TeraCrunch Solution Blade White Paper

Transforming Organizations
With Predictive Analytics & Artificial Intelligence

May 2016

TABLE OF CONTENTS

1. INTRODUCTION

According to Forrester most firms use less than 5% of available data. They want to tap into the power of advanced analytics, but cannot. The reason is, developing & packaging advanced analytics solutions to solve business problems is complex, time consuming and costly. TeraCrunch addresses this growing industry problem. TeraCrunch solves business challenges for its clients by providing innovative advanced analytics solutions and a highly experienced data science team to implement it right. Our approach is based on customizing *prepared solutions* for key problems in the Operations and Marketing domains across many verticals as a *software-as-a-service* (SaaS).

Since its inception TeraCrunch has pioneered the development of innovative, end-to-end data analytics solutions for a variety of industries ranging from insurance and finance, to retail and cloud infrastructure, to healthcare and legal. The specific applications are equally rich and varied, spanning objectives such as minimizing credit risk, optimizing staffing, reducing customer churn, increasing sales or optimizing hospital drug supply costs, to name a select few.

2. APPROACH

A fundamental component of our architecture is the concept of a *Solution Blade* as illustrated in Figure 1 below.

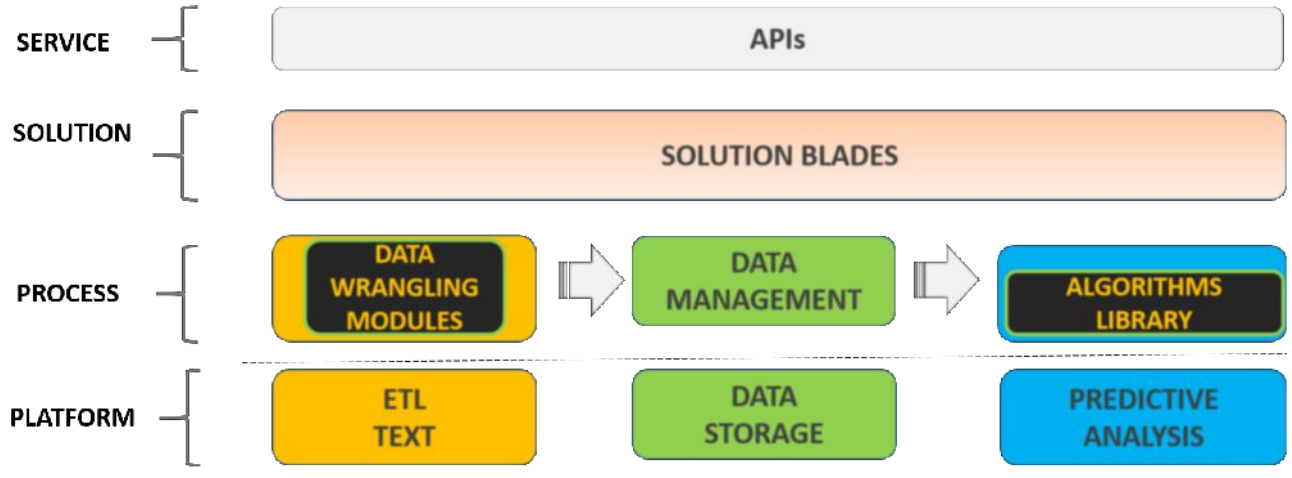


Figure 1. Socratez ™, The Solution Blade Architecture

A Solution Blade is [RECIPE+SOFTWARE] where the recipe comprises of a series of steps and also best practices to identify, select, assemble, process, integrate, and analyze data to provide insights and/or predictions to address a customer problem at hand. For example, we have solution blades for customer analysis, predicting resource usage (of medications & other equipment) for a large hospital, and so forth. We have developed a large library of such solution blades that are highly configurable. Figure 1 highlights several important aspects of this architecture. This architecture enables our data science team to quickly tailor advanced analytics solutions based on each client's business needs. Our solution architecture is extensively tested & optimized, and therefore yields predictably accurate results.

