PRISM INFORMATIX

EMPOWERING COLLABORATIVE ANALYTICS

Real-time Prediction with Distributed Big Data

Whitepaper: Core Platform, Draft September 2016

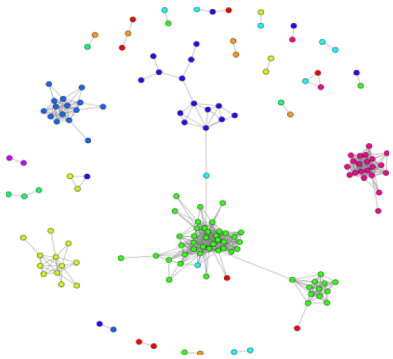
ABSTRACT

The most significant barrier to realizing the value of combined data from high variety sources such as mobile devices, social networks, and sensors is the need to integrate data to perform analysis. Data integration for predictive modeling is complex, expensive, and challenged to respond to data dynamics. In addition, communicating and assembling large integrated datasets for machine learning is increasingly in tension with proprietary and security considerations. What is needed is a cost-effective means of quantifying the value of information and integrating diverse analytics into a unified model, *without ever integrating the data*.

Here we describe a totally new approach for real-time decision making from high variety and distributed big data that avoids the costs of data integration. The method is based on a novel system of collaborative analytics in which the data is analyzed globally but in distributed fashion, in place, organically and privately. Because the raw data are never integrated, substantial acceleration in learning and prediction, enhanced agility to adapt to dynamic data, and significant reductions in cost and complexity are obtained. The approach makes it possible to explore rapidly many combinations of data to make better decisions, even when sources cannot be directly accessed due to privacy or proprietary restrictions. This is accomplished through a unique analytics platform that designs itself to answer user queries, and can run asynchronously, at global scale, over the Internet.

In this system, the currency of exchange is information, in the form of compact analytics and models, not data. And as a result, pathways are opened to new business models and marketplaces for information that allow both data consumers and data producers to trade confidently and profitably in all varieties of information.

A Revolutionary New Approach to Managing Complexity and Realizing ROI with Big Data



Alan I Chao, John W. Wissinger

Keywords: predictive analytics, big data, distributed analytics, stream processing, decision systems, team decision making, real-time analytics, risk management, data integration, collaborative analysis

TABLE OF CONTENTS

EMPOWERING COLLABORATIVE ANALYTICS

Whitepaper, Core Platform, Draft September 2016.

1. THE "BIG" PROBLEM

A relentless increase in the volume and variety of potentially useful datasets has brought about the industrialization of data on a massive scale, driving large investments in infrastructure for storing, managing, preparing, and computing on all this data. The goal is to capitalize on opportunities to derive insights with business or scientific value. There is increasing need to exploit more and different sources of data to improve decision making, and in particular, to use a variety of different source types that provide complementary and up-to-date information related to questions of interest. This data may include sources as diverse as sensor feeds, social media, digital exhaust from mobile devices, documents, images, and so on. The problem is that accessing, gathering together, and preparing all this data for analysis is expensive and time consuming, and the situation holds little promise of scaling well into the future.

This challenge of big data is widespread. All businesses are driven to use data to improve customer experiences through tailored or personalized content, to improve products and services, and to improve efficiencies and lower costs. According to a recent Harvard Business Review study[1], a full *two-thirds* of organizations are already trying to blend together five to fifteen different sources of data for analysis, and the majority who use manual analysis via spreadsheets realize that it is not a viable solution anymore, driving them toward automated analytics. Furthermore, because the desired raw data content itself is often proprietary, or has critical business value for competitive advantage, there are significant barriers to sharing it in a B2B setting. While the emergence of the data economy and trends such as the Internet of Things (IoT) ensure that more and more potentially useful sources of information will become available, they will likely continue to be difficult to share in practice due to incompatibilities in business models, infrastructure variabilities, high data velocities, and proprietary concerns.

What is needed is an efficient way to analyze a wide variety of sources taken together, automatically within an accelerated cycle time and infrastructure-agnostic way, to provide users a *cost-effective approach to navigate* their way to analysis solutions that have good value and return on investment. Data consumers need help to answer questions such as:

[1] "Data Blending: A Powerful Method for Faster, Easier Decisions", Harvard Business Review Analytic Services Report, Aug 2015.

- *"How do I make use of all this data, it is distributed all over the place, it's in different formats, and speaks to different aspects of what I want to understand?"*
- *"What data should I put together? What will be the value after I incur the cost to do so? If I can only put together a subset of it, what sources should I use to get the most benefit? This looks like a great new source of data, but how can I know it's really going to help answer my question before the cost of a full data integration?"*
- *"What if it is restricted and proprietary, could I still use it? Is it worth buying, and if so, for how much? What kind of follow-on governance framework will I need?"*
- *"How can I take advantage of new sources that arise without relearning all my models?"*
- *"Since I cannot control data quality of external sources, how can I manage my risk in using it?"*
- *"How can I respond rapidly to changes in my data? How can I track variations in quality and meaningful content?"*
- *"How can I establish data relevance, and determine obsolescence? Can I readily disconnect from sources? If I can only keep around some of the data, what should I keep?"*
- *"How can I experiment and try all the things I imagine doing with the data quickly and affordably?"*

And on the other side of the coin, the key to successful monetization for data producers involves answering questions such as:

- *"How can I establish the value of my data? How much should I charge for it?"*
- *"How can I share my data without undermining its value? How can I share while keeping control? Could I ever confidently share my sensitive data with competitors?"*
- *"How can I avoid creating a different version of my dataset for every customer that wants it?"*
- *"How can I sell my data while conforming to regulations and privacy concerns?"*
- *"How can I even find and reach the customers that might have interest in what I am collecting?"*

The following illustration places these challenges into a more specific business context.

Vignette: Risk Reduction for Real-Time Decisions with Dynamic Data

A Problem of Value

Problems of automated decision making with uncertain information appear in areas such as online credit approval, mortgage loan approval, electronic trading, and insurance policy issuance, to name a few. Data-driven analytics can support the decision process to reduce risk and establish fair pricing by creating models based on typical consumer profiles and histories. Leveraging multiple sources of data that are timely and relevant can make decision models even more accurate by *tailoring* to specific individuals and current circumstances, and taking advantage of the *most recent* data available (to overcome staleness). While decision models are commonly derived from large bodies of historical data, ***it is often when deviations from that history occur, i.e., when change is happening, and shocks are occurring, that the risk profile is also rapidly changing***. Modeling systems that can recognize and adapt to that change will perform better. This illustration will describe using multi-source data to answer the question of whether or not to extend an online customer credit, e.g. to make a spot purchase or to issue a credit card, although the themes will extend to many other situations like those just mentioned above. This vignette is not intended as a case study, but rather as an illustration of the challenges and opportunities.

The Actors and Their Interests

The data consumer in this application wants to leverage multiple sources to help answer the business question "Approve Credit?", meaning to predict[2] a binary-valued outcome variable "Yes" or "No" with associated confidence for the applicant. The objective is to answer this question with the highest expected probability that the decision will be correct, in the sense that the individual who is extended credit will pay off that credit within the established business rules, and will not be late on payments or eventually default. This outcome is directly linked to the profitability of the business extending the credit. A constraint is that this good decision needs to be made in "real-time", where that means that the automated decision system returns the result with a latency of a few seconds from the time that the requested user data has been provided. This is commonly referred to as an "instant" approval decision.

The data producers may include companies that have collected data that could be useful to help make the credit approval assessment. While historically credit approval has focused on the customer's past history of creditworthiness, mostly as reflected in the credit score, there is increasing interest in improving the models with respect to a customer's projected future ability to pay, and bringing in factors related to the customer's future financial stability. The data producers would like to sell their data to the decision maker, but to do so, they need to establish its utility, its quality, its overlap

[2] See appendix for a brief backgrounder on terms such as predictive analytics.

