rapid commercial development near forests at risk of wildfire?

How can we harness advances in AI to make it possible to easily answer these questions? EO FMs, trained on vast amounts of EO data, have learned to represent data differences. The significance of their understanding is that the model can perceive differences in things like moisture, vegetation health, urban infrastructure, forests, and grasslands. At a technical level, the model represents its understanding in a dense representation known as a vector embedding, essentially a list of floating point numbers.

The nature of vector embeddings allows them to be directly compared to identify areas on the Earth that are similar or where significant changes have occurred, such as drought or increases in urban development. A vector embedding catalog would enable the planetary-scale search of data, classification, or trend analysis without need for specialized technical knowledge of machine learning. Questions like those listed above could be answered rapidly due to indexed views of data by feature, spatial data, and temporal data.

Organizations like NASA are uniquely positioned to lead this effort due to their expansive collection of open EO datasets, existing infrastructure, history of promoting trusted standards, and experience both constructing and working with these foundation models. A NASA led effort would establish a completely new way for users without experience in machine learning or advanced geospatial data processing skills to exploit the huge wealth of NASA Earth science data.

We outline strategic steps for assessing the practical efficacy of embedding-distribution through targeted prototyping and case studies. This approach explicitly aligns with NASA's open science policies and the FAIR (Findable, Accessible, Interoperable, Reusable) data management principles. By investigating the feasibility and utility of embedding-centric infrastructure integrated with existing EO standards, such as the SpatioTemporal Asset Catalog (STAC), the proposed initiative aims to fulfill federal mandates regarding open data access and interoperability.

State of EO

Earth Observation (EO) data collected by satellites, aircraft, and ground-based sensors is large and projected to grow at an accelerating pace. NASA's Earth Observing System Data and Information System (EOSDIS) alone archives hundreds of petabytes of data from dozens of current satellite missions, with total holdings increasing by hundreds of terabytes each day. Similarly, the European Space Agency's Copernicus program generates on the order of petabytes per year from its Sentinel satellites. These numbers can be expected to rise significantly over the next decade, given both new missions and advances in sensor technologies.

Barriers to discovery, access efficiency, and interpretability remain despite significant investments in Earth data solutions. The vastness of these archives often overwhelm traditional search and analysis methods, rendering information effectively inaccessible or too expensive to analyze at a global or regional scale.

Current EO practices enable asset discovery by indexing data by ancillary details in the metadata (e.g. lat/long boundaries, time range, spectral properties) rather than the semantic content of the data itself. Users seeking specific phenomena or trying to gain deeper insights thus have to constrain their searches over these massive archives by cross referencing metadata (e.g. avoiding cloudy images over their area of interest) followed by exhaustive analysis of those results. Every person running global analysis has to read every pixel of data—potentially across multiple timesteps—even if they are only interested in finding wildfires. This presents a financial barrier for would-be users and dramatically limits the effectiveness of resource-constrained researchers and institutions. Since global analysis is challenging, users end up taking short cuts and artificially spatially and geographically constraining queries, potentially introducing bias and missing important results.

The Need for AI Integrated Solutions

> *The volume and exponential growth of digital data and of the ability to mine and generate those data provide rich opportunities for progress. This growth has led to **quantitative** change in the way research is conducted. Pairing advances in artificial intelligence (AI), computing, and automation of laboratories and observations can also lead to a **qualitative** step change.*
>
> Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop. National Academies of Sciences, Engineering, and Medicine.

Using AI to improve data discoverability, as suggested in this whitepaper, is strongly aligned with NASA's goals. The most recent recommendations of the Applied Sciences Advisory Committee (ASAC), which provides advice and recommendations for NASA's Earth Science Division, published on August 19, 2024, include:

- "Developing a Comprehensive Strategy for GeoAI"
- "Supporting Data Discoverability, Accessibility, and Utility for End Users"
- "Enhancing Private Sector Collaboration"

The recommendations note an urgent need for NASA's Applied Sciences Program to "*create a strategy for incorporating GeoAI into its programs*" and recognize that "*ensuring [the] accessibility and usability [of NASA-produced data and analysis] is crucial for empowering end users with the tools they need.*" The proposal for an EO vector embeddings catalog laid out in this whitepaper provides a roadmap to make progress on each of these recommendations: it describes a way for NASA to incorporate advances in GeoAI and AI in general into its product offerings; it provides a plan for productionizing a data discovery feature that would greatly help users find and make use of NASA's vast data archives; and it identifies an opportunity for NASA to collaborate with the private sector to solve a challenging problem.

More recently, in May 2025, NASA's Chief Science Data Officer, Kevin Murphy, presented on the need for "Accelerated Discovery" and NASA's vision of AI for Science at the ESA-NASA International Workshop on EO Foundation Models.

He outlined the “AI for Science Goals” as the following:

1. “Accelerate and Advance Scientific Discovery - unlock new scientific insights, enabling faster and more accurate discoveries.”
2. “Empower and Embed the Scientist - Support every phase of scientific method and create models with the power of NASA scientists”
3. “Foster High Payoff Innovation - Pursue cutting-edge AI applications to push the boundaries of scientific research and exploration”

The creation of a catalog of vector embeddings is, again, directly in line with these goals for Accelerated Discovery and provides a democratizing tool that lowers the barrier for extracting insights from data.

What are Vector Embeddings?

Vector embeddings are compact numerical representations that summarize complex data—such as text, images, or audio—in a form computers can easily analyze and compare. Imagine reducing an entire satellite image into a concise vector in a space that captures all the essential features of such images; things like (but not at all limited to) patterns of vegetation, water, and urban areas. These numerical summaries allow researchers to quickly compare and find relationships among large amounts of data.

A useful analogy is to consider a vector embedding to be the “DNA” of a satellite image. In this sense, the embedding is a compressed representation encoding the essential information contained in the original image.This makes embeddings ideal for exploring massive archives, identifying similar images, or finding subtle trends without manually inspecting every item.

The following sections discuss practical ways embeddings can be used and why centralizing their creation and storage is strategically beneficial. Readers interested in greater detail about what embeddings are should refer to Appendix I: Vector Embeddings.