3. PLATFORM

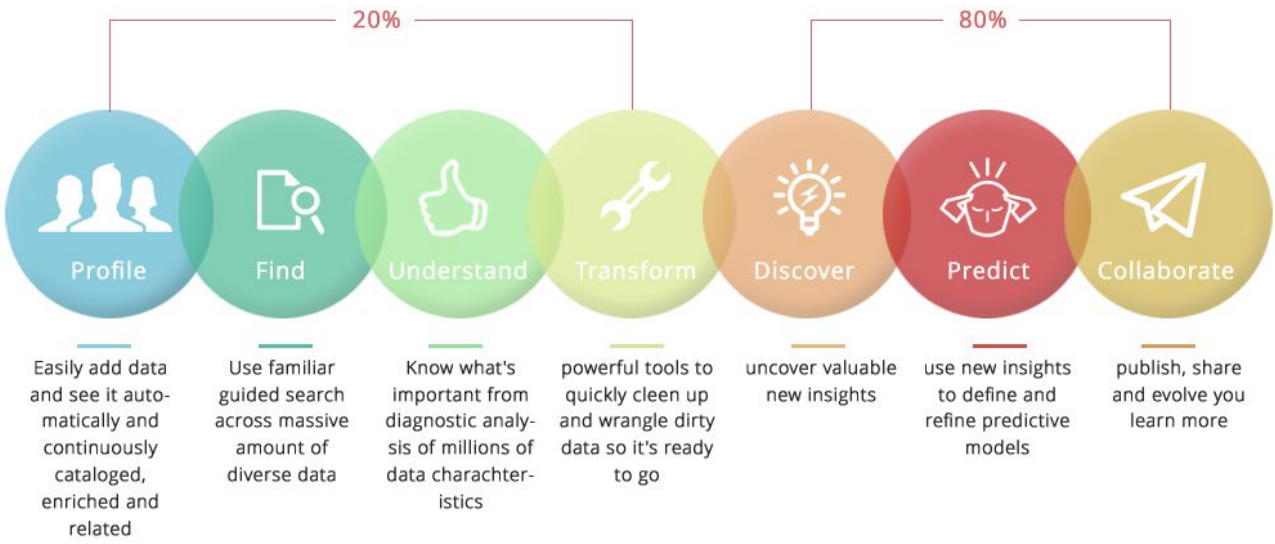
This layer comprises of the baseline software platforms and engines. For data preparation this includes tools such as ETL tools, data integration software, and also text analytics engines for understanding unstructured data. For data management this layer includes multiple different database engines and stores including relational databases, NoSQL databases, graph stores, and Semantic (Triple) stores. One of our key strengths is in the area of understanding *unstructured data*, a modality of data that has seen an exponential growth in not only content but also relevance to predictive tasks in recent years. TeraCrunch has developed a patent-pending platform called Socratez™ for extracting structured information and concepts from unstructured natural language text. Socratez™ can extract a wide variety of concepts, entities, and expressions in a normalized manner from unstructured text. We have leveraged decades of R&D expertise in the areas of natural language processing, machine learning and semantics towards the development of this platform.

4. PROCESS

The process layer is composed of software modules and libraries we have developed for tasks found to be common and pervasive across multiple Operations and Marketing solutions. The process layer software lies primarily in two phases: data preparation and data analysis, which we discuss below.

4.1 ***Data Preparation***

PARETO PRINCIPLE IN BIG DATA ANALYTICS: *Rarely is there a scenario in which new data sets*



arrive and the data is instantaneously "analyzed." Data almost always has to go through a preparatory stage or series of stages. In today's world of big data in which the data may have unknown value, there can be as many as seven stages as indicated in the figure.

While organizations aspire to have a completely integrated data management system, the majority of data required to make strategic business decisions still resides ***outside*** *their IT-controlled data environment.*

In fact if we are dealing with high volumes and high velocities of Big Data, this 80/20 ratio may be more like 90/10 as Big Data exacerbates this problem. Once data has been properly prepared, then and only then can it be used for the actual analysis activity.