(intersected coverage), and possibly overcome barriers to sharing related to privacy, regulations, or proprietary concerns. And of course they need to be able to deliver it on-time and continuously to support instant decisions.

Sources of Data

Automated systems are being fielded with simplified front-end customer interfaces, to minimize the burden of data entry by the customer to try and maximize sales conversion by avoiding a complex application process. These systems are supported by increasingly sophisticated analytics back-ends that augment customer-entered data with additional supporting information, in real-time, and eliminate any human-in-the-loop review of applications. The specific algorithms employed by different companies that extend credit constitutes their secret sauce, and their approaches are under continuous scrutiny and improvement as small improvements in performance, meaning correctness of approval decisions, can lead to big impacts on profitability.

What changes with Big Data is the organizations can now construct more complete and up-to-date views of risks by incorporating a wide variety of sources together. A host of different data sources could be used to support credit decisions, and this data may originate in different systems, in different locations, in different formats, and typically would evolve along different timescales, as illustrated in Figure 1-1.

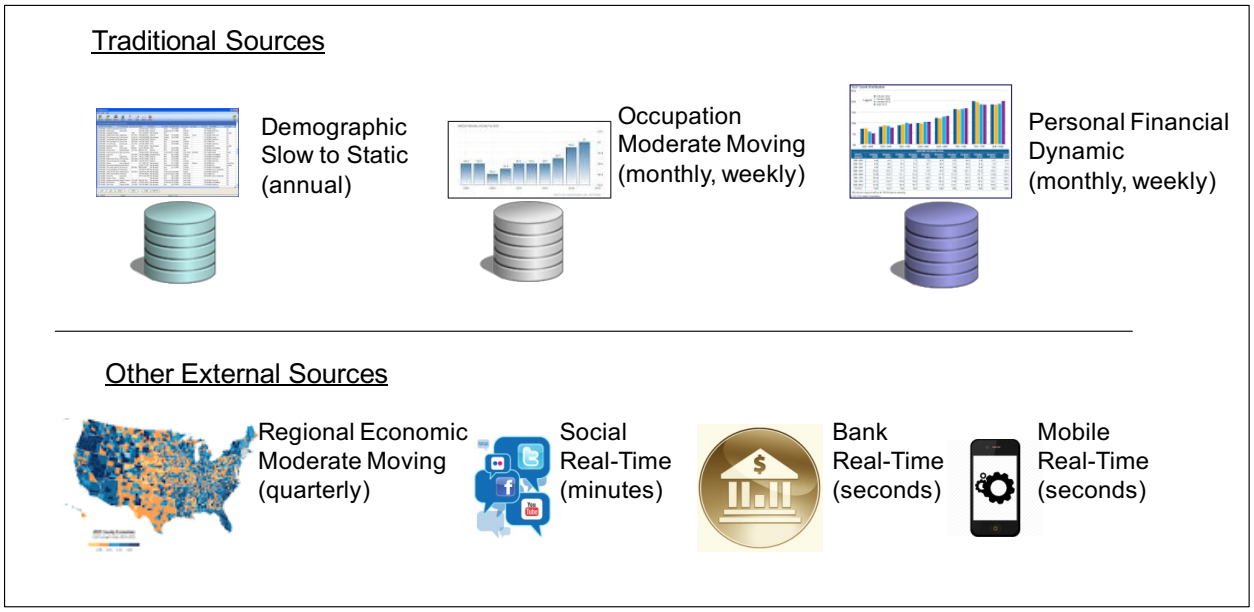


Figure 1-1: Sources of data that can support credit approval decisions are distributed and high variety and evolve along different timescales. The top panel shows the kind of slower moving information that is typically aggregated for analysis in conventional methods. The bottom panel represents new opportunities that are moving faster.

Historically, the single key driver of credit approval decisions has been the credit score produced by the nationwide credit reporting bureaus (Experian, Equifax, TransUnion). That information may be complemented with other demographic and personal financial history information, such as the specific types of information described in the "traditional" column of Table 1-1 below. This information is largely historical and semi-static, and because it is often gathered up and maintained in internally-controlled systems, the frequency with which it is updated (snapshots taken of it) can be quite low.

Table 1-1: Sampling of the kind of detailed information that can be exploited to improve decision making models for credit approval processing. Moving from left to right across the columns, the information becomes more personal and time dynamic, finer resolution.

Category	**Decision Factors – Features**		
	Traditional (core models)	Alternative (augmented models)	Emerging ("wild" new opportunities)
Demographic	age, education: level attained, gender, marital status, address	home: rent/own, address changes, dependents	marital status volatility, family legal troubles, education: field of study and institution attended
Occupation	current employment and income	employment and income history/volatility	name of company, health claims filed, performance ratings
Personal Financial	credit score (bureau), bankruptcies, history prior default	credit score (alt.), utility payment histories, debt-to-income ratio	early withdrawal from retirement accounts, stock holdings and movements
Regional Economic			local jobs outlook by profession, median incomes, regional default histories
Social Media			education: history including professional degrees societies honors, size of professional network, tweets/postings/blogs regarding purchases and purchase plans, financial status, relationship status, health and health status of immediate family members, pregnancy, gambling habits
Bank Data	co-signs on other loans	outstanding credit and loan applications	sufficiency of current cash balances, spending status on other cards, stability of savings accounts
Mobile Data			personal behavioral profiles, calls from collection agencies, location patterns, online shopping habits, application usage patterns

A host of firms have recognized that other information that is not traditionally reported to the credit agencies could be applied to benefit, and so there are various alternative credit scores now being produced. The alternative column in Table 1-1 summarizes

some of the specific types of information starting to be used to produce these augmented scores. For example, while a strong history of on-time payment of utility bills (gas, electric, water, etc.) could be considered a solid indicator of financial stability and responsibility, it is not reported to the credit bureaus, and so does not impact their assignment of credit scores. The augmented models seek to exploit such complementary factors that are more personalized and more time dynamic. A challenge is that such information is often owned by other organizations, and may have restrictions for its distribution, so it can be difficult to acquire and bring into indigenous systems.

The emerging column at the far right of Table 1-1 includes other potential factors that are much more personal and targeted to individuals, indeed many would be considered as personally sensitive "early warning indicators" or "red flags", but they are in very obvious tension with privacy, and would likely be only occasionally available. Because of this, one can speculate on the utility, but it is hard to even test them, and they are volatile and constantly moving. However, if they could be included in the analysis, these factors would likely usefully complement the other sources of information to improve a model's ability to predict future financial stability and ability to pay. Consider the impact on assessing future creditworthiness from the higher resolution information in these examples:

- Education
    - Core data only: master's degree attained
    - Enhanced data: MS in Bioinformatics, Stanford (e.g., mined from LinkedIn)

- Occupation Status
    - Core data: employed, Acme machines, senior service manager, $75k / annual
    - Enhanced data: 3 job changes over last year, former VP sales, changed geography twice, previous average salary was $90k+, machine industry now on the decline in current geography with major plant closures

- Bank Account Status
    - Core data: not available
    - Enhanced data: owns two accounts with large balances (>$100K), but they have diminished 50% over the last year, and another large withdrawal was just made (diminishing cash reserves)

- Personal Financial Sentiment
    - Core data: not available
    - Enhanced data: lost a big bet in Las Vegas last weekend, spent the night in jail, wife just filed for divorce, may have to sell the house! (e.g., mined from Facebook)

The education example speaks to future earning potential and employability, the occupation status speaks to job stability and career trend, the bank account status speaks directly to financial reserves and ability to pay, and the social media extraction creates visibility into "breaking news" about the applicant's current personal circumstance and financial confidence.

The Big Problem is to bring the right information together without bringing all the data together

The essential problem is that traditional methods are making decisions using data that is largely static in nature or sampled way too infrequently, whereas the key information that could lead to better decisions is actually captured in factors that are highly targeted and dynamic. But the difficulties in leveraging data in this way are as obvious as the opportunities: How can you gain access to that kind of data? How do you know a source of information will even provide valuable levels of performance gain without spending the time and effort to try it? How can "data snapshots" be kept from getting stale when the information is always changing? How could the barriers to personal privacy ever be overcome, and do we want them to be?