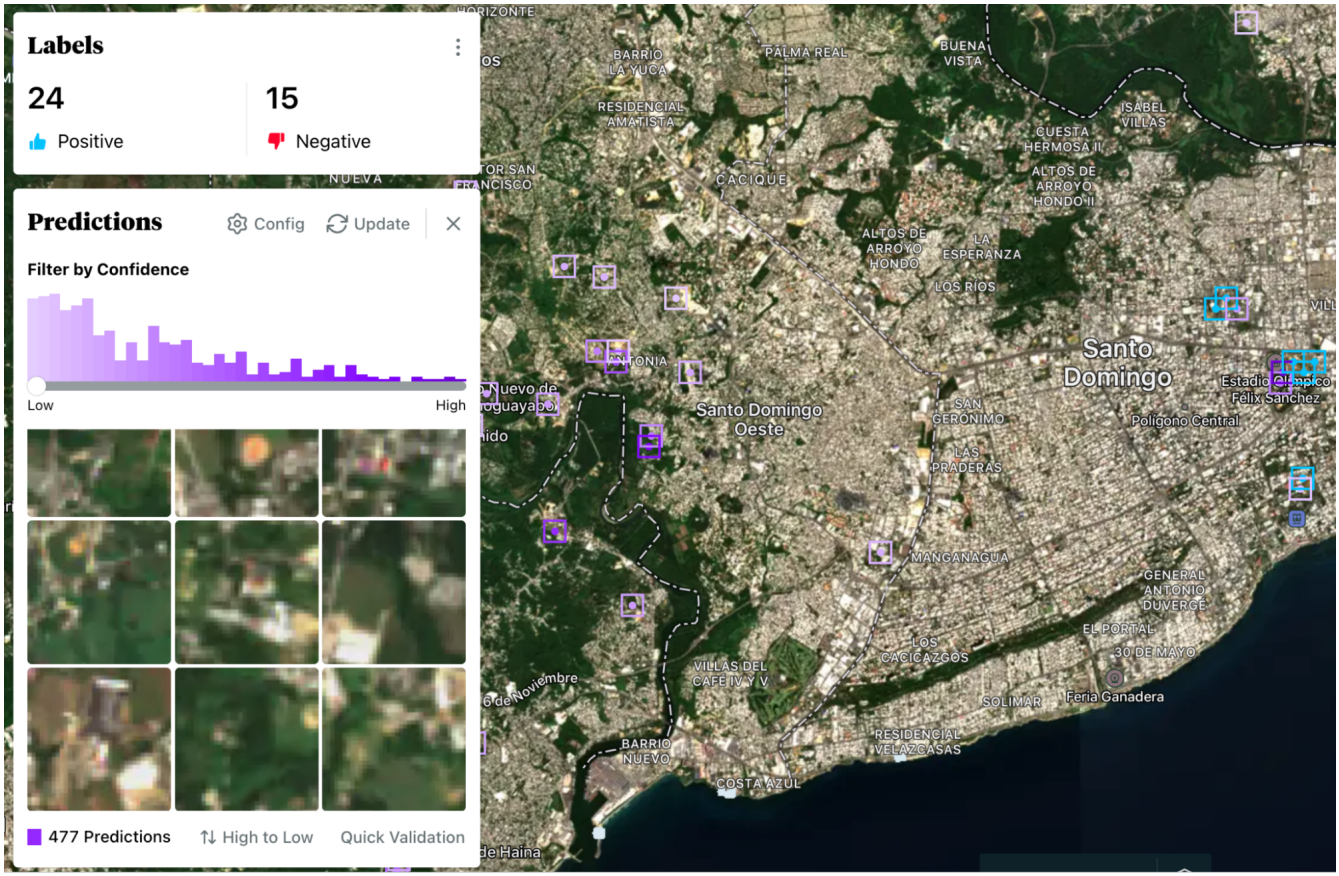
Vector Embedding Uses

Above, we discussed how vector embeddings can be used to compress highly complex data. Here, we focus on some of the uses for these compressed representations when applied to multiband, EO imagery. In each of these cases, these small lists of numerical values help to capture the essential character of the observations from which they are derived. In downstream analysis, vector embeddings can be used in lieu of, or as a complement to, actual EO imagery making them an extreme form of analysis-ready data.

In this section, we look at some exciting use cases that geospatial vector embeddings would enable if they were made widely available and easily accessible. We consider embeddings generated by both image-only foundation models as well as vision-language foundation models (“text-aligned embeddings”).

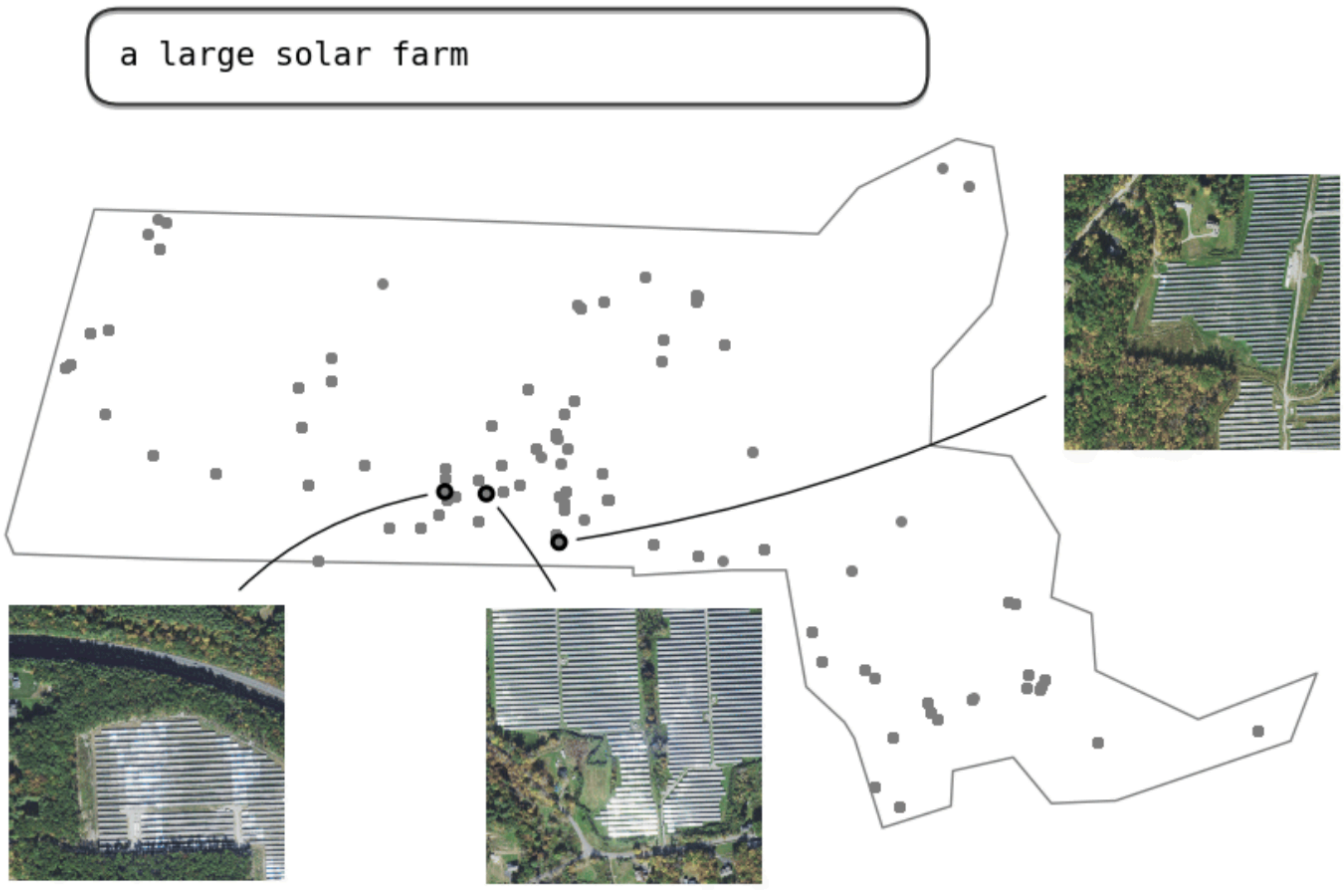
1. Searching for EO data based on semantic content