More often than not, organizations do *not* have a clear understanding of the specific data that any proposed analytic solution will be based on. This includes data within the organization, which may have multiple different sources of potentially relevant data, as well as external sources of data such as open data or data from the Internet and social-media. Further, significant additional processing is required to make the data amenable for analysis. The table below lists the key processing operations we encounter frequently in our applications.

Table: Fundamental Data Preparation Operations

DATA PREPARATION PROCESS	DEFINITION
Data Selection	The process of identifying the set of data from amongst the databases with the enterprise, information systems, unstructured data repositories, other sources in the enterprise, as well as external data such as on the internet, social-media and open data that will be used in the solution.
Data Blending	The process of combining data from multiple sources to create an actionable analytic dataset for business decision-making.
Data Cleaning	The process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. We also extend data cleaning to data inference in certain cases, when the data is incomplete or missing.
Data Aggregation	Roll-up, aggregation, and transformation operations on the data so as to get it to the right level of abstraction for analysis.
Data Structuring	The process of curating structured data to the level from unstructured, natural language text using text analytic technology.
Deduplication	The process of resolving terms representing the same entity but which is represented differently at different points in the data.

WRANGLER™

TeraCrunch's WRANGLER™ *suite of data wrangling modules is an combination of several man decades of research expertise and experience in data transformation, integration, and synthesis research.*

Wrangler™ is a suite of modules for data preparation developed by TeraCrunch. We have generic modules for each of the above data preparation tasks including data deduplication, entity resolution, data harmonization, and data transformation. The development of WRANGLER™ has also leveraged decades of our own R&D expertise in key data management problems. This includes pioneering work in developing general data integration systems such as "mediators" and "federated" database systems in a wide variety of domains including government data, aviation and aerospace data, digital government, and biomedical informatics. We have pioneered the development of algorithms for entity resolution (or data de-deduplication) including algorithms and implemented approaches based on graph network analysis as well as machine learning. Also (as mentioned above) we have particular expertise

in the distillation of information from unstructured text, having developed multiple next-generation systems for text analytics in domains ranging from clinical data to social-media conversations.

4.2 ***Data Analysis***

We have employed a rich variety of analytical models in our applications including Regression algorithms over historical data, a variety of Classification algorithms including Decision Tree, SVM, GBM and hybrid models, and time-series algorithms (such as ARIMA) to name a few. With the development of such applications has emerged a rich library of (implemented) predictive models that we reuse and further configure for new similar class of applications.