Getting at the Value – What Would it REALLY Take?

In an ideal solution, the information could be brought together and brokered between data consumers and data producers without compromising the private nature of the raw data itself. The analytics system would quantify the value of information, and suggest which sources were most worth combining. The system would be agile enough to keep pace with rapidly changing data sources, and make it easy to add new sources as they became available. And finally, sources could be included when that data happened to be available, or even just selectively invoked on an "as-needed" basis to resolve ambiguity in the decision process. Hold those thoughts!

In the sections that follow, we will continue with the credit approval illustration to contrast the challenges that conventional methodologies face to deliver on the ideal solution, with the advantages of the new approach we propose here.

2. Current State-Of-the-Art

As just illustrated, there is increasing interest in developing predictive analytics that involve using less conventional and fast evolving sources of personal information about customers, such as may derive from sensors in her car or on her appliances, or even her recent social media postings[3], and combining that with more traditional sources to make

[3] "Variety, Not Volume, is Driving Big Data Initiatives". MIT Sloan Management Review, Data & Analytics Blog, March 28, 2016.

better decisions. We will see that kind of challenge is well-suited to the new technology that will be presented here.

The most common approach in use today is to bring the various types of data together into a "data lake", often hosted in the cloud and maintained in a system like Hadoop/HDFS or NoSQL. To enable proper assessment of coverage, a matching problem[4] must be solved to resolve entities across the silos to properly associate the feature data and establish that a sufficient number of training samples can be assembled to learn the model. The data is subsequently prepared and fed to a centralized machine learning (ML) algorithm as illustrated in Figure 2-1. However, bringing together heterogeneous big data for machine learning from geographically and phenomenologically diverse sources can be difficult due to access restrictions including from proprietary or privacy considerations, from costs and delays to acquire, and high local arrival velocities. And the simple act of gathering the data together and associating it can create a privacy issue in itself. It is sometimes said that "all the data is in the cloud now anyway", but even if there happened to be close physical proximity between sources of data (and there most often is not), that does not equate to the data being more integratable in the sense of being amassed into a single consistent dataset.

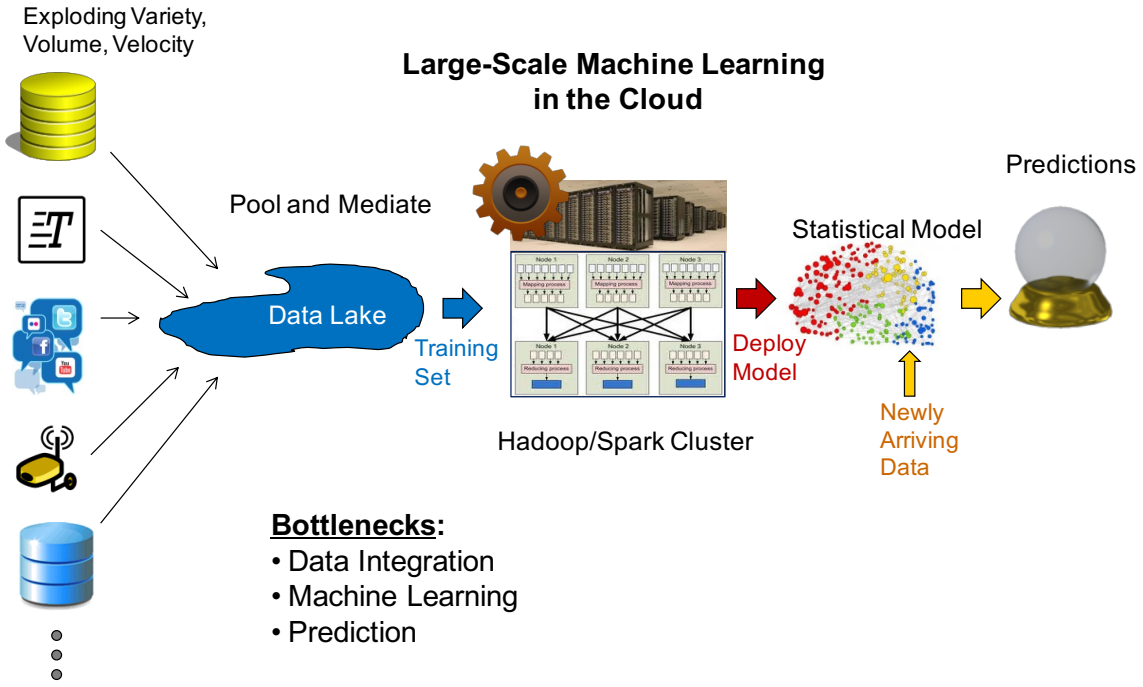


Figure 2-1: Performing data integration to support centralized predictive analytics is complex, gates the analytics workflow, and is expensive.

[4] Fellegi, Ivan; Sunter, Alan (December 1969). "A Theory for Record Linkage" (PDF). *Journal of the American Statistical Association*. **64** (328): *pp.* 1183–1210. doi:10.2307/2286061. JSTOR 2286061.

The subsequent process of cleaning, preparing, and organizing a massive integrated data set for analysis, for example on cloud-based analytics platforms such as Hadoop/Spark clusters as shown in Figure 2-1, has costs that increase exponentially with the addition of more feature-bearing sources, until the clean-up becomes the dominant contributor to the total turnaround time in the workflow. Larger and higher variety data exacerbates the costs and the technical challenges associated with this step, which is often quoted by experts as consuming as much as 70% of the total effort and resource in the analytics workflow[5]. The labor costs alone are painfully high. Estimating just 3 man months of effort per silo, at a burdened rate of $100/hr, gives $50K per silo, and those costs do not scale linearly but exponentially as more silos are combined. And it can be difficult to get beyond the spending as the projects have a tendency to go on and on, with goals undergoing modification, the data changing, and so the spending just continues to grow in size.

A key technical issue that must be addressed in this phase is that of missing data. During the integration process, missing data is typically resolved through imputation to "fill in" what is missing through various estimation processes, and if that cannot be done satisfactorily, records for which too much data is missing may need to be discarded. This means that the size and quality of the resulting master training can begin to be gated by what is available from the sources with the least overall support.

The machine learning of a large centralized model also presents a potential bottleneck. Increasing the size of the data by adding more and more features, or in other words making the data table wider in terms of columns, drives exponential complexity in the feature selection process within the machine learning algorithms. In particular, the amount of training data required to properly fit the predictive model typically grows exponentially with the number of features, leading to fallout in terms of similarly painful trends in run times and/or compute resources required (the so-called "curse of dimensionality" or Hughes effect in machine learning[6]). The impact is on turnaround time in the analytics workflow. There are several reasons why this is a critical limitation in practice. First, there are typically many varieties of models worth exploring (e.g., boosted trees, random forest, SVM, logistic regression, neural nets, …) each of which may have many potential parameter settings, so the ability to quickly experiment to get a combination that meets requirements is essential, but the conventional approach of bringing data to the compute has difficulty scaling up to support the necessary amount of iteration to explore the solution space adequately. Second, if new data comes available at a source that could support learning a better model, then the entire centralized learning process needs to be repeated to take advantage of that data by learning a

[5] "For Big Data Scientists, 'Janitor Work' is Key Hurdle for Insights", NY Times Technology Section, August 17, 2014.
[6] Hughes, G.F. "On the mean accuracy of statistical pattern recognizers". *IEEE Transactions on Information Theory*. **14** (1): 55–63, 1968.

whole new model. It is not possible to perform an "incremental" update to refine the model, which limits the time dynamic adaptive response of the system.

Time to results may be further be gated by the prediction step. Once the centralized predictive model is deployed, for example in the cloud, it will be used to process newly arriving test data to generate predictions. Assuming the new test input data arrive initially to their "point of origin" silo's, the data needs to be communicated to the collection site, conditioned and merged with data from other sources, all before it can be fed to the model.