One of the most basic things that vector embeddings enable is content-based data search. This follows straightforwardly from the fact that embeddings capture the essential meaning of the original data, so the embedding for an image of a solar farm will be similar to the embeddings of other images of solar farms even if they are located in different places. This can be used to build a search functionality where a user can easily look up images similar to one or more query images. The example below shows an existing demonstration of this kind of search functionality.



A demo from Earth Genome showing how vector embeddings can be used to find similar images. Source: screenshot captured from https://app.earthindex.ai/

We can go even further if the embeddings happen to be text-aligned (some models produce text-aligned embeddings, some do not). In such cases, they can be used to build a search engine for EO imagery in which users can find images based on a natural language description of the contents of the images e.g. the query “a large solar farm” will find images containing large solar farms. See Element 84’s blog, *Building a queryable Earth with vision-language foundation models*, for a more thorough exploration of this topic.

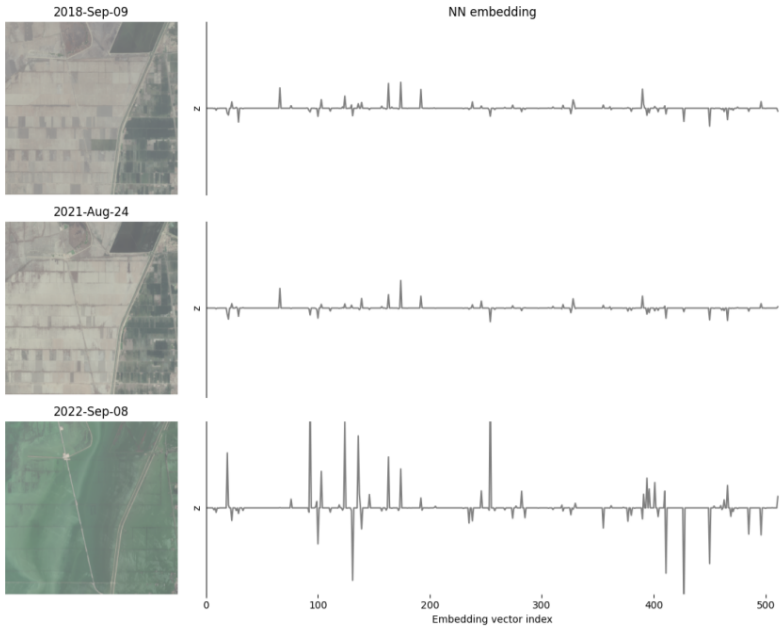


Using text-aligned vector embeddings to build a search engine for EO imagery, where you can search for images matching a textual description such as "a large solar farm". Source: https://element84.com/machine-learning/towards-a-queryable-earth-with-vision-language-foundation-models/

2. Detecting anomalies and long-term changes

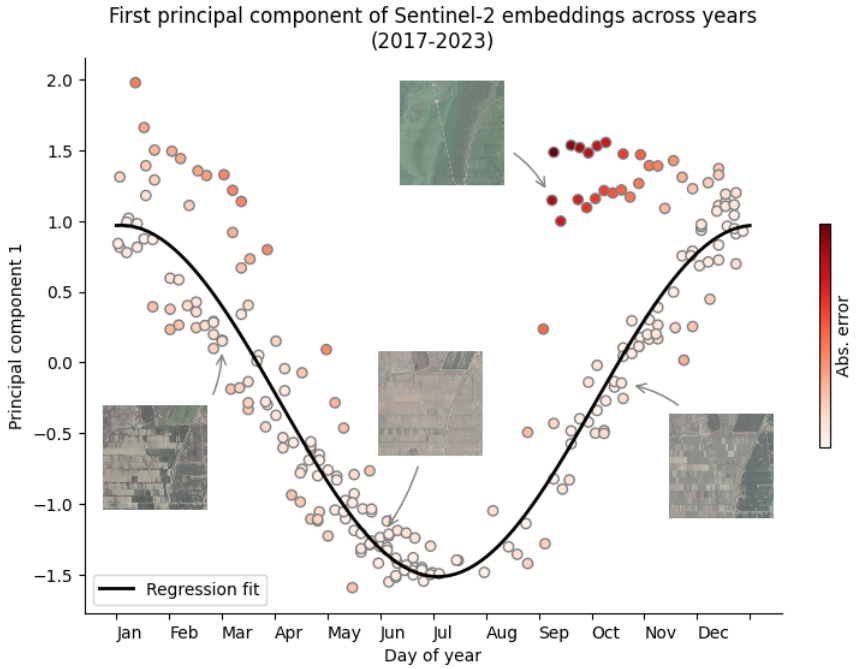
Above, we discussed how vector embeddings, as compressed representations of the contents of images, can be used to determine if images are similar to each other based on the similarity of their respective embeddings.