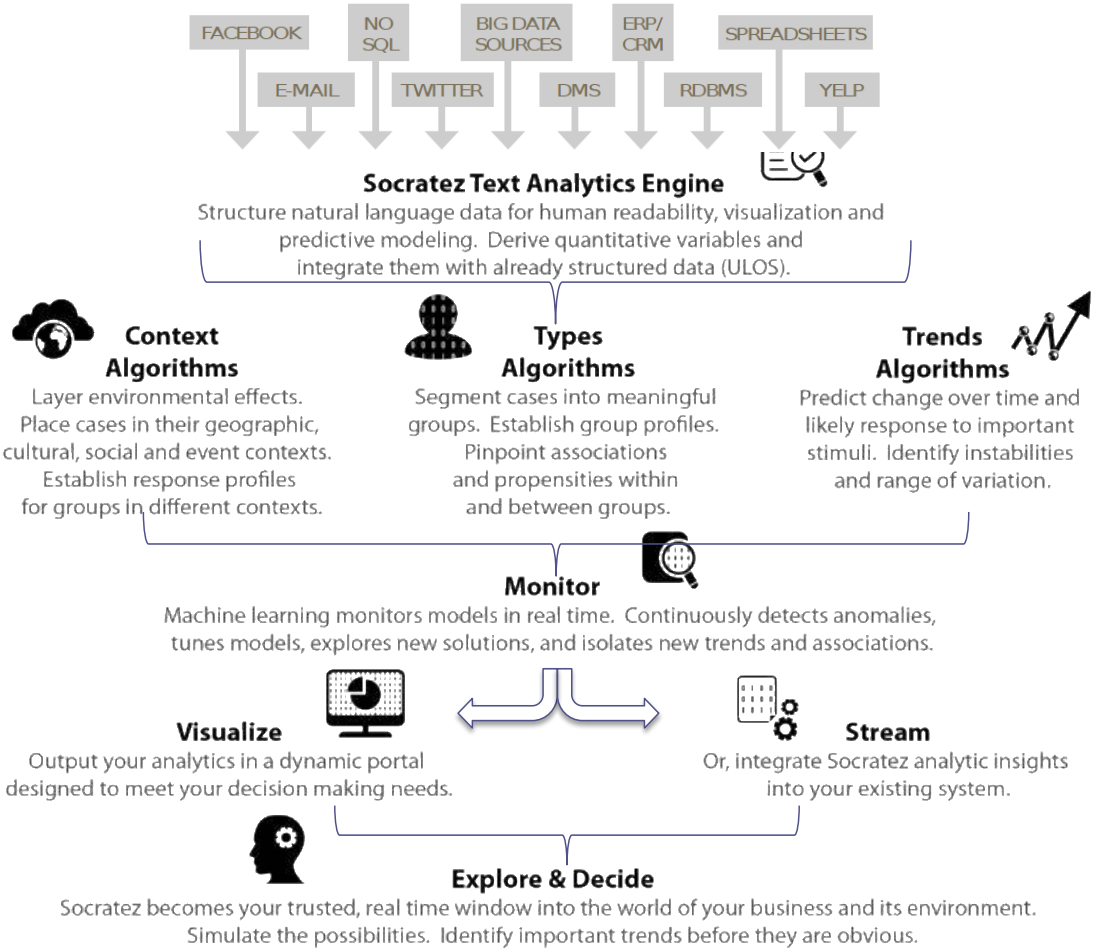


Figure 2. Socratez ™, LIGO ™ (Library of Algorithms) Overview

LIGO™ (Library of Algorithms) is a TeraCrunch curated library of data analysis and predictive models. LIGO contains predictive models, where the models can be reused for similar applications with minimal adaptation.

TeraCrunch has divided its primary algorithmic library into three groups: **Context**, **Types**, and **Trends** algorithms. Types algorithms establishes shared profiles between cases by classifying and clustering their characteristics. Context algorithms isolate individual sources and patterns of variation from their environment. Trends algorithms project these characteristics and behaviors through time to estimate what is likely to occur. They work together to generate complete, dynamic insights into your pressing business questions.

Types differentiates segmentations, clusters, profiles and personas. Classifies people, statements, businesses and other identifiable cases into meaningful groups, and evaluates data for these meaningful points of distinction using a variety of classification and clustering algorithms.

- How are my customers different from one another?
- How do my salespeople act differently from one another?
- Which products tend to go together, and for whom?

Context separates individual effects from their physical, cultural, social, network, organizational, and market environments. Uses the latest in hierarchical and variance components models to distinguish environmental from individual effects and partitions the variance accordingly so that you can choose the best strategy for the environment your business currently operates in.

- Are my products performing differently than others in its market?
- Do people who follow sports also like my products?
- How do people use my services differently as they age?

Trends projects cohort, longitudinal, and time series analyses into the future and separates dominant trends from anomalies. In a fast-paced, dynamic world, businesses need a real time understanding of what's changing and why. Our suite of trends algorithms detect what's changing, how it's changing, and offers deep understanding of why this is the case.

- What are my supply & demand likely to be next week? Next quarter?
- How are operational costs changing as we move into a new market?
- Does the customer equipment need maintenance in the coming weeks or months?

TeraCrunch's LIGO™ *algorithms library has evolved from rigorous experimentation with and evaluation of multiple machine-learning and predictive analysis algorithms applied to a variety of real-world analytics tasks.*

5. SOLUTION

Solutions are configured & tailored with solution blades, which is a data processing & analysis recipe and the associated software to achieve it. The software aspect of a solution blade is illustrated in Figure 3 below.