Traditional "Crank" = Exponential Pain[4]!

- *high cost*
- *complex*
- *slow turns*
- *not private*

Vignette (Continued): Analysis Strategy & Challenges of Conventional Approach

Consider how these challenges manifest in the context of our credit approval example. The conventional approach will attempt to integrate the various data together to support centralized machine learning, and there are several sources of pain in that process. Figure 2-2 highlights a few of the issues.

The first challenge that must be faced in learning a model is gaining access to non-in-house external sources of data that could be useful in supporting credit approval decisions, and ensuring the sufficiency of the data's coverage and quality so that it will prove worthwhile when taken in combination with other sources. Data producers have to be willing to turn the feature data over for integration, and subsequently prove the quality of their data and its value to the credit approval model. Even sources that are clean and high quality on arrival may have disparate formats and semantics. For example, the demographic data would typically be structured, whereas social media data would originate as unstructured text that needs to be post-processed and tagged. Onboarding of certain private sensitive data such as account balances may require that various governance and data management processes be followed, with attendant recurring costs. The snapshot of the resulting data is comprised of a mixture of data flows evolving on different timescales as shown in Figure 2-2.

To support machine learning of a global centralized predictive model, training data must be assimilated that combines features from the various sources of data into a single data set (conceptualized as a large spreadsheet with various source features as the columns, and instances as rows), and associates each instance with a desired target variable outcome, in this case whether the training instance represented a true "good credit" or "bad credit" case. The cycle time of the entire analytics workflow is gated by the time required to assemble this data across the sources and then learn the model. An intrinsic difficulty relates to the fact that the source data is evolving on different timescales, which can create "turbulence" in the mixed dataset (embedded non-stationarities on different timescales).

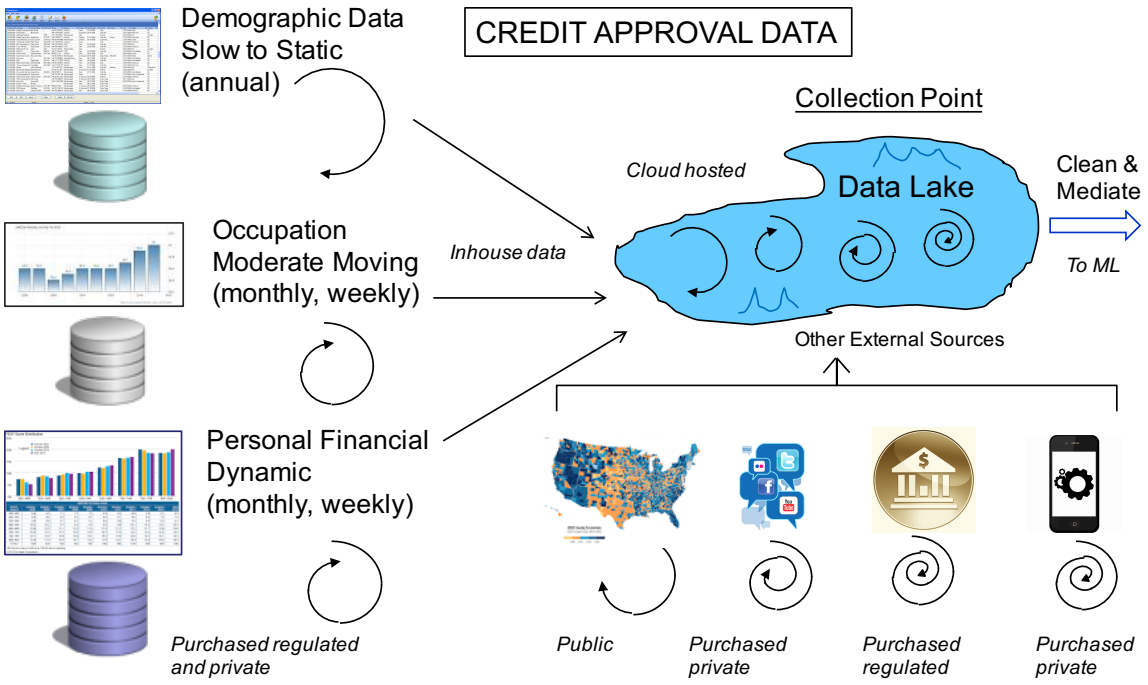


Figure 2-2: Integration of multiple sources of data faces challenges of accessing, gathering, mediating, and dealing with continual "turbulence" from data dynamics.

To illustrate with our example, let's assume that all the barriers of access, gathering and mediation of the data somehow have been successfully dealt with (!), and the features for a sufficiently overlapped training set have been assembled. An instance of that training data for developing the model, along with the currency of that information (timestamp of validity) at the time the record is created, might include contents such as:

Identifier	Credit Score	Employed Salary ($K)	Age (Yrs)	Edu.	Rent or Own	Zip	Utility Pymt History (>30 days)	Bank Cash Balances ($K)	Calls from Collect Agents (3 mo.)	Target
Doe, Jane	780	150	32	MS	Own	94027	None	450	0	1
updated	3 mo.	1 yr.	1 yr.	5 yr.	1 yr.	1 yr.	3 mo.	1 mo.	6 mo.	6 mo.
Smith, Bob	670	75	43	HS	Rent	39203	None	35	2	0
updated	1 yr.	1 wk.	1 yr.	3 yr.	3 yr.	3 yr.	n/a	9 mo.	1 wk.	1 wk.

The features in the table are dynamic and evolving along their own natural timescales. To make the best use of a source of information, it should be sampled and updated commensurate with its evolution, and handled consistently for all the records. For the modeling problem, assembling snapshots of features collected at different points in time, can create a hodgepodge training set that feeds the learning of a model with poor predictive power. If a source contains features with high importance to the decision

being made, and that source is in actuality highly dynamic, such as the utility payment history or bank balances might be in comparison to the demographic data, there would likely be benefit in updating the information and relearning the model to better track the dynamics, but that would require reassembling a new training set and relearning the entire centralized model. If that process takes days, while the data is evolving on timescales of minutes, the benefit cannot be realized, it is too late. The additional challenge is of course that obtaining training data with target variable outcomes on these timescales is also highly problematic.

For prediction, a query for a new user is going to be processed whose profile features are going to be gathered together from various sources and processed through the central model to make a decision. That assembled information may suffer from the same problems of "mixed staleness" that can affect the model build. For example, if the assembled feature profile indicates that the applicant is employed, but she really just lost her job, or the profile says she lives in a prosperous zipcode, but she has actually recently moved to a depressed area with limited job prospects, then the risks of awarding her credit will be improperly evaluated. In the case of our online credit application, the user could be prompted to enter more and more personal information to the web interface, but there are limits to what she may feel comfortable providing, be patient enough to provide, or that could be reasonably validated in real time anyway.

For the combination of data sources we have described, it is likely to be the case that certain sources of data are only "situationally valuable", or are frequently unavailable, or inject useful information only occasionally. This might be expected from features extracted from social media or mobile device exhaust. This creates challenges of missing data on both the model learning and prediction side. A significant amount of experimentation and iteration may be required to design an analytics system to make best use of that data, but the entire centralized approach suffers from a lack of agility in the workflow, and makes it expensive to try many approaches. It is hard to see how it can ever work to its full potential with so many rich varieties of dynamic data coming available. The next two tables summarize the challenges for the conventional method.