This provides a natural way to detect changes in the satellite imagery of an area: if the latest image capture differs significantly from what the area usually looks like, its vector embedding will also be very different. An example of this can be seen in the image below where the vector embedding deviates significantly from the norm when the area is flooded. The advantage of this, over more familiar approaches like image differencing or using change detection models, is that this does not require downloading the actual image data which may be orders of magnitude larger than the embeddings. We can tell that the *semantic contents* of the images have changed without ever accessing the actual pixels of the images at all.



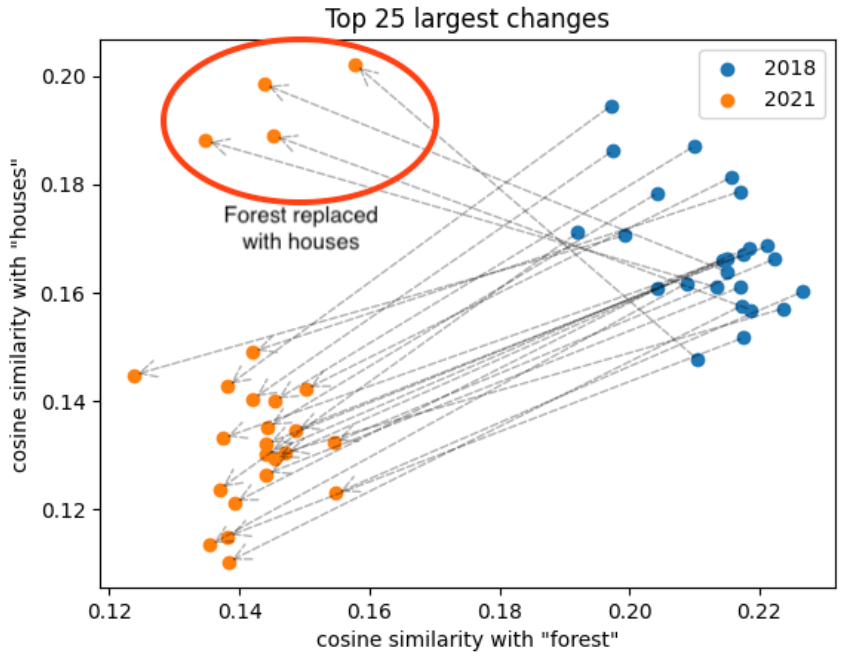
Similar images have similar embeddings while dissimilar images have dissimilar embeddings. Source: https://element84.com/machine-learning/exploring-unsupervised-change-detection-with-sentinel-2-vector-embeddings/

In practice, the analysis of the historical vector embeddings of a given area potentially offers much richer insights than mere anomaly detection. For example, as shown below, if the embeddings are compressed further into a single value using Principal Component Analysis (PCA) and visualized as a seasonal plot they reveal not only anomalous phenomena (such as flooding) but also the natural seasonal trend of how the appearance of the area varies over the seasons. See Element 84's blog, *Exploring unsupervised change detection with Sentinel-2 vector embeddings*, for a more thorough exploration of this topic.



Seasonal variation in the first principal component of Sentinel-2 embeddings (2017–2023) reveals an annual cycle driven by environmental changes. Source: https://element84.com/machine-learning/exploring-unsupervised-change-detection-with-sentinel-2-vector-embeddings/

Again, text-aligned vector embeddings provide some novel utility. Not only can we detect that changes have occurred but we can also figure out what the nature of each change is. Further, we can in fact *search* for specific kinds of changes. Going back to the example above of a search engine for EO imagery, not only would it allow users to search for “forests” or “houses”, but it would also allow them to search for “forests that have been replaced by houses”. In the visualizations below, we see how vector embeddings enable such searches. See Element 84’s blog, *Finding Changes on the Earth with Natural Language*, for a more thorough exploration of this topic.



Scatter plot showing changes in cosine similarity of satellite image embeddings to the terms "forest" and "houses" between 2018 (blue) and 2021 (orange). The circled points represent cases where forested areas in 2018 were replaced by residential housing by 2021, as reflected in an increase in similarity to "houses" and a decrease in similarity to "forest." Gray arrows indicate the direction and magnitude of change over time. Source: https://element84.com/machine-learning/finding-changes-on-the-earth-with-natural-language/

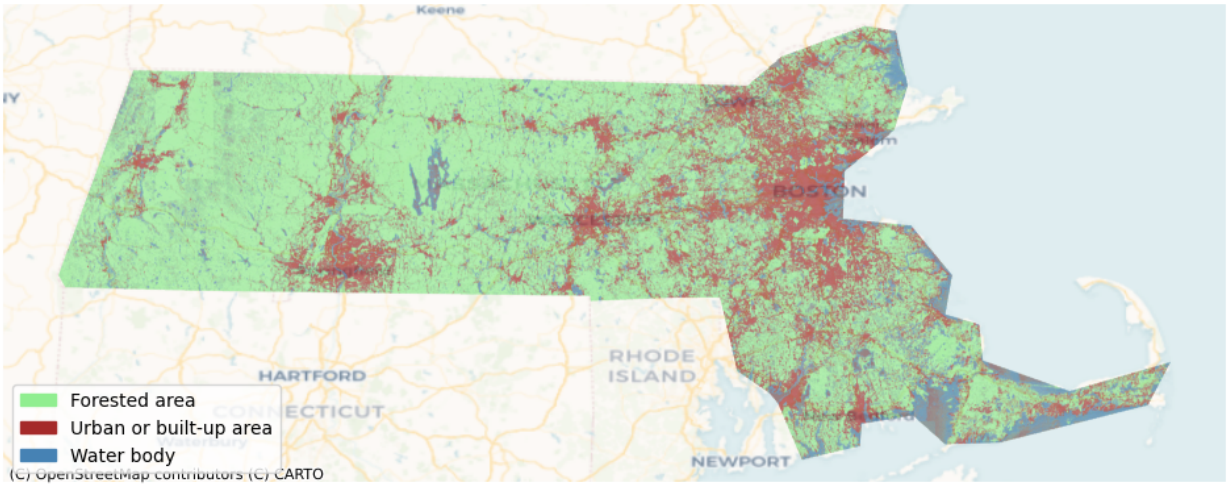
Satellite image pairs showing examples of forest-to-house transitions from 2018 to 2021, corresponding to the circled points in the scatter plot above. The 2018 images (left) show dense forest, while the 2021 images (right) reveal residential developments or construction activity, explaining the increased similarity to "houses" and decreased similarity to "forest." Source: https://element84.com/machine-learning/finding-changes-on-the-earth-with-natural-language/

3. Classifying EO Imagery

One of the most common uses of EO imagery is to generate land cover maps to analyze and monitor how land is being used. This is usually achieved by training specialized classification models (today, these tend to be deep neural networks) that classify patches of EO imagery into one of a fixed number of categories such as “forest”, “urban”, “desert” etc.

