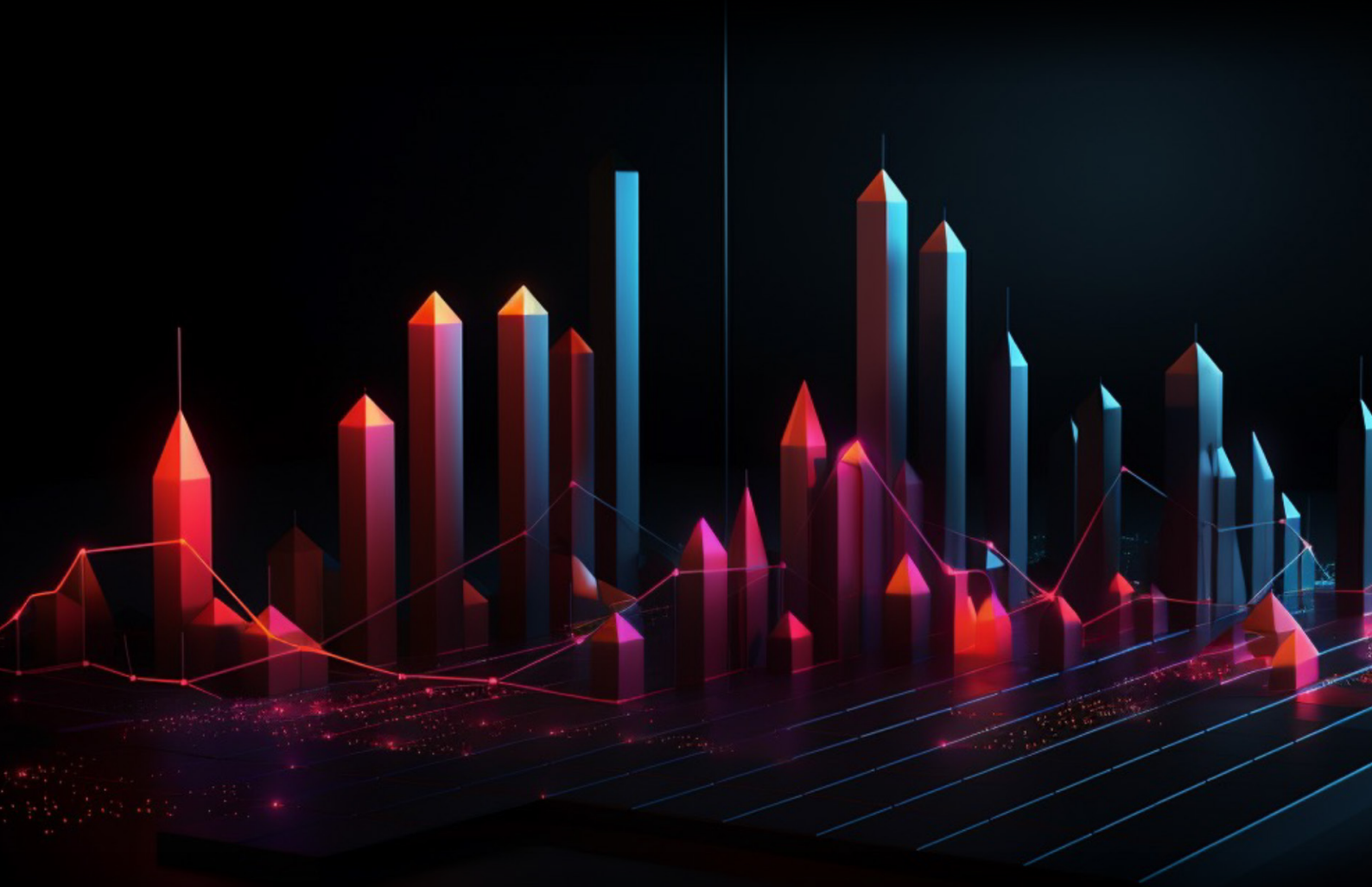




Base Model

Apply science to behavioral data. Automatically.



BaseModel is a private foundation model for behavioral data

Foundation models like **ChatGPT**, **GPT-4**, **Dall-E 2**, **StableDiffusion** have revolutionized Text and Image processing. A single large model trained on massive datasets can replace thousands of specialized models. For the first time, BaseModel allows to apply the same principle to behavioral data.

