Sisu Technical Whitepaper

How Sisu gets the facts

Diagnosing critical business metrics using complex data at scale

Over the last 10 years, the ability to capture structured data across an enterprise has grown exponentially, and the opportunity for both analytics and operations teams to benefit from this accumulated knowledge has never been greater. Unfortunately, most companies capture far more data than they can effectively use, and even the most data-savvy organizations don't have enough resources to tackle anything beyond their highest priority questions.

That's where Sisu comes in. An operational analytics platform, Sisu helps companies diagnose the drivers behind their most important metrics, in real time and using all their structured data. Instead of relying on descriptive dashboards and after-the-fact analysis, Sisu continuously assesses information as it arrives to help answer a business's toughest questions.

To accomplish this, Sisu relies on a combination of automatic feature engineering, model selection, and personalized ranking and recommendation. Proven at scale at Microsoft and Facebook, and building on years of research at the Stanford DAWN project, Sisu is an ideal platform for:

- **Diagnosing the factors responsible for driving business metrics**
- **Testing and exploring complex, high-dimensional hypotheses at scale**
- **Harnessing the latent potential in enterprise data**
- **Monitoring KPIs for actionable changes**
- **Automating the translation of analyst outputs to business stakeholders**

In this paper, we will detail Sisu's unique approach to operational analytics, how the platform ingests and analyzes large-scale, high-dimensional data, and the mechanics behind Sisu's ranking, relevance, and presentation systems. Through these methods, we'll illustrate why these kinds of analysis are challenging, if not impossible, for traditional business intelligence (BI) and analytics services.

Delivering the fastest time to why

The first step in tackling the challenge of diagnosing these complex metrics in an organization is to re-think the interface to large-scale data.

Today, when a business metric changes, it's up to analysts to manually dig into the data, specify and test hypotheses, and generate reports to capture their findings. This process is labor-intensive, slow, and fails to scale to the kinds of data found in modern data stores, which can contain hundreds of columns and millions of rows, representing a massive space of hypotheses to examine and assess.

This inefficient, error-prone process is reinforced by conventional OLAP engines and BI tools. These platforms require users to carefully select factors for analysis in advance, a process that's limited by the analyst's knowledge of the business and ability to manage multiple variables.

For example, an analyst might compute a cross-tabulation of sales by region, make, model, and customer age, either by issuing a SELECT-FROM-WHERE-GROUP BY query in SQL, or by using an interface like Tableau.

While effective at reducing the computational complexity of the task, this approach is subject to the bias of the analyst and will often exclude dozens of potentially interesting combinations of variables. Even interfaces to data that layer natural language functionality on top of existing OLAP queries ("select the sum of sales by region") are subject to these constraints. Knowing what question to ask is the biggest barrier to getting useful answers.

> **"Knowing what question to ask is the biggest barrier to getting useful answers."**

To address the speed and usability challenges created by conventional interfaces to data, we've completely re-built the analytics interface by taking an objective-based approach. For each objective or KPI, an analyst using Sisu identifies the metric or measure (e.g., column of interest) they wish to analyze. Subsequently, Sisu uses the other factors contained in the data (e.g., other tables and columns) to automatically search the space of hypotheses to rank and highlight the most relevant factors driving the metric.

A familiar declarative approach: Search engines

To draw a familiar parallel, look no further than the modern search engine. Today's search engines allow users to locate highly relevant web content, drawn from a corpus of billions of documents, with only a few simple search terms. When users search for "Big Apple" on Google, they don't need to tell Google that they're interested in learning about New York and not large fruit. Rather, Google uses the large amount of data available on the web to learn the semantic association between "Big Apple" and the city of New York. In addition, Google improves over time, learning from behavioral signals like clickthrough rates to improve over time.

While Sisu is not a natural-language search engine, it uses a similar high-level declarative query model that only requires users to highlight the metric of interest within their data, then automates the process of identifying the concepts (i.e., features, and derived features) within the data that are most important to the metric. This is made possible by the enormous amount of data available in modern data lakes, and by leveraging user interactions to continuously improve over time as new data arrives, and as users interact with Sisu's results.