Widely available vector embeddings for EO imagery would super-charge this workflow in two ways:

1. Direct classification of embeddings themselves:
   Relatively simple classification models could classify the vector embeddings directly rather than the much larger and more complex multiband EO imagery. Instead of downloading large amounts of EO imagery, preprocessing it, and then passing it to a vision model, users would be able to download just the corresponding vector embeddings, which would be much smaller in size, and pass them directly to a simpler embedding classification model, resulting in network and compute cost savings. These same cost savings will also apply every time the model is used for inference in the future.
2. Heuristic classification by embedding proximity:
   Classification without spending time and money on training a classifier should also be possible. This approach comes in at least two flavors.
   a. Extremely strong examples of forested areas can be used to generate a ‘forested’ coordinate while examples of urban use could be used to generate an ‘urbanized’ coordinate. Classification would then involve looking to see which coordinate new observations are more proximal to.
   b. Using text-aligned embeddings is perhaps simpler yet. Users would be able to directly classify imagery into arbitrary categories with natural language and generate classifications using vector similarity scores. An example of this can be seen in the image below.

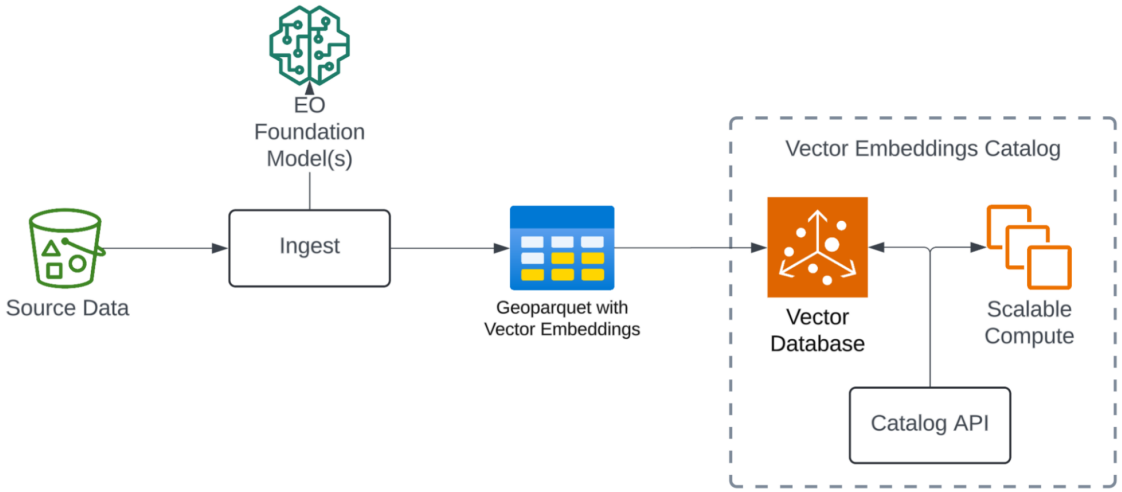


A land cover map generated directly from text-aligned image embeddings without training a land-use classification model.

A Vector Embeddings Catalog

We have made the case for how vector embeddings can be useful when available at scale. This section describes the system that would actually enable their widespread use. For this, we recommend building and maintaining a Vector Embeddings Catalog. Unlike a standard data catalog which allows filtering data based on associated metadata, a Vector Embeddings Catalog would additionally allow filtering data based on its semantic content. For example, a user who might have used spatial and temporal constraints to look up satellite imagery in a STAC Catalog (Spatio-Temporal Asset Catalog) would now be able to narrow down the search even further by constraining the *content* of the imagery e.g. "forest".

Below, we describe the high-level architecture of such a catalog.



Ingest

This is the data transformation step that produces the embeddings. This will involve gridding the data and then running inference on each grid tile (or "chip") using an EO foundation model. The outputs of the model (the embeddings) will be stored along with associated metadata in the cloud using a cloud-optimized data format such as GeoParquet or Zarr. The associated metadata for each embedding will include, among other things, the bounding box of the chip used to generate it which will allow users to easily visually inspect the source images for embeddings. Initially, this pipeline will have to be run over a selected portion of the archive of data to populate the catalog, but once that is done, it will only need to be run periodically as new data becomes available.

There are several critical open questions here that do not have established solutions such as which foundation model to use and how exactly to grid the satellite data. The choice of model will depend on factors like the quality of the model (as measured by appropriate ML benchmarks), the modality and resolution of the data, the number of dimensions in the embeddings, and so on. The choice of the gridding strategy will include considerations like the size of the chips, the amount of overlap between chips, possible reprojection of georeferenced data, and so on. Another common concern with generating embeddings is what happens when a better model comes around. Will the entire catalog of embeddings need to be regenerated?

The answer, for now, is yes. Embeddings from one model are not directly compatible with those from another model, although this might change in the future as overcoming this limitation is an active area of research. As such, a versioning scheme for the embeddings would also need to be implemented. See the *Technical Risks and Mitigations* section from more discussion on this.

Vector Embeddings Catalog Components

Catalog API

Here we discuss possible functionality that the catalog's API might offer users, some of which go beyond content-based search:

- Find items with embeddings most similar to a given embedding vector.
- Find regions with a temporal profile matching a given sequence of embeddings. E.g. regions that were forested in 2020 but are urban in 2025, or regions that are vegetated in the summer but barren in winter, etc.
- Find items matching a small, known set of attributes e.g. cloudy, forest, water, urban, etc. This will match based on pre-computed embeddings corresponding to each of these attributes.
- Find items matching a free-form text description, if the embeddings are text-aligned or if there is a model to map text embeddings to image embeddings.
- Find regions with the most or least stable embeddings over some period of time. E.g. for least stable: a city destroyed by an earthquake or a wildfire; most stable: a water reservoir whose water level has stayed nearly constant.

All of these can be implemented using a vector database. However, in some cases or at a future time, it might be desirable to offer more powerful functionality where users can upload positive and negative samples for what they are looking for and receive more precise classification results as opposed to mere vector similarity scores. This can be beneficial because vector similarity alone is sometimes not enough to distinguish finer details in the data. Such a feature would require some form of scalable compute to train light-weight classification models on top of embeddings.