Reduce your modeling life-cycle to days instead of months

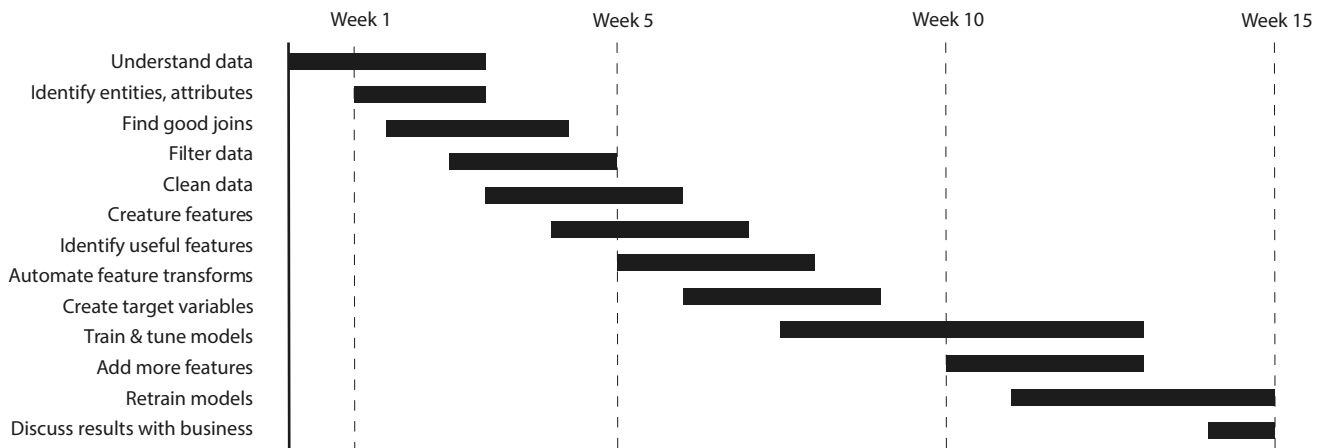
Understanding complex & intricate patterns of interactions is a super-human challenge. Imagine a single model could learn from all your raw data. Such a model could form a foundation for solving any applied task with unparalleled efficiency and quality. **This is exactly what BaseModel does.**

Until now:

- each ML project required careful manual labor, starting with analyzing available data sources
- countless handcrafted features had to be created using expert knowledge
- despite best efforts, important behavioral cues were often lost due to human limitations
- the information content of raw data was orders of magnitude richer than the actual input of models

BaseModel eliminates these problems and supercharges behavioral ML.