This objective-based approach is a declarative means of data diagnosis. Instead of requiring users to specify the exact set of factors and to manually explore the enormous factor search space, Sisu abstracts the entire search process, including data enrichment, feature engineering, order of operations, hyperparameter tuning, and ranking. By starting with the objective, Sisu aligns the analysis with the analyst's goal, eliminates the need to hand-pick features for analysis, and dramatically reduces the time it takes to generate useful answers.

From objective to relevant facts: Diagnosing KPIs at scale

Once a user defines an objective, Sisu begins the process of querying the data, featurizing the inputs, selecting and ranking relevant facts, and ultimately presenting them to the end user.

In this way, facts capture relevant and actionable behaviors – both positive and negative – within a dataset. Facts can represent diverse phenomena including user demographics and behavioral cohorts, promotional and marketing campaigns, feature flags and product lines, and other groups captured in the input data.

As new data arrives, Sisu re-evaluates each objective created in the platform and returns a set of updated facts that reflect the latest changes. In addition, users can subscribe to a Sisu objective to receive push-based notifications as facts change.

Step 1: Querying and processing the data

For each objective created, Sisu evaluates structured data that is stored in tabular form – database tables, transaction-level details, dataframes, and structured events. Sisu provides standard connectors for a variety of common data warehouses and file formats, including Amazon Redshift, Google, BigQuery, Snowflake, PostgreSQL, and CSV.

While Sisu is not an ETL engine, it supports lightweight data transformations (e.g., joins and algebraic manipulation of columns) via SQL. Sisu accesses input data securely, via industry-standard Open Database Connectivity (ODBC) connectors provided by each major database provider. Sisu is not a storage engine, and with the exception of CSV uploads, the platform does not store input data within the platform. Instead, Sisu loads data, processes it to extract the facts, then discards the raw inputs.

As new data arrives, Sisu continues to ingest the data and performs mini-batch computation to update the set of facts over time.

What's a fact?

Sisu delivers results in the form of facts, or explanations of key phenomena affecting business metrics. Each fact consists of a set of factors (e.g., columns, like "age between 20 and 30" or "state = CA") present in the data, along with statistics that capture the impact on a user's objective.

For example, if an analyst configures an objective to reduce churn, Sisu might return facts similar to:

"Where `trial_duration is less than 7 days` and `device_type = iPhone`, `churn` is 22% lower than usual. This decreased `churn` by 4% overall."

"Where `promo_code = SMB AUG19` and `account_type = premium`, `churn` is 10% higher than usual. This increased `churn` by 7% overall."

"Where `acquisition_type = self-serve` and `discount > 20%`, month-over-month `churn` increased by 15%. This increased overall `churn` by 9%."

Step 2: Making the most of data: Data quality, enrichment, and featurization

Next, Sisu performs a series of automated transformations, or featurizations, to expose factors of statistical interest during the search.

The key challenge here is to identify high-quality statistical features that generalize across use cases and can be generated with minimal user intervention. In a conventional analytics environment, analysts often manually select these features, requiring hours or days of tedious and error-prone experimentation. At the other end of the spectrum, deep learning helps automatically engineer high-quality features but in turn famously suffers from a lack of interpretability – it is hard to tell when we can trust a model's predictions.

Sisu leverages experience from years of research and production deployments at Google, Facebook, and Microsoft, and derives features using a proprietary optimizer for data-dependent transformations that leverages structural properties of the input to determine the appropriate features to engineer. Sisu inspects properties of the input data, including the schema, data types, cardinality, and related tables found in the input source. Using this information, Sisu performs a range of transformations, including:

- **Computing bins** (i.e., discretization of) **continuous attributes with sufficiently high cardinality**
- **Computing conjunctive features** (i.e., higher-order features, or feature crosses)
- **Selecting relevant time ranges based on temporal features in the data**
- **Enriching common fields** (e.g., zip code) **with demographic, financial, and behavioral data**

These transformations are critical to result quality because many features found in tables are not statistically relevant in their input form. For example:

- Instead of treating individual ages like 18- and 19-year olds as separate, relevant behaviors may be found in specific age ranges, like 18-25 year olds and 26-35 year olds.
- Teenagers may have a high conversion rate overall, but that same demographic may convert poorly when exposed to a specific advertising campaign – a factor that is only evident when examining both age and referring campaign together.
- Analyzing the total number of purchases made over a customer's lifetime may hide effects present only for newly acquired customers who have only made a small number of transactions – instead, the number of purchases made in the last week may be more relevant.