Vector Database

Traditional database types and indexes are ill-suited to storing and searching for vector embeddings, especially at the scale of millions or billions of records. Instead, one must turn to *vector databases*. Vector databases implement storage indexes and search algorithms that are optimized for efficiently comparing vector embeddings. They have exploded in popularity due to the widespread interest in AI approaches like Retrieval Augmented Generation (RAG) that make extensive use of vector embeddings. Using vector databases will allow us to leverage cutting edge advances in the wider AI community for EO. In choosing an appropriate vector database, one will have to consider tradeoffs involving cost, scalability, and functionality. See the blog post *3 Billion Vectors in PostgreSQL to Protect the Earth* for a detailed case study of precisely this problem in the context of Earth Genome's Earth Index platform.

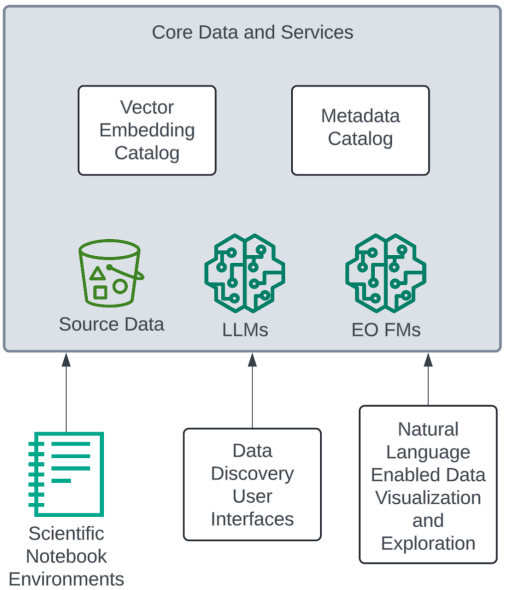
Scalable Compute

Being able to quickly compute vector similarity at scale enables a lot of use cases as seen above, but vector similarity on its own may not be the best tool for use cases where finding every match is essential. A similarity search using the embedding for a satellite image of a construction site, for example, might (in addition to valid matches) return false positives like images of mining sites. Moreover, similarity search results usually don't have an obvious similarity-score "cutoff" marking the point where correct matches end and false positives begin, necessitating manual visual inspection by the users. One way to address these limitations is to allow the user to provide multiple samples, both positive and negative, of what they are and are not looking for. These can then be used to train a very light-weight model, such as a small Multi-Layer Perceptron (MLP) that would take in embeddings and classify them. In machine learning terms, this would be equivalent to training a classification head on top of a frozen encoder. This approach would not only be more powerful but would also output classification scores which are more easily interpretable.

Being able to train models on demand like this is a powerful feature but would add additional complexity and impact costs. Therefore, it might be prudent to defer this functionality until future iterations of the catalog.

Part of an Interoperable Ecosystem

It is natural to imagine such a catalog enhancing existing scientific workflows that use programmatic interfaces. But a catalog like this could also power additional tools, applications, and services – exposing new search capabilities to a wider audience. These could include data discovery user interfaces like NASA's Earthdata Search, as well as LLM-driven workflows that allow users to query data through natural language. Together, these would constitute a rich ecosystem of interoperable tools and modules that bring together EO data and the latest advancements in AI to allow both expert and non-expert users to find answers to their most important scientific questions.



A diagram illustrating the ecosystem of catalogs, data, and AI models that enable accelerated discovery and science.

Why NASA Should Create a Vector Embeddings Catalog

We've already identified ways in which a catalog of vector embeddings will be useful, but why does it make sense for NASA to create an embeddings catalog, and why now?

An Urgent Need for Accelerated Solutions

Current approaches to Earth observation data analysis, while previously effective, are insufficient for today's challenges. As noted in the Earth Science to Action strategy and Earth Science at Work, "with increased frequency of drought and flooding, these tools, while previously effective, no longer suffice". NASA must make "changes and upgrades to remain successful under [these] new conditions" and "help Americans respond to challenges and societal needs".

Vector embeddings offer a complementary solution to existing approaches like fine-tuned foundation models. While fine tuning is effective for creating models focused on a specific task, such as land classification or object detection, it requires specialized knowledge, lots of training data, and compute resources. Vector embeddings, in contrast, can be:

- Generated once and used for many different purposes
- Accessed through browser-based interfaces by non-experts
- Applied to rapid searches across billions of records.
- Used to quantify impacts of natural events in near real time.

By providing EO FMs and cataloged vector embeddings NASA will have created a solution that combines the benefits of both approaches.

NASA's Unique Position vs. Community Efforts

While other organizations have made progress with vector embeddings, their efforts have limitations which NASA is uniquely positioned to address.

ESA Generated Vector Embeddings from Sentinel Data

The European Space Agency (ESA) has created 40B vector embeddings from their Major TOM dataset from Sentinel data. This is a great start but ESA's Major TOM is a *single time subset* of Sentinel data with a focus on experimentation and model training. Additionally, the embeddings are not cataloged which means that they can't be used easily for fast search.

Clay Foundation Generated Vector Embeddings from NAIP and Sentinel

The Clay Foundation generated vector embeddings for both NAIP (2010-2021) and a selected subset of Sentinel-2 data but without consistent tiling schemes or global coverage, limiting their utility for systematic change detection.

Earth Genome Created a Vector Embeddings Catalog

Earth Genome, a non-profit organization, created a catalog of vector embeddings for their application, Earth Index, which allows assigning labels to known locations which are used to find similar areas. This has already been successful for many use cases like finding unmaintained quarries, uncovering illegal mining, and tracking narco airstrips. Non-profit organizations like Earth Genome can create their own purpose built vector embedding catalogs but these kinds of organizations often have to make hard choices on scale and features due to their limited resources. Earth Genome is only able to publish yearly embeddings generated using an annual cloud free composite. This limits the usability for using the catalog to track changes such as damage from storms or wildfires. It would be cost prohibitive for them to publish embeddings on a smaller temporal scale.

A NASA-Led Effort

While the existing efforts of other organizations are valuable, their limitations highlight a gap NASA could fill. NASA has a full archive of multi-mission, multi-decadal data with a full range of sensors beyond optical imagery. A vector embedding catalog would become part of the foundational data infrastructure aligned with NASA's core mission to maximize scientific return and make Earth science data accessible. If NASA builds it, they can ensure that it is focused on the most important use cases for the United States. They can ensure that it is truly open with free access and with scientific neutrality in design choices.