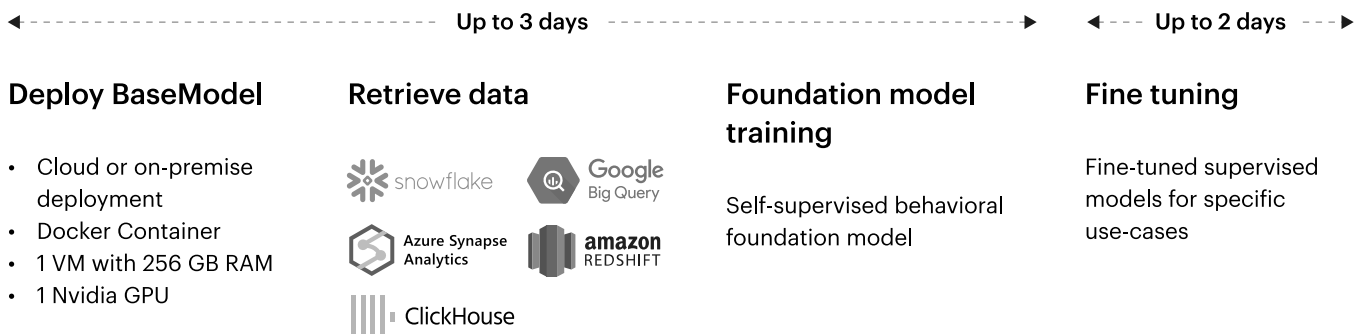
Current standard approach



BaseModel



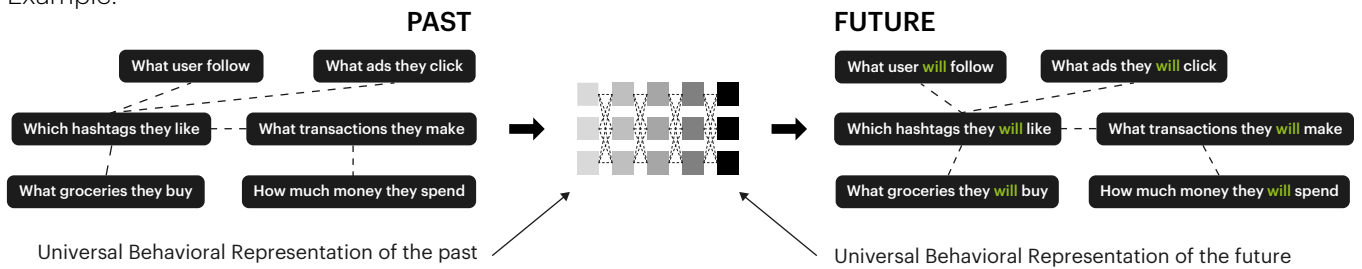
Just 5 days: from BaseModel configuration to first trained model



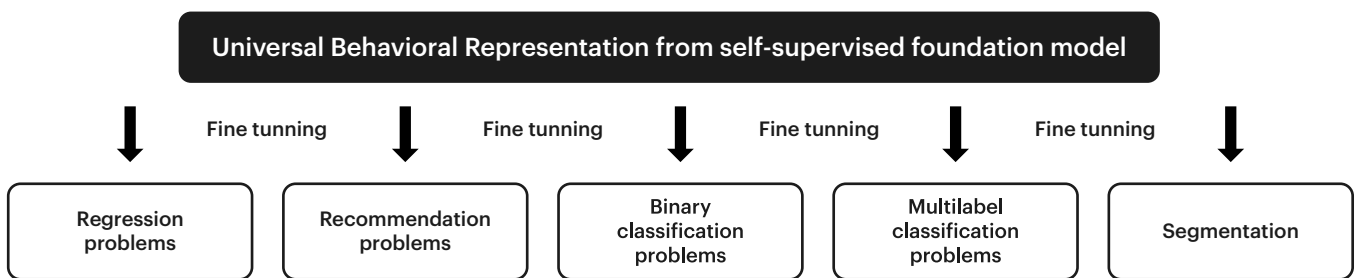
BaseModel: unprecedented accuracy of behavioral modeling

BaseModel looks at all human interactions with your organization and learns to predict future behaviors. Automatically. At scale.

Example:



For BaseModel, solving applied problems becomes a piece of cake thanks to understanding the full spectrum of behaviors. The foundation model is trained only once, and then adapted to specific tasks in a process called „fine-tuning“. Not having to re-learn behaviors from scratch, a fine-tuned model only needs a target to zoom in on.



Your private foundation model can bootstrap business applications like:

- propensity prediction
- churn prediction
- recommendations
- anomaly detection
- customer scoring
- customer matching
- behavioral hyper-segmentation
- and many more

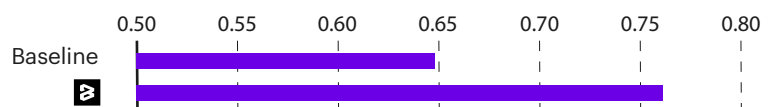
Just like large language models can be fine-tuned to any application, **BaseModel is not restricted to a predefined set of use cases** – you can unleash your creativity freely.

No more „plumbing pipelines“ and „handcrafting model inputs“. Ideate, prototype, evaluate!

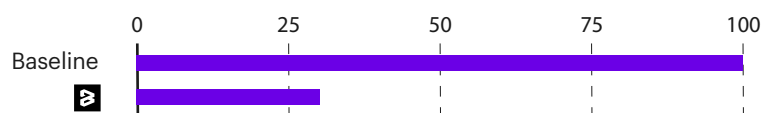
Thanks to BaseModel, applied models train faster, require less labelled data and perform better than classic approaches.

- BaseModel models are trained **2x faster** than baselines
- BaseModel needs **3.5x less** labelled data than baselines
- BaseModel’s quality metrics **significantly** outperform classic baselines
- BaseModel performs advanced feature engineering, representation learning & training **automatically**

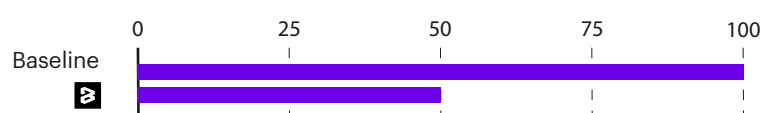
Model quality (AUC)



Datapoints required (%)













Training time (%)



Challenging top AI powerhouses

We take science seriously. We verified BaseModel's power at the most prestigious research venues, where early versions of **BaseModel took the podium** competing with global giants like **Google DeepMind, Baidu Research, NVIDIA, Intel, Oppo Research, Xerox PARC, Rakuten** and 800+ teams from leading universities worldwide.