Notably, all of Sisu's automated featurizations are designed to be interpretable by humans – typically, in the form of natural-language descriptions of the features. In contrast, many approaches to automated featurization lack this capacity for easy interpretability – for example, deep networks automatically learn a feature representation consisting of millions of parameters coupled via often nonlinear interactions. These deep networks excel at predictive tasks on unstructured data but result in feature representations that are highly complex and are difficult for humans to understand. By delivering explanations in the form of phenomena within the input data (i.e,. features and derived features), Sisu sidesteps the challenge of model interpretability, which **remains an open problem** even within the research community.

Step 3: Automated model and feature selection

Given this large set of candidate features of interest, Sisu performs a model and feature search to identify statistically significant factors both point-in-time and over time.

The key challenge at this step lies in the extremely large search space of facts and hypotheses to consider, often numbering in the hundreds of millions of total possible combinations. For example, simply throwing these raw feature vectors into a common analysis package like scikit-learn or XGBoost would lead to prohibitively expensive runtime due to the combinatorial explosion of higher-order features and, in some cases, redundancy across features. Similarly, performing an exhaustive CUBE or GROUP BY operation within a relational database would result in a prohibitively expensive computational cost and an enormous set of results to sift through. For reference, on production workloads, these approaches can take more than three days to complete a single query.

Sisu is purpose-built from the ground up to quickly search over the feature space. The platform combines both algorithmic improvements in the form of dynamic pruning and ranking of the search space and hardcore systems-level optimizations in the form of a custom Rust-based dataflow runtime and dynamic data encoding. The result is a feature selection engine that is over three orders of magnitude faster than commodity approaches and two orders of magnitude faster than the fastest research prototypes we have encountered.

Specifically, based on the characteristics of the KPI, Sisu trains several models to examine particular population-level statistics of interest. For continuous KPIs, this includes factors that affect the average, density, and extrema of a distribution, as well as shifts in these metrics over time. For categorical KPIs, this includes factors that affect the rate and distribution of discrete variables, as well as shifts in these metrics over time.

“More traditional analytics approaches can take more than three days to complete a single query.”

Sisu subsequently performs feature selection to assess the most relevant features for each statistical model. Conceptually, this is similar to **LASSO**-style feature selection, but requires several modifications for scalability and usability, including:

- **Hyperparameter tuning** (e.g., regularization) **to balance feature complexity and accuracy**
- **Dynamic programming to prioritize computation over the enormous feature space**
- **Low-level machine-friendly vectorization, columnar execution, and parallel dataflow**

This feature selection, or search process, is the most computationally-intensive portion of Sisu’s runtime execution. This is where we’ve spent hundreds of hours of R&D in algorithmic and systems engineering, building a cloud-native dataflow engine in Rust from the ground up. This engine outperforms our prior research and prototypes that run in production at companies like Microsoft and Google by orders of magnitude, and is instrumental in allowing Sisu to run at interactive speeds.

Step 4: Just the (most relevant) facts: Personalized ranking and relevance

Because Sisu's feature selection process frequently results in a large number of factors, Sisu must determine the most important and relevant factors for each individual use. Just like services like Netflix and Facebook learn relevance based on user feedback and engagement, Sisu learns a personalized model for each user that improves over time Sisu ranks the facts derived from customer data by several measures, including:

1. **Effect on population-level statistics**
   e.g., averages and extremes
2. **Time period of interest**
   e.g., point in time, week-over-week
3. **Prevalence**
   i.e., occurrence within population of interest
4. **Prior user engagement**
   e.g., fact saves, shares, scrolling, clickthrough
5. **Strength of association and confidence**
   e.g., odds and risk ratio
6. **Counterfactual impact and confidence**
7. **Number of factors and factor type**
   (e.g., discrete or continuous), factor name

By leveraging both explicit feedback from users curating, associating, and sharing individual facts as well as implicit feedback like scrolling and clickthrough, Sisu harnesses all available training data to inform ranking and relevance.