NASA is also in a position to integrate generation of vector embeddings into their existing data processing pipelines. This single additional processing step would allow the creation and indexing of vector embeddings into the catalog in near real time. This would further expand the number of applicable use cases like being able to use vector embedding differences to automatically detect evolving conditions for wildfires.

By leveraging innovative ideas like this from academia and the private sector NASA will be helping users understand impacts of natural hazards, support national security through enhanced Earth observation capabilities, and ensure that the solution provides trusted information backed by science.

Infrastructure Cost Reduction and Democratization

Vector embeddings offer dramatic reductions in data storage and transfer requirements compared to raw EO imagery. Terabytes of imagery can be represented as compact numerical summaries measured in megabytes and gigabytes, while preserving the semantic content necessary for many analyses. This enables researchers to quickly access and analyze data without downloading massive datasets, greatly reducing redundant I/O operations and minimizing transfer costs and compute costs for researchers.

Terabytes of imagery become compact numerical representations measured in megabytes and gigabytes. Take Prithvi-EO-2.0 as an example, a single input image will have 224 width x 224 height x 12 band fields. Each such field will have 4 bytes with the standard floating point precision of 32 bits. This yields 12 × 224 × 224 × 4, which is 2,408,448 bytes or roughly 2.41 MB. The embedding for this input is comparatively tiny; 768 parameters x 4 bytes per parameter means 3,072 bytes or roughly 3 KB. Going from 2.3 MB to 3 KB is almost 784x compression. **This means that for a 1 TB of imagery (i.e. about 415,000 224×224×12 images), just about 1.28 GB of embeddings will need to be stored or a compression ratio of 784x.** Quantization, an approach of reducing the size per parameter, can further reduce the size without sacrificing much of the encoded information.

These efficiency gains have profound implications for democratizing access to EO data analysis. Currently, sophisticated EO research often requires substantial infrastructure and computational resources, creating significant barriers for smaller research groups or businesses. With pre-computed embeddings readily accessible, analysts no longer need powerful local computing resources to perform preliminary assessments, validation, or exploratory studies. Smaller institutions can thus participate meaningfully in EO research that would otherwise be cost-prohibitive.

For NASA, this aligns perfectly with goals of making Earth science data accessible for research and societal benefit. By reducing the computational threshold required for sophisticated analyses, a vector embeddings catalog would foster broader participation, innovation, and democratization of data-driven discovery.

Research Catalyst Effect

Additional value of a vector embeddings catalog emerges through network effects—the principle that a resource becomes increasingly valuable as more people use and contribute to it. A NASA-led vector embeddings catalog would accelerate research not just by making analysis more efficient, but by providing a standardized foundation that enables new types of scientific collaboration. When embeddings become direct subjects of study, a centralized catalog offers researchers a wealth of standardized data to investigate the semantics of what these embeddings encode and their relationships to observable Earth phenomena.

Standardized embeddings create a shared scientific vocabulary that enables cross-fertilization among research efforts. Instead of each research group generating their own embeddings using different models and approaches, a common foundation allows researchers to build directly on each other's work, compare results meaningfully, and develop more sophisticated analytical techniques.

This standardization effect extends beyond academic research. By offering consistent, accessible embedding resources, the catalog would empower the broader technological community to rapidly prototype, test, and deploy novel applications. Developers could build sophisticated embedding-aware applications without first having to solve the fundamental challenge of generating reliable, high-quality embeddings at scale.

Technical Risks and Mitigations

Limitations of vector embeddings and vector similarity

Given the fact that vector embeddings represent a lossy compression of the source data and the coarse-grainedness of vector similarity measures, it is possible that similarity search does not work well enough for most use cases.

As discussed earlier in this paper, there are limits to the kinds of analyses one can do using similarity searches on vector embeddings alone. For example, this kind of analysis might not be sufficient to differentiate between construction sites and mining sites. This kind of limitation is a combination of the limitations of:

- the foundation model (i.e. it is unable to generate sufficiently distinct embeddings for the two phenomena), as well as of
- vector similarity measures like cosine similarity (i.e. they may be unable to tease out small differences between high-dimensional vectors).

The first of these can be mitigated by extensively testing and benchmarking the quality of embeddings produced by different foundation models during the process of selecting a foundation model. This will ensure that there are use cases where the embeddings are proven to work well before we undertake the task of generating them at scale.

The second one may be mitigated in a future iteration of the catalog by implementing a feature that allows users to perform more fine-grained searches by training small classification models in the background. It may be further mitigated by the fact that users can simply download the embeddings if they want to perform more sophisticated analyses that are not supported by the catalog API.

Choice of model and embedding longevity

Given the rapid pace of research in EO foundation models, choosing the optimal foundation model to use is difficult and it is possible that whichever foundation model we choose will be superseded by significantly better newer models soon after.

There are a number of EO foundation models that have been released in recent years, and more are coming out all the time, and will continue to do so, at least in the near future. Not all of them are interchangeable – some are trained on specific sensor modalities and resolutions while others claim to be more general; some work with individual images, while others work with time series of images; some have been trained using the additional external modalities of geographic coordinates or text, while others have not. Moreover, different models may excel on different subsets of benchmarks making it hard to pick out a clear winner.

All of these are important concerns that must be carefully considered when choosing a model to use and it is likely that the answers will vary depending on the target data archive. However, these risks are mitigated by the fact that having any embeddings at all, even if they are not objectively the best, is so much more desirable than having no embeddings that we should proceed with this effort despite the potential tradeoffs. Additionally, newer models usually represent only incremental improvements in benchmarks over previous models, so it is unlikely that the generated embeddings will be rendered obsolete in the near term. We recommend choosing a model from one of the current state of the art models and then revisiting that choice at regular intervals of 1-3 years.

Scaling for Large Archives

Given the vast scale of NASA data archives, creating a comprehensive vector embedding catalog of all NASA data that is also performant would be both challenging and expensive.

There are multiple embeddings per individual scene or input granule. That multiplicative factor means that a collection of tens to hundreds millions of granules would result in billions of vector embeddings. The size of the embeddings themselves would be much smaller than the input as noted below in our Infrastructure Cost Reduction and Democratization section but the sheer number of rows would be very large, especially if the initiative is successful and NASA wants to expand it to more collections and more data types.