The difference? Our competitors spent weeks hand-crafting their solutions for every problem separately, with zero re-usability. **We just applied BaseModel to all the problems.**

Organizer	Competition name	Description of the challenge	1st Prize	2nd Prize	3rd Prize
KDD2021	ACM KDD Cup Stanford OGB-LSC 2021	Predict the subject of scientific publications on the basis of edges contained in the heterogeneous graph of papers, citations, authors and scientific institutions.			
Booking.com	ACM WSDM Booking.com Data Challenge 2021	Provide the best suggestion for the next trip most likely to be bought by a booking.com customer based on millions of real, anonymized bookings of existing clients			Team Dasou
	ACM RecSys Twitter Challenge 2021	Based on 1 billion historical tweets predict 4 possible actions (likes, retweet, comment, retweet with comment) of the next tweet provided in a random language			layer6
Rakuten	SIGIR eCom Rakuten Challenge 2020	Identify the images of a huge data base of products without being given pairing of descriptions and images		Undisclosed	

Example questions BaseModel can answer

General How do daily customer interactions influence their future behaviors?	Retail How much will the customer spend in a specific category next week?	Travel What is the customer's expected number of trips this year?
Ecommerce What is the customer's likelihood of using a special offer?	Telco How much data traffic will the customer use this month?	Health How many diagnostic tests will the patient need this year?
Insurance How many insurance policies will the customer subscribe to this year?	Gaming How many power-ups/bundles will the gamer buy this month?	Banking What is the customer's projected profitability in the next quarter?
Ecommerce Which products/promotions/offers/categories the customer is interested in?	Home & Furniture How to split the customer population into behaviorally distinctive groups?	Automotive What kind of product/category is the customer interested in and why?
Insurance Will the customer churn in the near future and what events had an impact on that?	Banking What is the utility of customer for your business and what are the behavioral and sociodemographic factors affecting it?	Fashion Will the customer make a purchase next week? What steps need to be taken to increase the chance of purchase?
Payment Is recent behavior of the customer inconsistent with past habits?	Compliance Are there outlier customers in the population, who might be worth looking into?	News & Publishing Will the reader subscribe to a premium plan?



Concept used by leaders

„We are currently using EMDE2 for generating candidates to facilitate downstream recommendation systems. **It generates recommendations using density-based rich customer representation.** It allows us to trace **customer look-alikes** (‘People Like You’) to find similar users with similar cuisine/taste preferences as well as price affinity. We used Cleora for customer-restaurants graph data [...] And to our delight, the embedding generation was superfast (i.e <5 minutes). For context, do remember that GraphSAGE took ~20hours for the same data in the NCR region. **Cleora + EMDE gives us a generalised framework for recommendations** [...] We are exploring ways to use it in other applications such as search ranking, dish recommendations, etc.”

Zomato.com Data Science team

Zomato is an multinational restaurant aggregator and food delivery company founded 2008 owned inter alia by Uber and AliPay. Zomato provides information, menus and user-reviews of restaurants as well as food delivery options from partner restaurants in select cities. The service is available in 24 countries and in more than 10,000 cities. In financial year 2021, the average monthly active users for Zomato were 32.1 million users.

Technical Appendix

Examples of behavioral data BaseModel can be fueled with

Events:

web interactions	page views, searches, transactions, product returns, support queries, ...
offline interactions	transactions, contracts signed, customer support calls, ...
mobile app interactions	clicks, scrolls, push events, location, sensor data, ...
financial	invoicing, credit card payments, wire transfers, ATM withdrawals, ...
telecom	phone calls, text messages, internet usage, ...

many more

Attributes and metadata:

customer attributes	socio-demographic, location, loyalty program, subscriptions, ...
product attributes	brands, titles, descriptions, colors, sizes, styles, images, ...
employee attributes	seniority, specialty, skills, ...
marketing action attributes	channel, format, description, text content, ...
financial	merchant categorization, merchant description, transfer destination metadata, ...

many more

Thanks to extremely efficient algorithms and optimizations, BaseModel is blazingly fast. No clusters needed

Our mission to equip every organization with a private behavioral foundation model makes us very conscious about resources. That's why we take cost-efficiency seriously & optimize our algorithms to be lightning-fast and extremely scalable.