Table 2-1: ***Summary: Difficulties in Model Learning, Conventional Method***

Challenge	Negative Impacts
Accessing, Gathering, and Integrating Data from Multiple Sources	Bringing the raw data together has high cost and complexity, compromises privacy
Handling Time Dynamic Information (No incremental model update)	Decision performance likely suffers due to model staleness
Difficult to Scale Up	Turnaround time and cost grows exponentially with addition of new sources
Limited Ability to Iterate / Experiment (high cycle time)	Decision performance likely suffers due to lack of optimization and inability to track changes in data; limited model builds
Testing Value of a Data Source Requires Integrating It	It is high cost to even establish the value of different source combinations
Sensitive to Missing Data	Missing data may require extensive imputation and/or reduction in training samples

Table 2-2: ***Summary: Difficulties in Online Prediction, Conventional Method***

Challenge	Negative Impacts
Data Must be Gathered	Data originating in different locales needs to be assembled to process centrally, and may arrive asynchronously, which can create response latency and waiting on data
Sensitive to Missing Data	Centrally-learned model that was trained using all the data may not robustly handle data that is missing during prediction

3. A New Approach

Create teams of informative analytics, not lakes overflowing with data

A spectrum of possible approaches resides between the extremes of i) analyzing data that originates in silos by integrating the data together and computing on it centrally to create a single global model as just described, or ii) simply leaving the data in separate silos and combining the results of analyses performed locally and separately, as illustrated in Figure 3-1 below. In the case of centralized processing, a global model is constructed across all the data sources by sharing all the raw feature data. In the case of purely local processing, each of the sources is processed without regard to the others, i.e., no data is shared, and the processed outputs are gathered together in a post-processing step. We advocate a

novel approach that sits in-between these extremes, which offers the benefit of providing a global model, but without integrating any feature data.

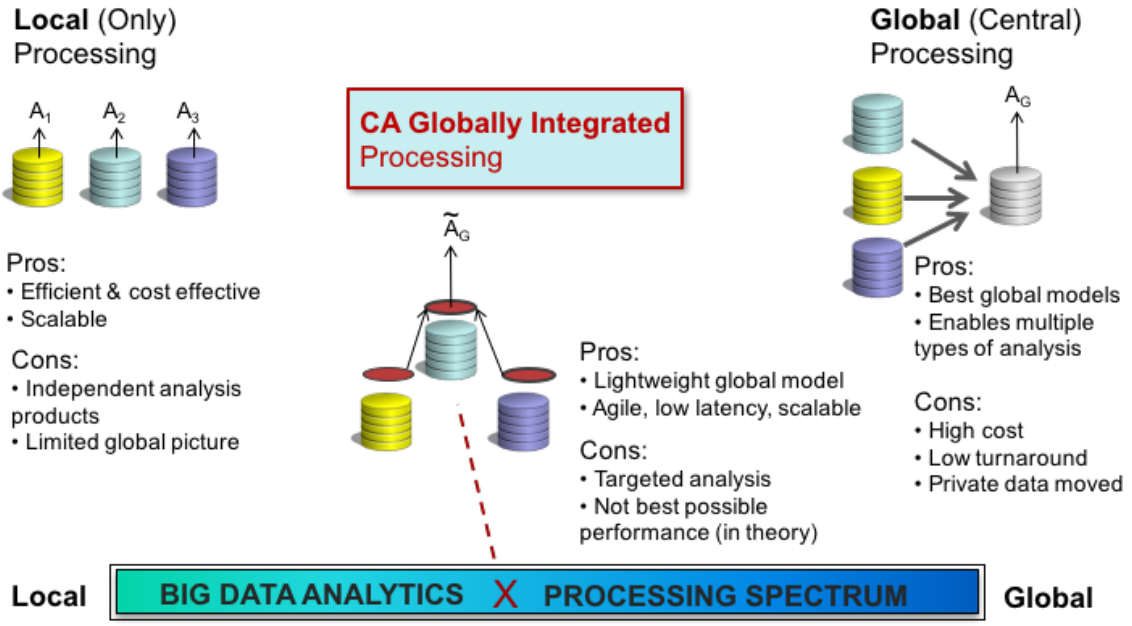


Figure 3-1: Collaborative Analytics (CA) can provide a global model across all the data with low cost and complexity.

Collaborative Analytics (CA)

We propose an approach in which locally processed analytics are coupled into an overall global predictive model, by design. This is not a process of attaching separately derived analytics after the fact, e.g., by combining "votes" from the different sources, but instead the local analytics are interpreted with respect to optimizing the performance of the entire system as a whole. We refer to this as Collaborative Analytics (CA) to convey the notion that the global analytic is derived from a team-optimized decision process, one that provides a holistic treatment of all the silo'ed data taken in totality. Local private data is effectively reduced into compact messages which serve as signals for the other team members, and which have meanings that are obscure without knowledge of the entire system's computation.

Our approach exploits a tradeoff between the granularity with which information is represented, and the cost and time required to communicate and model that information. The design objective for our system was to constrain the amount of information that needs to be communicated to the absolute minimum amount possible, while optimizing performance to push as close to the centralized full-sharing solution as possible. The premise relative to achievable performance is that if the system can be made responsive

enough, through mechanisms of optimal quantization of information and rapid adaptation, then the benefits of being able to operate with more timely data, and more frequently updated models, can offset the penalties from a coarser representation of information. Said another way, ***more aggregate and more agile can compete favorably with finer grain and cumbersome***.

CA operates as a distributed system in the form of a network of cooperating agents that are embedded with the data (organically). The agent networks provide a solution fabric that rides logically on top of the physical infrastructure and native software applications, abstracting away the complexity of the implementation details below, as suggested in Figure 3-2. In Section 5 "How It Works" we provide more explanation on how the system operates.

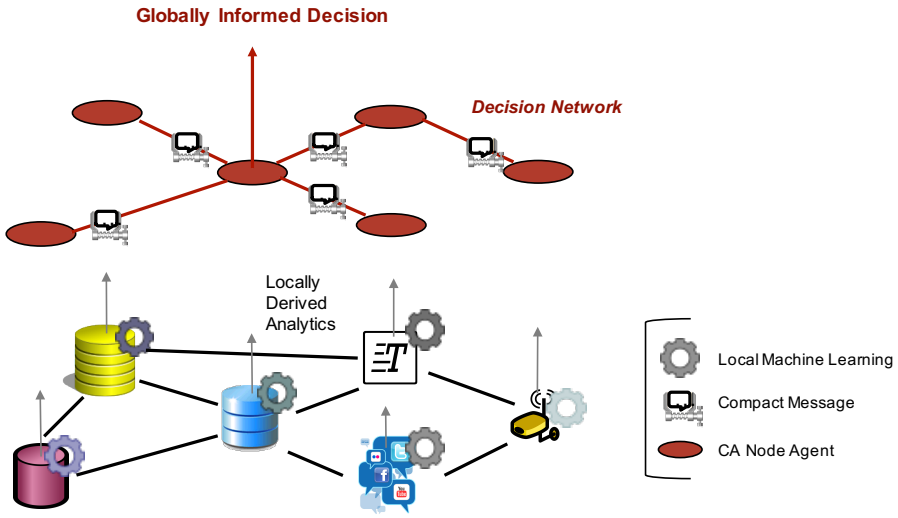


Figure 3-2: Collaborative Analytics bring together information extracted from locally derived analytics to develop a global decision informed by all the sources, while leaving the data in place and not sharing any raw data.

In terms of the analytics workflow, what CA effectively accomplishes is a bypass of the most costly and complex steps driven by data integration, as illustrated in Figure 3-3. Note that CA does not bypass the problem of matching to associate data across the silos. That problem must be solved in the CA system in similar fashion to approaches that integrate data (see Section 5). But the raw feature data resulting from those matches (the column data) is never communicated. The subsequent steps in the workflow of actually acquiring and integrating that data and learning a model over it are high cost and represent the areas of big win for CA, as emphasized in the graphic.