These scaling risks might be mitigated by:

- Being selective about the number of years and collections to support initially. The initial offering might only support data from the most recent few years and might be gradually expanded to include older data over time.
- Optimizing the gridding strategy (i.e. how we break large scenes into smaller chips to create embeddings for) in a way that the number of chips scales reasonably. This would involve tuning, among other things, the chip size and overlap.
- Partitioning the embeddings into separate databases that are searched in parallel and their results combined. Each collection's vector embeddings are essentially a different database.

This also presents an opportunity to more thoroughly investigate promising approaches to storing, indexing, and querying vector embeddings at scale. These include:

- Using paid vector embedding database vendors.
- Using partitioning to separate collections of vector embeddings into separate databases.
- Storing embeddings in the GeoParquet or Zarr data formats with their associated database indexes which can be searched in parallel with serverless approaches.

This is an active area of investigation within the geospatial open source community where NASA can help make the push to find a solution. NASA is no stranger to large scalability problems.

A Plan to Get Started

While a catalog of vector embeddings presents a clear value, initial work should focus on identifying target use cases, quantifying success criteria, and mitigating risks.

Identifying Supported Use Cases

Identifying use cases is the most important step prior to the development of a prototype and a larger effort as it will drive specific tests to perform in a prototype. A catalog has many potential applications but the needs of your organization and users will have unique constraints. While a vector embedding catalog provides the ability to find similar features or perform change detection, it's important to understand the kinds of features that are expected to be found or types of changes you want to detect. For example, features that are smaller, like detection of individual buildings, will benefit from vector embeddings with smaller foot prints. Ideally, your vector catalog would support searching for variable size features so testing with multiple chip sizes will help identify the ideal size or sizes to use.

Create a Demonstration Application

An important vector embedding catalog selling point is its ability to help users. Facilitating adoption within an organization requires the ability to make somewhat abstract concepts like vector embeddings tangible to stakeholders. An interactive demo helps make the capabilities real for stakeholders and potential users and allows you to build a vision for the final product. Existing community efforts to accomplish a similar prototype have lacked elements necessary for practical usability. For example, some lack consistent tiling schemes, consistent and global coverage, or are not completely cataloged which compromises the potential for fast and easy search.

Create a Demonstration Application

There are a few important options to consider when creating a vector embedding catalog. These include:

- Identifying the source datasets that are relevant to the specific use cases and would be visible within the optical bands or measurements of the data.
- Selecting an appropriate EO Foundation Models that will be able to discern important features and translate them into the embedding space.
- Chipping/Tiling approaches as noted above.
- The actual vector database to use with open source options like the pgvector extension for Postgresql or OpenSearch, as well as commercial vector databases.
- The methods for computing vector similarity to determine which embeddings are similar or dissimilar from an input embedding.
    - Vector databases support multiple methods for computing similarity such as L2, L1, and cosine similarity. Additionally you may want to evaluate training small models that can work directly with embeddings.

Evaluating these options and how well the use cases can be supported requires creating a suite of evaluations with metrics to measure success. You should develop evaluations that will measure the accuracy of results. As an example, for use cases that require identifying new construction you should select a set of input data that includes known cases of new construction. Then evaluate how many of these cases are correctly identified using vector embedding similarity, how many are missed, and the number of incorrect identifications.

Some use cases might not be well supported and some will have limitations. For example the accuracy for vector embedding cosine similarity may not be appropriate for very early detection of certain plant diseases or early drought detection. Those may be more appropriate for use with a fine tuned foundation model. This will be dependent on the input data, the model selected, and the choice of similarity metric. Running evaluations help identify the limits here.

Estimating Costs

Costs for creating and operating a vector embedding catalog can be broken down into three categories.

- **Ingest Costs** - The cost to run the model on input data and generate the vector embeddings
- **Development Costs** - The engineering resource time to develop the evaluation tests, the ingest process, and use of the vector embedding catalog.
- **Operating Costs** - The ongoing costs of storing vector embeddings and running the vector embedding database and other infrastructure.

These costs will vary based on the size of the input data, model selected, and other factors including time spent optimizing these processes. It's important to note that initial estimates often tend to be overly conservative for ingest costs while underestimating development costs. Development of a prototype will ensure that you have realistic numbers for estimating the cost of the final system. We've developed initial estimates for processing the entire Sentinel 2 catalog but these can vary quite a bit with the tiling scheme that is chosen.

Conclusion

In this whitepaper, we have presented the case for NASA to build a centralized vector embeddings catalog for EO data that will enable users to more easily extract insights from NASA's vast archives of data and is strongly aligned with NASA's declared goals of accelerating data discovery and accessibility and democratizing data access.

We have demonstrated that dense vector embeddings derived from modern EO foundation models efficiently capture land-cover semantics, significantly outperforming traditional pixel-level methods in terms of accuracy, speed, and storage efficiency. The empirical evidence presented clearly positions vector embeddings as effective tools for addressing the bottlenecks of petabyte-scale EO archives.

Treating these embeddings as first-class assets offers a path to accelerated discovery. Curated in an open, versioned catalog, there are benefits across a number of dimensions:

- **Democratized science and analytics** — Capturing all of the relevant information with a fraction of the footprint of normal EO data means that individuals and institutions lacking large reserves of capital have fewer barriers to entry.
- **Accelerated Science** - Instead of spending weeks downloading and processing terabytes of imagery to identify relevant datasets, scientists can query a catalog in seconds to find exactly what they need.
- **Standardization and communication** — A single, centralized strategy enables reproducibility and communication which benefits scientific progress and adoption by industry.
- **Reduced infrastructure costs** — At ≈700 × compression relative to raw imagery, sometimes better, researchers retrieve kilobytes instead of gigabytes.

We have additionally outlined a practical architecture for making these embeddings broadly accessible. Based on these findings, we recommend the following phased program:

1. Prototype ingest & similarity API
2. Evaluate cost and accuracy against reference tasks
3. Scale to the full historical archive with versioned embeddings
4. Release a public API and governance roadmap

Executing this program positions NASA as the neutral steward of a global EO-embedding standard while enabling government, academic, and commercial stakeholders to build novel applications on a shared, scientifically grounded foundation.

For more information about our team's perspective on the subject of a centralized vector embedding catalog for EO data, please reach out to Element 84 directly on our contact us page.

Acknowledgements

We'd like to thank the Earth Genome team, Dan Pilone (Element 84), Julia Signell (Element 84), and Sara Mack (Element 84) for their support in providing input and reviews for this paper.