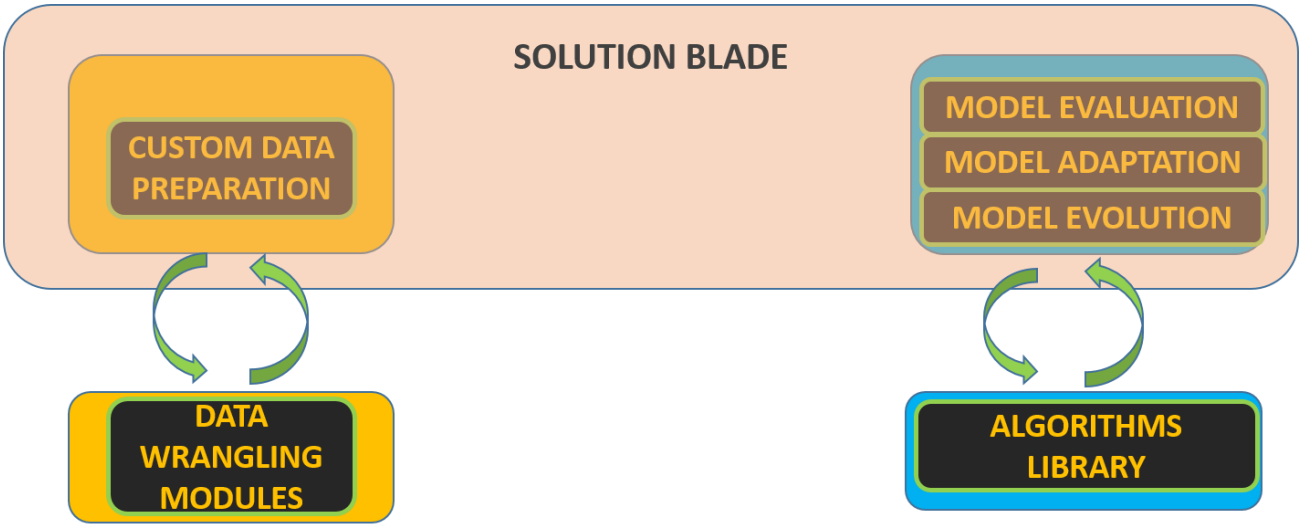


Figure 3. Socratez ™ - Solution Blade Overview

Every solution requires some custom data preparation. For instance many solutions require some level of data blending from different sources, the kind of data cleaning required is different in different solutions, and solutions which have a component of unstructured data utilized require data structuring. Each solution has some custom data preparation modules that are developed on top of the (relatively) more generic data wrangling operations in the process layer.

Similarly, predictive models must be customized to particular solutions. While the process layer would contain general purpose models for, say, time series analysis, further customization is required to adapt this time series model for a network server outage prediction solution versus a retail out-of-stock prediction solution.

5.1 **New Solution Blade Development**

For a first instance of an application for us, we design and instantiate a new solution blade. The first step is to identify the (kinds of) data collections that the predictive solution would be based on. Then, we identify the data wrangling operations that must be done to prepare such data for analysis. Typically many of the required wrangling operations have already been implemented in our Data Wrangling modules though some level of customization is usually required for a new application.

For predictive analysis we then consider the algorithms library and select implemented models and algorithms as appropriate. Again while more generic implementations of the predictive algorithms are already present in the Algorithms Library, additional customization may be required for instance in areas like feature engineering for the models. Further, we do R&D with multiple relevant algorithms and conduct parameter and hyper-parameter tuning to hone in the best possible predictive algorithm (or ensemble thereof) for the problem.