A full self-supervised training pipeline for 10M+ customer profiles with 1 year worth of data can complete within a few hours on a VM equipped with a single GPU. Subsequent fine-tuning for supervised applications is even faster, and requires very little labeled data.

Datasource	Timespan	Number of rows	Fitting time
Events 1	1 year	> 3 billion	100 minutes
Events 2	1 year	~ 630 million	50 minutes
Client Attributes	-	10 million	20 seconds

Simple, secure and quick deployment, with zero overhead

We've made sure that starting the journey with BaseModel is smooth:

- BaseModel can be deployed on your own **cloud subscription**, or on-premise (NVIDIA GPU-equipped server required)
- BaseModel is available as a hardened **Docker image**. It doesn't require clusters, multiple micro-services or standalone databases.
- BaseModel does not duplicate your data. It reads the necessary data **directly from your data-warehouse** in a streaming way, during model training & inference.
- Thanks to the above requirements, BaseModel is easily auditable and can pass the most **stringent security requirements**.

How does BaseModel work under the hood?

BaseModel automatically finds proper representations suitable for aggregation of data, such as:

- graphs
- texts
- images
- numbers
- categorical variables

It utilizes a mix of Graph ML, differential geometry and Deep Learning.

BaseModel uses proprietary research to represent complex multi-modal, multi-source histories of behavior in the form of sparse vectors, called **Universal Behavioral Representations**. Technically, these vectors represent probability density estimates over Riemannian product manifolds and can serve as both inputs and targets for neural network training. In simpler words, BaseModel compresses multi-modal event series into very wide **fixed-length sparse vectors**.

The key property of BaseModel's representations is that they are **approximately reversible** - which means, that it is mathematically possible to query a Universal Behavioral Representation about the elements aggregated within, with high accuracy. This allows neural models to „ask“ very specific questions about the user's historic activities, without the need for encoding precise knowledge as manual features. This unique property also allows for **fine-grained interpretability** of models (down to the lowest level of raw data).

Cleora + EMDE

Some of the algorithms used in BaseModel like **Cleora** and **EMDE** have been open-sourced and can be found in the last section of the Appendix.

We're committed to scientific transparency by publishing and open-sourcing parts of our research

OSS Cleora (graph + hypergraph embeddings):

Cleora

Scalable learning of stable and inductive entity embeddings for heterogeneous relational data

<http://cleora.ai>

OSS EMDE (multi-modal behavioral fusion):

EMDE

Efficient Manifold Density Estimator for high-dimensional vector spaces

<https://github.com/Synerise/kdd-cup-2021>

<https://github.com/Synerise/booking-challenge>

Publications

Theory

Theoretical underpinnings of technology behind BaseModel

- An efficient manifold density estimator for all recommendation systems

<https://arxiv.org/abs/2006.01894>

- T-EMDE: Sketching-based global similarity for cross-modal retrieval

<https://arxiv.org/abs/2105.04242>

- I know why you like this movie: Interpretable Efficient Multimodal Recommender
<https://arxiv.org/abs/2006.09979>
- Cleora: A Simple, Strong and Scalable Graph Embedding Scheme
<https://arxiv.org/abs/2102.02302>
- On the Unreasonable Effectiveness of Centroids in Image Retrieval
<https://arxiv.org/abs/2104.13643>
- A Strong Baseline for Fashion Retrieval with Person Re-Identification Models
<https://arxiv.org/abs/2003.04094>

Applications

BaseModel applied to real life use cases

- Synerise at KDD CUP 2021: Node classification in massive heterogenous graphs
https://ogb.stanford.edu/paper/kddcup2021/mag240m_SyneriseAI.pdf
- Synerise at RecSys 2021: Twitter user engagement prediction with a fast neural model
<https://arxiv.org/abs/2109.12985>
- Modeling Multi-Destination Trips with Sketch-Based Model
<https://arxiv.org/abs/2102.11252>
- Synerise at SIGIR Rakuten Data Challenge 2020: Efficient Manifold Density Estimator for Cross-Modal Retrieval
https://sigir-ecom.github.io/ecom20DCPapers/SIGIR_eCom20_DC_paper_1.pdf
- Synerise at SIGIR Rakuten Data Challenge 2020: Efficient Manifold Density Estimator for Multimodal Classification
https://sigir-ecom.github.io/ecom20DCPapers/SIGIR_eCom20_DC_paper_3.pdf