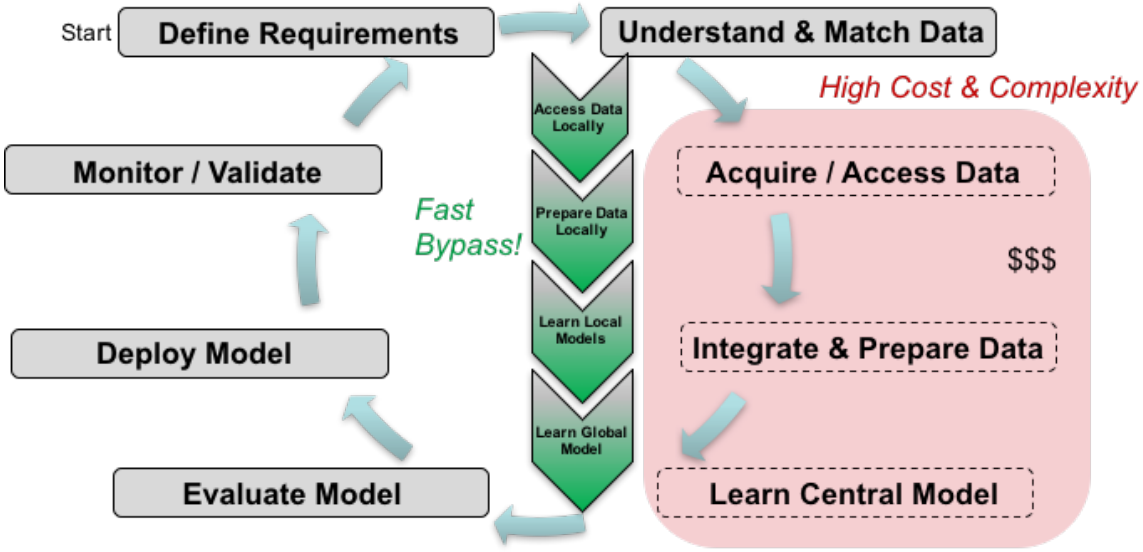


Figure 3-3: CA bypasses the most costly and complex steps that data integration introduces into the analytics workflow, providing accelerated cycle time with lower cost.

An additional derived benefit of the bypass enabled by the decoupling of the model learning problem frees local data scientists to focus on issues related to the data quality and sufficiency in the silos they are concerned with, and they can identify the best techniques and parameters to mine that data locally, and ensure that the information is as solid and current as possible. They don't have to stretch to understand the details of the data and algorithms in use across all the silos. And as data scientists learn better models locally, or adopt better techniques and algorithms, those can immediately be made available to the CA system, and because the learning of the global model is automated and fast, it can be *incrementally refined*, on the fly and while the system operates, without ever having to learn a whole new model. This means that the learning cycle of the system can be greatly accelerated, and if certain sources of data have higher time dynamics than others, those models can be relearned and plugged into the overall system at higher rates to track the changes. New sources can be brought online, or sources removed, and the system will automatically reconfigure itself to make use of them in real-time.

Vignette (Continued): Analysis Strategy with New Approach

Our illustration of using multi-source data for risk reduction in credit approval processing looks *very different* in the context of a collaborative analytics solution. With CA, the problem is transformed from integrating data into combining the results of local queries that have bearing on the global query (awarding of credit), through the aggregate relationship between decision statistics, as illustrated in Figure 3-4. Now the information is being brokered through local models answering local questions, but with *answers communicated in a way that optimizes the overall collective.*

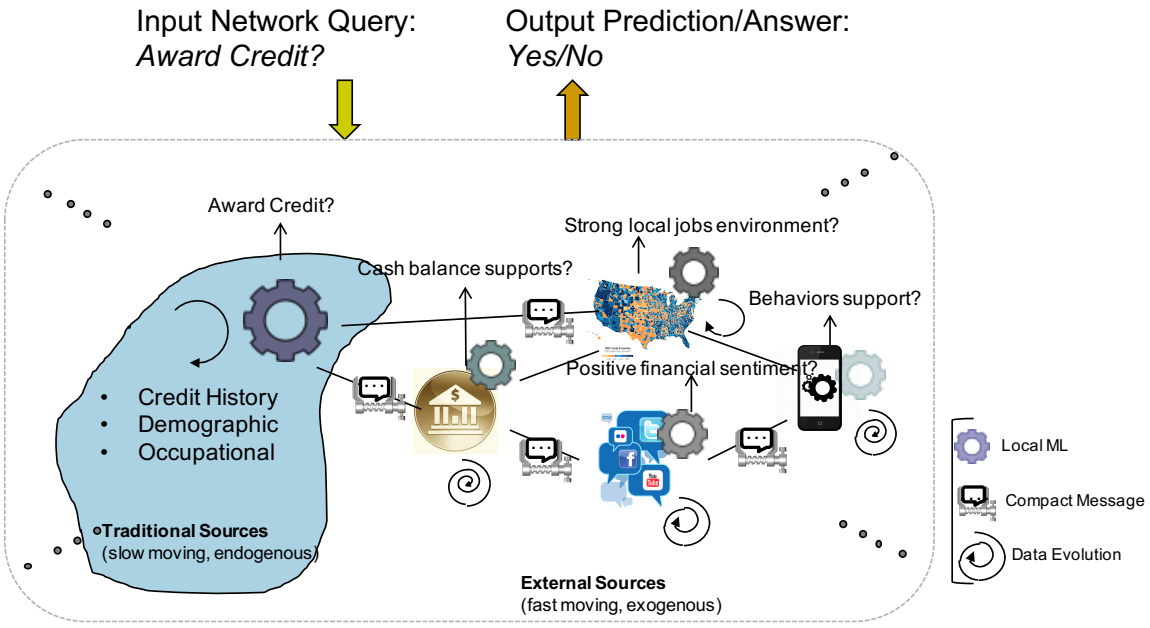


Figure 3-4: CA creates a global distributed information model from a collaborative network of locally learned models. The global credit model is brokered through the answering of local questions that have bearing on creditworthiness. External dynamic sources can be connected to augment in-house traditional analyses.

To explain further, suppose that the traditional sources have been brought under "house control", as suggested in Figure 3-4, and these data are being mined to make credit award decisions. But there are other external sources of information that could provide additional information gain and usefully augment the pending approval decision if they could be accessed. For example, the banks would have information on the current state of balances and trends, social media might have information on your personal sentiments and circumstances based on recent posts, and your mobile device would have information on last few months of buying patterns and call history, especially from credit and collection agencies. Could this information possibly be tapped to support the credit decision with a total view across all these sources, but without compromising your privacy? Yes, and here is how.

It is not uncommon for banks today to offer traditional banking services (checking, savings accounts) as well as credit card services. So they are in position to construct a model that relates cash balances in the accounts they hold, and the augmentation / depletion trends, with the payment histories on the credit they extend. They can build a decision model that evaluates a local-only question like *"Cash balance supports credit award?: Yes/No"* and make a projection of creditworthiness based on the correlation they can see based on the data they have.

Social media sites such as Facebook are a dumping ground for human sharing of emotions to friends and family about relationships, job status, and important life changes

that may represent the most up-to-date information someone has publically released about what is going on in his or her life. Could a company that possesses that kind of information possibly build a model that reliably predicts *"Personal financial sentiment?: Upbeat/Downbeat"*. They probably could. What they need is outcome variables, some indicators as to whether an individual can legitimately be described as financially stable or unstable, based on his or her real world profile and posted sentiments. They may have the data, or be able to get that data. But if someone makes a lamenting post that "I haven't been able to pay my bills for the last 4 months", that could suggest strongly they might be getting into trouble, and not be a good credit risk at this time. Of course, that piece of information would have to be weighed properly, given its uncertainties, against all the other information available.

The professional networking site LinkedIn, with its online resumes and professional networks, is in position to develop detailed career assessments, at least for its most diligent users that keep it up to date. The types of jobs held, educational background, job turnover rate, location of jobs (moves), career trajectory and upward/downward mobility, and network activity level are all obtainable. That data can be mined to answer the question *"Positive job outlook?: Yes/No*". The higher resolution factors are rolled into an answer to that question.