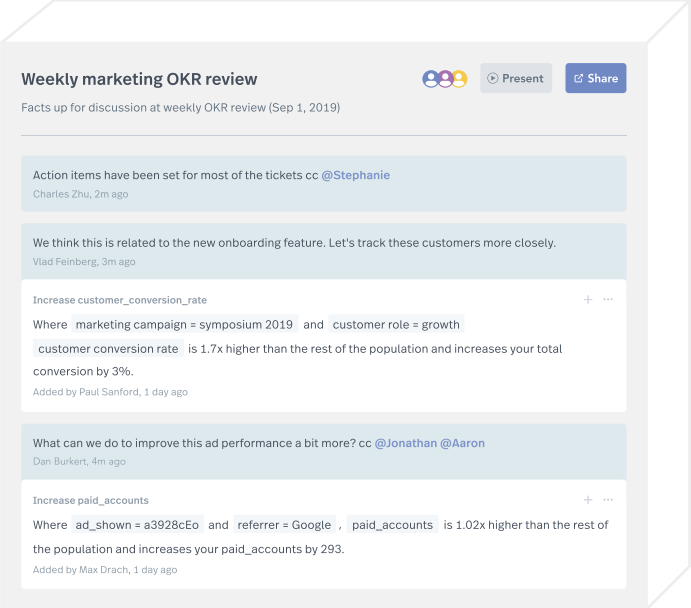
In doing so, Sisu prioritizes the scarcest resources in modern analytics: human time and attention.

Step 5: Letting the data speak: Presentation and sharing

Given the top facts for a given objective, Sisu provides an easy and interpretable explanation of the behavior and its impact on the objective of interest.

The key challenge is to precisely describe the statistical phenomena underlying the KPI shift without overwhelming a user with a barrage of low-level statistics like counterfactual impact and odds ratio. Many existing tools for feature exploration assign an arbitrarily-scaled and opaque "feature importance" score (e.g., "state=CA has importance 2.7") to individual features, and complex measures like R-squared value to assess the goodness of fit (e.g, "R^2=0.92"). These scores are difficult to interpret, and the resulting net effect ("so what?") on the business metric is left as an exercise to the user (or, most commonly, the analyst team who must go and compute additional supporting statistics over and over).

As a result, analysts and business users rarely collaborate within the same tools, and it can take many iterations to arrive at a precise statement of the business impact and opportunity for action supported by the data.



How Sisu lets the data speak: Digging into fact details

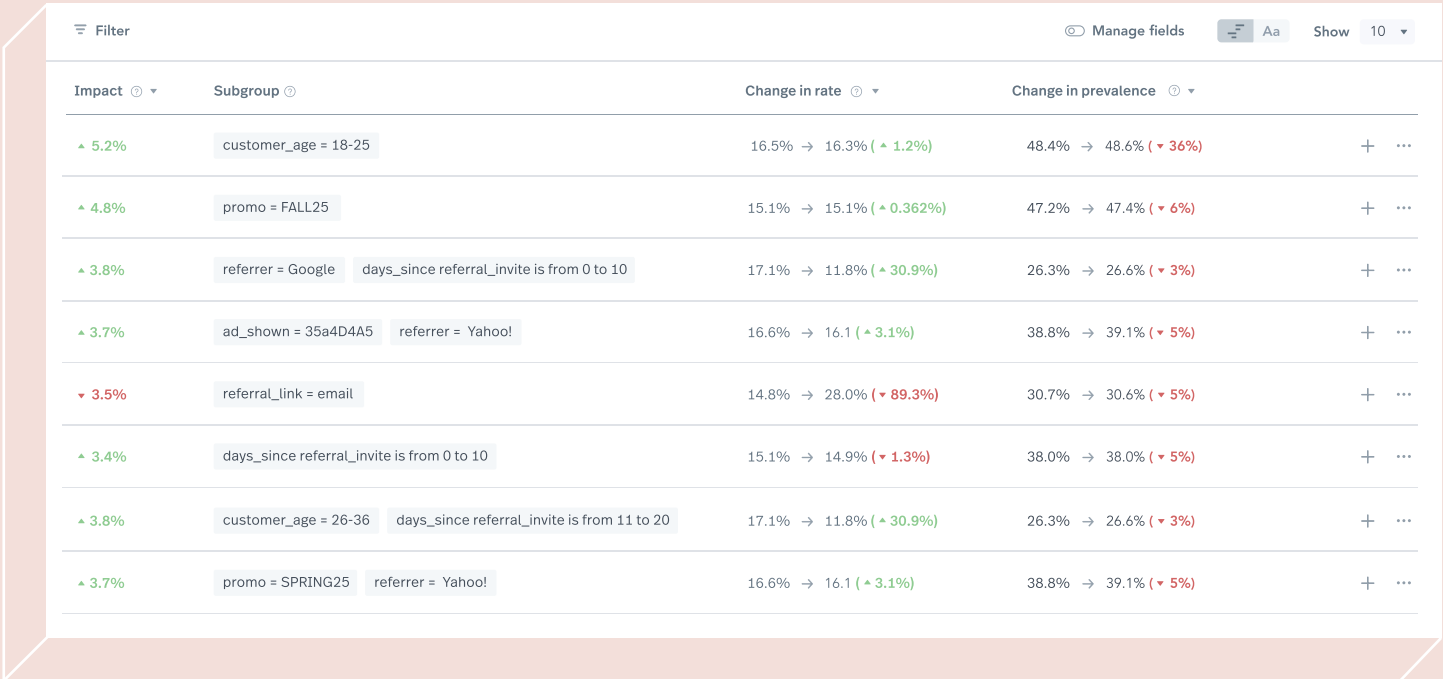
Natural language processing
A natural-language overview of each fact, describing the overall impact on the objective and why the fact was selected