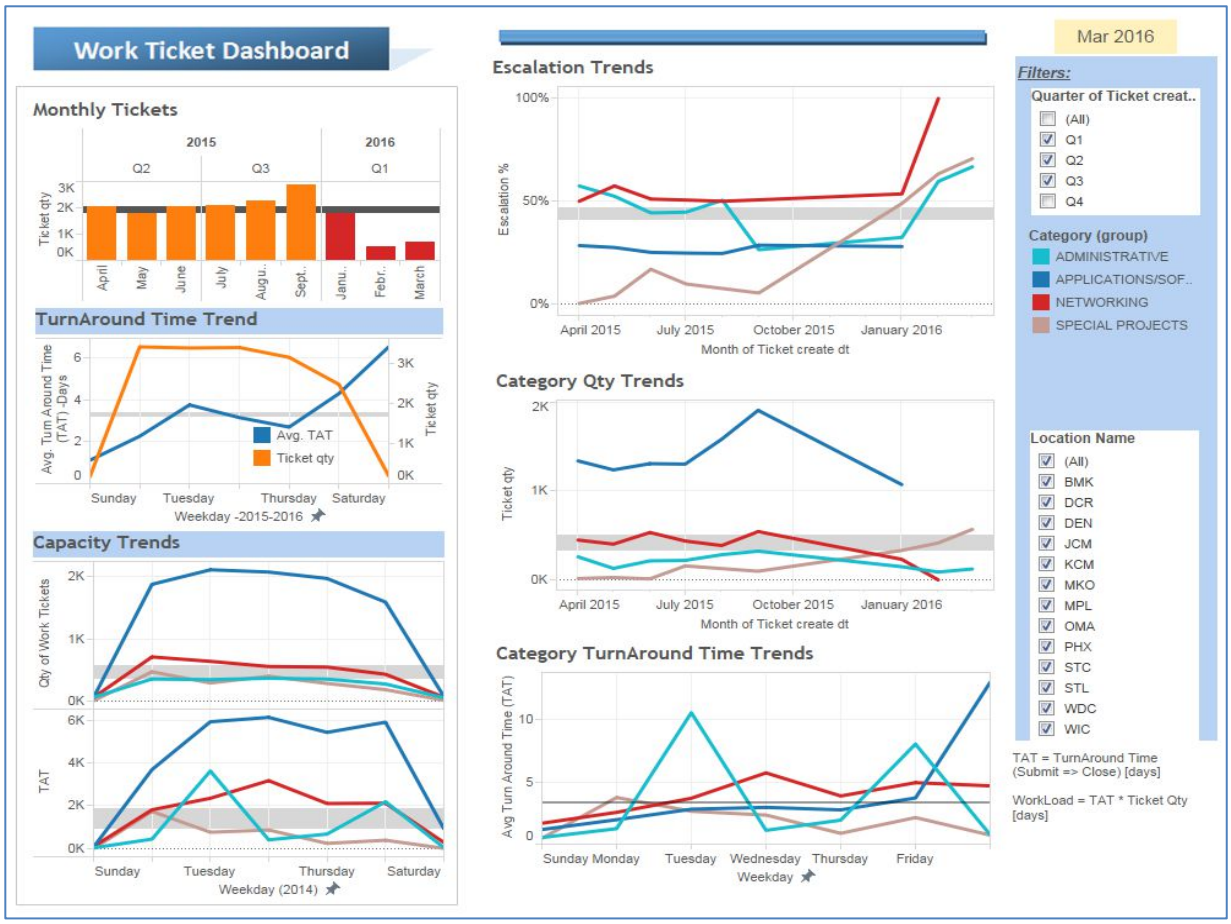
5.2 **Solution Blade Reuse**

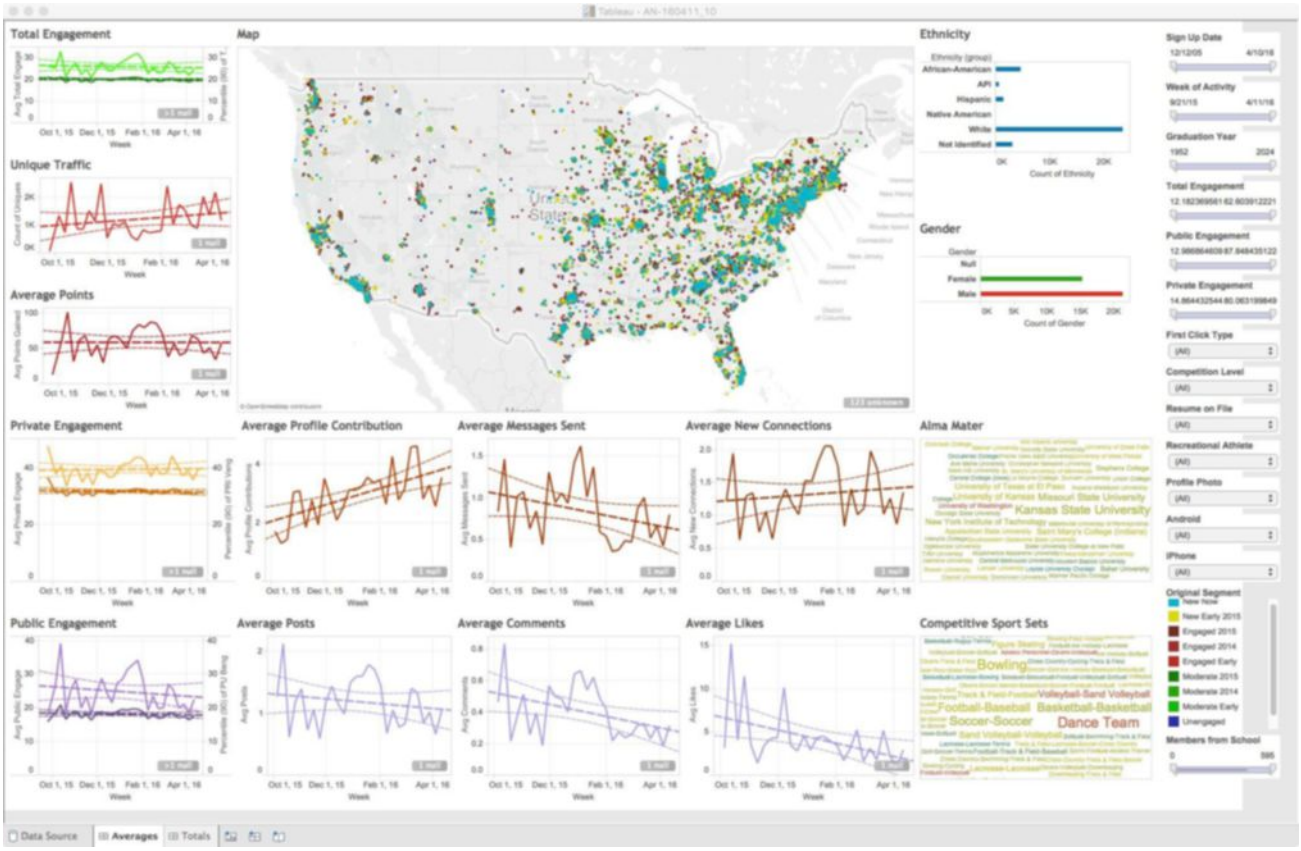
In many instances we encounter problems and application for which we have pre-assembled solution blades. As one example we frequently develop applications in the marketing domain and have pre-assembled solution blades for applications like customer retention, promotion targeting, and others that we frequently engage in.

Some level of additional customization is required, starting from data selection; a new client and organization may not necessarily have the same relevant data for an application that we assumed for previous clients. The actual client data itself may require additional customization in data wrangling. Also, while the solution blade may contain the best applicable predictive algorithm(s) for the problem, the predictive models may have to be reconfigured and optimized for that particular application and instances of relevant data.

5.3 **Data Visualization**

Despite inherent complexity, it is key that our solution analytic insights can be readily understood by customers and in an intuitive manner. We place significant emphasis on representing the analyzed data and uncovered insights with powerful, rich, but easy-to-understand visualizations using state-of-the-art tools.





6. CONCLUSION

Analytic solutions that enable end-to-end capabilities are complex to create. However the commonality of issues across multiple solutions, 100+ years of data science R&D and the demand for similar solutions across different verticals has led us to the solution blade approach with which we are able to offer solutions with maximal pre configuration, and an efficient methodology to perform customizations for each client.

TeraCrunch's Socratez Solution blade architecture coupled with its Data Science team, offers highly accurate analytics that can be iteratively expanded, scaled and continually optimized.