Mobile data provides an interesting case because of all the sources of data available, it is the most "real time" and reflective of what is going on *right now*. Suppose as the credit approver you knew that the individual had received 10 calls from 3 different collection agencies in the last 2 weeks, and suppose you also knew that the applicant had just bought $2K worth of oddball items online in the same time period, spent significant hours playing games on his phone during standard work hours, and also had been an average of over three months late paying the mobile phone bills. That information might raise a question about fiscal responsibility and stability that needed to be considered. That kind of data cannot be shared, but if the mobile data could be privately mined to answer a question like *"Fiscally responsible profile?: Yes/No",* it could provide a useful input to the overall credit awarding process.

An important point is that what is actually communicated when a source is plugged into the overall collaborative decision network is not the best local answer to that question. It is a message that is derived based on having that local model in place, but that is appropriately hedged to consider the performance contributions of all the other source team members in the network. And the question being answered locally never has to be communicated.

Returning to our example, suppose that a significant external event impacts the risk profile and how the various decision factors should be weighed because the environment has changed. A once prosperous coastal town is devastated by a hurricane, obliterating local industry, or a nationwide financial crisis precipitates massive layoffs across segments of the economy, or an election causes cancelation of major

initiatives that directly impact the applicant. The meaning of living in a certain zipcode, or having a specific job title, or working within a certain industry, as it applies to the question at hand about credit risk can change quickly. And in the CA system, local models that are deriving answers to questions such as "Positive job outlook?" are far more quickly updateable than building an entirely new centralized credit decision model.

It may be the case that some sources only have relevant data part of the time. During the model learning phase, overlapped records for the silos included Mary, Jane, and Joe. But during prediction of a decision for Bob, who is not a social media user, no supporting information from that silo can be provided. In this case, the CA system will automatically configure itself to make the best use of whatever information it has, and report the decision along with performance metrics regarding its strength of support in the data. So missing data is dealt with automatically and explicitly rather than requiring artful imputation. Conversely, if some sources have a cost to acquire the information, then they may be selectively invoked to reduce ambiguity only when needed, or paid for only when the data is available and used. This could be the case with an emerging real-time source such as mobile data. It may be the case that as the data evolves in the various silos, certain sources increase in relevance, e.g., because new applications are deployed allowing the collection of new kinds of data, or the timeliness and quality of the information improves, etc. As a byproduct of its operation, the CA system naturally quantifies the "value of information" to the decision process, and so can guide users on what sources are most valuable to include in the decision process. Finally, the entire system scales readily to the addition of new sources as they come available, without requiring any redesign of the existing infrastructure, in fact while not requiring the system to ever even be brought offline to integrate the new information.

CA provides a mechanism to explore entirely new credit scoring algorithms based on new sources of data. One way to structure that effort would be to keep the traditional and regulated models in place and left unmodified, and explore augmenting them with additional sources of information to create derived scoring systems that could supplement the core approach, or otherwise provide risk flags. It is important in this process that *consistency* be maintained, so that decisions across consumers are made on a level playing field and can be compared and justified. CA provides this through a quantitative calculus that is rigorous in the way it weighs information and normalizes results (Section 5). *Transparency* is provided to the participating silos, but not down to the feature level data, by design, but could be selectively investigated if it could be accessed. *Trust* in the information is an issue as it is with any approach, but information that proves unreliable will naturally be de-weighted in the decision process, and ultimately be dropped if not contributing value.

The CA system provides agile mechanisms of information sharing and real-time model update that are just not possible with any other approach. In the context of the real-time credit approval example, the business benefits that derive for data consumers include:

- Reducing bad decisions to award credit (improved decision making through use of more sources of data that are specific and timely)
- Creating the simplest and least costly analytics system that makes use of wide ranging data to support credit award decisions at the required performance level
- Remaining continually flexible to adapt to new opportunities to incorporate new sources of data in the credit award decision process
- Being in position to take advantage of sources for which the raw data could never be acquired and integrated to a data lake, but where information extracted from that data could provide high value in the credit awarding process

And data producers also derive the following benefits:

- Having a means to establish the value of their data to a potential customer
- Being able to sell their data product without giving up control of the raw data itself

CA provides just the capabilities needed to really get at the value in the data.

Table 3-1: ***Summary: CA's Advantages in Model Learning***

Advantage	Positive Impacts
No Data Integration Required	The cost and complexity of data integration is bypassed; private source data can be exploited for better models; heterogeneity in local formats and methods is abstracted away; data scientists can focus on optimizing local performance
Exploits Time Dynamic Information (with incremental model update)	Model is readily updated with new sources and to keep track with time dynamics and volatility in sources being used
Scales Up to High Variety Big Data	New sources are easily added; system can scale to large numbers of sources, providing exponentially more benefit with each source added
Accelerates Analytics Cycle Time	Allows for rapid experimentation, trial and error exploration, fast iterative refinement of objectives and models, rapid build of many models
Efficiently Assesses the Value of Information	The value of combining different sources can be established without a complex data integration step to find out
More Tolerant of Missing Data	Global network model learning requires fewer samples than learning of a high dimensional central model

Table 3-2: ***Summary: CA's Advantages in Online Prediction***

Advantage	Positive Impacts
No Need to Bring Data Together	Highly accelerated response time of system through in-situ local processing, reducing time between signals appearing in data and decisions made
Robust to Missing Data	Explicitly handles missing data by re-optimizing decision performance with whatever data is available
"Sample When Ready" Mode	User can query system at arbitrary times and will receive response along with confidence based on source availabilities
"Sample When Needed" Mode	Sources can be invoked selectively and only when needed to improve the accuracy of decisions

CA's Positioning in the Analytics Landscape

The biggest competition for the CA system comes from approaches that integrate data. Substantial effort is being applied today to accelerate the cycle times associated with integrating and processing very large datasets[7]. The problem is being attacked at multiple levels: faster and more scalable hardware, accelerating data integration and cleanup, better tools for visualization, superior algorithms for scaled-up machine learning. That said, there are practical limitations in terms of cost and complexity that would appear to present obstacles to proceeding endlessly with centralizing data, and at some point even fundamental limits such as the curse of dimensionality ultimately become severe. Because of this, we anticipate that distributed analysis approaches will be of increasing interest and ultimately become a necessity for many applications.

There is an entire continuum of distributed approaches possible, with a tradeoff between communication requirements and performance. ***We have chosen the most aggressively communication constrained approach possible***, to provide the many benefits of eliminating raw feature-level data sharing and running asynchronously. But a variety of alternative tradeoffs could be explored. In the context of the CA system, this would make sense to consider in the case where there are features that reside in separate silos, that when brought together and optimized by machine learning within one single bucket of data provide a meaningful boost of performance. In this case, and if it is possible to do, it would make sense to aggregate the features in those buckets of data into a new bucket, in order to exploit the opportunity for additional performance by selectively eliminating the silo constraint.

[7] "Data Integration Déjà vu: Big Data Reinvigorates DI", SAS Whitepaper, Aug 2015.

Augmenting Data Integration to Improve ROI

While CA provides a competing approach to integrating data for centralized modeling, if an organization is already underway with data integration projects, CA can still provide valuable complementary capability as a low cost and efficient "front end" to make subsequent data integration efforts more efficient and targeted, as suggested in Figure 3-5. The benefits derive from i) CA's ability to highlight which information is most worth combining in the context of specific business questions, and to project the likely performance when those silos are combined, and ii) CA's ability to augment an existing model, built via data integration, with additional sources that do not have to be integrated even to perform the test. These steps can be performed to inform a subsequent full-scale data integration if that is the business' plan.