Visualization
A custom visualization depicting the impact of the fact on the objective's metric – for example, for a fact describing month-over-month change, a time series highlighting the decrease relative to prior months, and the impact on the overall metric

Support statistics
Supporting statistics include prevalence (number of rows affected), counterfactual impact (the effect on the metric due to this population), and relative change (the difference between this cohort and the rest)

Related facts
Related facts that illustrate why this fact was chosen including:

- Fact marginals that indicate why a conjunctive fact such as "version 50 in California" is more relevant than a fact containing just "version 50 or just "California"
- Fact competitors that indicate why other features were not selected – for example "version 50 in New York," "version 50 in Colorado," and "version 51 in California"



Ongoing: Running continuously

While Sisu enables extremely fast answers to "why" on demand, it's even more valuable to proactively monitor the factors driving change to metrics over time. As a result, rather than requiring users to re-define and re-run analyses on-demand, Sisu continuously processes each objective as new data arrives. As a result, Sisu can automatically inform users about the most impactful changes reflected in their data.

To do so, Sisu allows users to subscribe to personalized notifications about their objectives, delivering push-based fact updates via email and Slack. Sisu's internal ranking and relevance-based measures determine what facts are most important to each user and when, ensuring that users never miss a key fact that impacts their objectives.

To enable this continuous analysis, Sisu incorporates several key techniques, including:

- **Mini-batch continuous dataflow execution**
- **Incorporation of temporal ranking and relevance measures, including result hysteresis**
- **Incremental model retraining to incorporate the latest training data from each user and their respective organization**

This ensures Sisu's customers never miss a key factor affecting their business, especially in dynamic environments with changing business dynamics, user preferences, and competitive environments.

About Sisu

Sisu is an operational analytics platform that helps businesses rapidly diagnose the factors driving changes to their businss. Sisu's ML-powered approach empowers anyone to get answers to their toughest business questions and makes complex analysis easy and accessible. Instead of relying on analysts playing detective with data, Sisu continuously monitors all an organization's information and automatically recommends meaningful facts in seconds.

The technology behind Sisu has been proven at scale at Microsoft and Facebook, and builds on years of Stanford research led by Sisu's founder, Peter Bailis.

To learn more about the technology behind Sisu, and to start a free trial for your analytics team, contact us at hello@sisu.ai or visit www.sisu.ai.

Recap

An enterprise's ability to capture and structure detailed information about its business has far surpassed its ability to utilize that information for daily decision making. In fact, **IDC's David Reinsel has postulated that in 2019,** "more data will be stored in the enterprise core [storage environments] than in all the world's existing endpoints."

Without fundamental changes to how advanced analytics teams and daily operational leaders in these organizations access, investigate, and explain this data, enterprises may miss the opportunity to shift their data warehouses from static systems of record to active sources of advantage.

By taking an objective-based, declarative approach to data exploration and analysis, Sisu can continuously assess information as it arrives and never stops asking any business's toughest question: "Why are my key metrics changing?"

Get in touch

hello@sisu.ai
www.sisu.ai
Sisu Data
@sisudata