CAMBRIDGE SEMANTICS
THE SMART DATA COMPANY®

ANZO® for Machine Learning

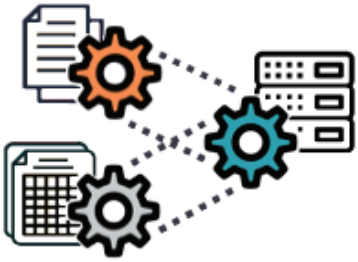
Fast-track your Machine Learning projects with rapidly integrated, high-quality data and optimized feature engineering and selection processing

Incomplete, inaccurate or overly complex data can throw off your best Machine Learning efforts. Data scientists often spend 70-80% of their time on manual, one-off data preparation and feature engineering development.

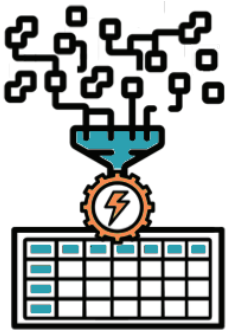
The Anzo for Machine Learning solution replaces this tedious, error-prone work with a modern data platform designed to rapidly integrate, harmonize and transform data from all relevant data sources into optimized Machine Learning-ready feature datasets.

Anzo provides the advanced data transformation functionality essential for fast and effective feature engineering to help separate key business signals from irrelevant noise.

Anzo enables data scientists to train, test and operationalize their machine learning models faster and more effectively than ever before:



Unify and utilize all structured and unstructured data



Expedite and optimize feature engineering

Manage data sets and retain data lineage/provenance

Rapidly deploy and continuously improve model performance

Unify and utilize all structured and unstructured data. Anzo freely integrates and harmonizes all relevant data sources – structured and unstructured data alike – using its built-in graph database and semantic data layer. Anzo conveys the business context and meaning of your data, making it easier for business users to understand and properly utilize.

Anzo enables the practical combination of far more sources and varieties of data more rapidly and reliably than other methods allow, enabling the creation of optimized training datasets with the strongest predictive signals possible for your models.

Cambridge Semantics Inc. • 1 Beacon Street, 15th Floor, Boston, MA 02108 USA • CambridgeSemantics.com

(617) 245-0517 • info@cambridgesemantics.com

Expedite and optimize feature engineering. Anzo dramatically speeds up the feature engineering process. In addition to automatically generating all necessary complex graph queries and transformations, Anzo provides an intuitive Excel-like formula language to quickly and easily define a wide variety of data transformations essential for effective feature engineering, including:

- Compute data aggregations to derive new features not already explicitly within the dataset
- Quickly de-normalize data from multi-dimensional data sources (e.g., include a feature that must be derived from multiple tables of data)
- Range aggregation (aka “bucketing”) of numeric values into custom grouping dimensions
- Render alpha values into numeric values (e.g., convert “Yes” or “No” to 1 or 0), as preferred by many Machine Learning frameworks
- Pivot categorical values

New calculations are retained and presented on a per feature basis in a highly discrete and granular manner for easily managed feature engineering and feature selection, even if calculations result in very wide tables with many feature columns.

Manage data sets and retain data lineage/provenance. Anzo also makes it easy to reuse harmonized structured and unstructured data by managing catalogs of data sets as well as ongoing aspects of data integrations such as data quality processing. Anzo also retains end-to-end lineage and provenance for the data comprising machine learning datasets so that it is easy to find out what data transformations are required when it comes to using models in production.

Anzo also makes your optimized Machine Learning datasets available for export or easy access via OData/REST APIs and SQL ODBC/JDBC using any data science tool like R or Pandas, downstream system or algorithm.

Rapidly deploy and continuously improve model performance. Anzo serves as an enterprise data platform to rapidly put your trained and tested models into production under dashboards and web services. Anzo provides a horizontally scalable, operational analytics runtime environment, with support for external processing (or federation) callable via SQL JDBC, HTTP/S, and local command line processing.

Once a model is put into production, Anzo helps you optimize its ongoing performance using the Anzo Data Science Toolkit (DSTK). The DSTK provides access to the same data preparation pipelines used to create the original model training data, while also enabling easy access to new complimentary data sources. Anzo can be used to monitor the performance of your model against new testing data for continuous model improvement.

Anzo can be installed on your preferred cloud environment or on-premise infrastructure, to ingest and connect all your data sources for comprehensive data integration, feature engineering and feature selection.

In a matter of days, your data scientists, business analysts and enterprise data architects will be training, testing and deploying machine learning models using high-quality, optimized datasets faster than they ever thought possible. Contact Cambridge Semantics and arrange for a personalized demo today.