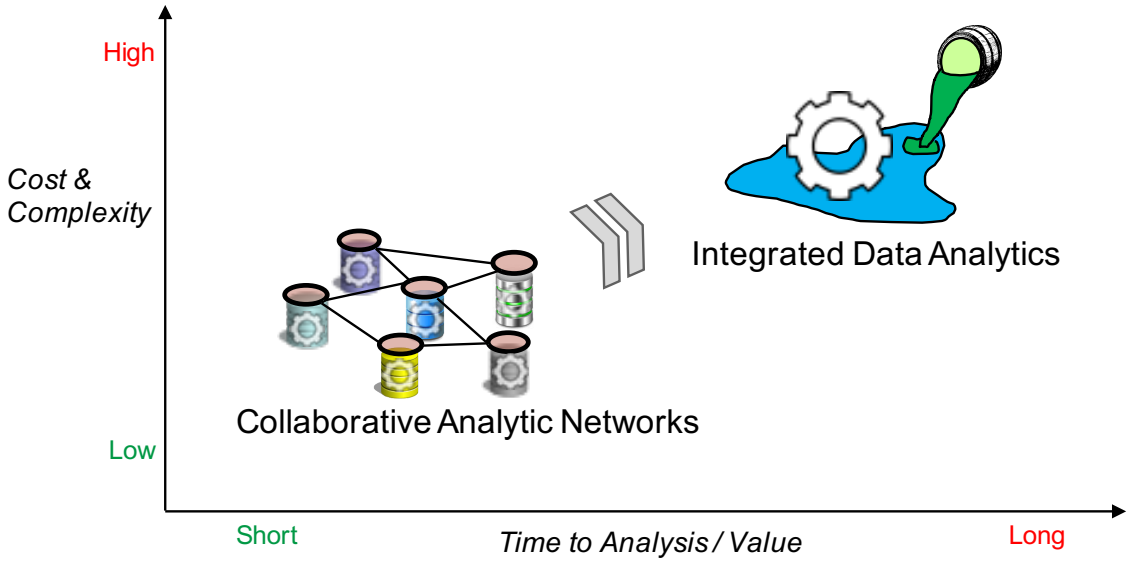


Figure 3-5: The ROI for data integration projects can be improved by adding a supporting CA capability that provides front-end rapid data evaluation and fast time-to-initial results.

To summarize, some of the specific capability enhancements CA can provide include:

- Provide a quick up-front way of doing model selection across data sources to inform subsequent activities of data integration and machine learning
- Provide a mechanism to get up and running fast with a distributed model while the effort to perform full data integration and centralized learning catches up
- Augment an existing (centralized) model that has already been constructed with a new source incrementally, and without rebuilding the original model
- Augment an existing model with a new source for which the data cannot ever be accessed

- Provide ongoing validation of the value of sources and worthiness of expenditures to integrate data, as that data is continually changing

Composing the Analytics Stack

CA plugs into the analytics stack at the top end, as illustrated in Figure 3-6.

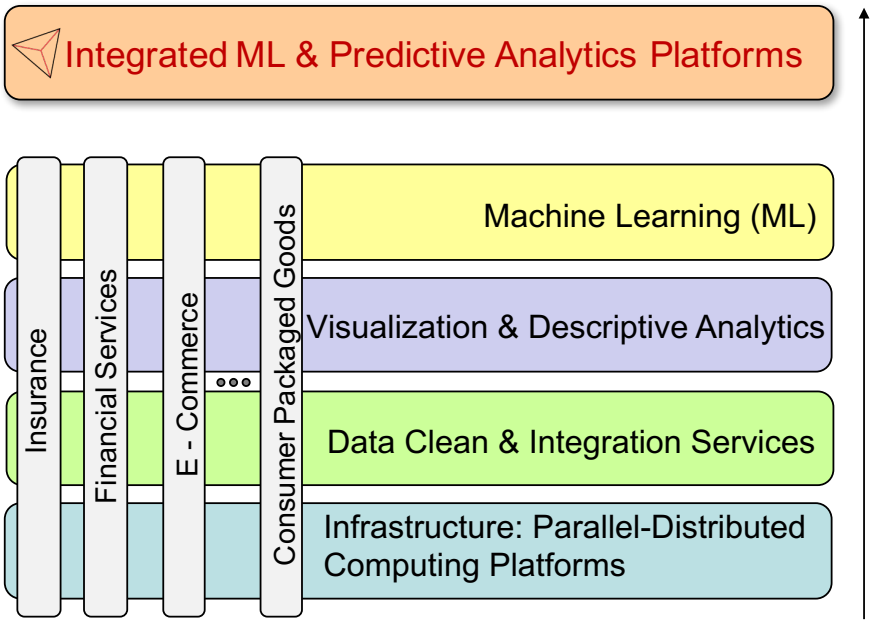


Figure 3-6: CA provides a framework for integrating analysis capabilities through the analytics stack and across application market segments.

CA composes diverse analytics platforms into a unified system

The stack in Figure 3-6 builds upward from infrastructure, to ETL and data integration, to visualization and descriptive analytics, and finally to machine learning algorithms to support predictive analytics. Shown at the base of the diagram is computing and storage infrastructure to support analysis of big data. This would include parallel-distributed computing platforms such as Spark/Hadoop and systems for managing data lakes. A number of companies have emerged to support the substantial work required for data cleaning, aggregation, and augmentation to create consistent and analyzable sets of data. Once the data is assembled and organized, it is typically investigated in manual or semi-automated fashion via descriptive analytics methods and visualizations to understand its content, quality, and interrelationship. If automated analysis engines are to be employed, the data may then be processed using machine learning methods. There are companies that are providing platform/general purpose capabilities at one or more tiers of this stack, and other companies that are providing more end-end capability that is targeted on specific market niches. Prism Informatix is offering a platform capability that serves to combine machine

learning and predictive analytics across multiple installations and silos of data, and so its capability conceptually sits atop of this stack, and leverages the capabilities underneath. This position is uniquely occupied.

4. Business Value

Superior business intelligence at lower cost and latency

The core of the business value proposition is that the CA platform enables users to make better decisions, using more varieties of data, without incurring the costs and complexity to integrate that data. The business benefits that accrue derive from higher quality decisions, accelerated workflows, insensitivity to local data mining infrastructure and interoperability, and the lightweight and privacy-preserving nature of the system's communications.

4.1 Increase Competitiveness

Many organizations see data analytics as an opportunity to increase competitiveness, by helping them better understand customer needs, create better products and services, or improve efficiencies. They may see opportunities to use a variety of different sources of data to improve decision making, but that data is difficult or expensive to access. Control and governance of the data may reside in the hands of different functional groups within an organization, or may belong to other organizations that are reluctant to share it. Even if it is obtainable, by the time it can be accessed, moved over, integrated with other data, cleaned and prepared, the opportunity for advantage may be passed and the costs incurred have driven the bar on successful ROI of the analysis project very high.

An approach based on CA can enhance a data/insights-driven organization's competitiveness in a number of ways:

1. Facilitate experiments with business questions and the availability of support in the data to quickly determine project viability and project attainable levels of performance to ensure efforts will lead to actionable insights
2. Empower different functional groups (e.g., sales, service, support, finance, operations) to query available data from its perspective concurrently, and also tap the data from other parts of the organization without modifying it or moving it, thereby facilitating cross-functional collaborations
3. Exploit more sources of data, including those that cannot be easily accessed, or accessed to obtain the raw data at all, to make more informed decisions quickly and efficiently
4. Reduce the analytics cycle time required for model/evaluate/deploy/validate to accelerate time-to-analysis results

5. Keep talent and expertise requirements manageable by assigning personnel to become expert in local sources without requiring them to understand everything about all the data across the silos
6. Validate other models that are combining multiple sources of data
7. Eliminate the recurring costs of data management and governance by enabling use of data without "onboarding" it; save cost by avoiding integration of information from disparate systems and in different native formats into data lakes; use sources as long as they are valuable and then disconnect from them

4.2 Quantify the Value of Information

It can be expensive and time consuming to bring the data together to figure out whether or not it is worth bringing the data together! What is possible in principle may not be realized in practice due to the quality or timelines or specific content of the data. The CA system provides value by helping users to understand the benefit of putting data together without incurring the cost of actually integrating data to find out.

To illustrate, Figure 4-1 diagrams a scenario in which an organization is operating an in-house data lake to support analytics, and four additional sources of data are available that may be useful, but with different associated costs and timescales to acquire. The question is what sources to put together to best answer the user question, in the most cost effective way?

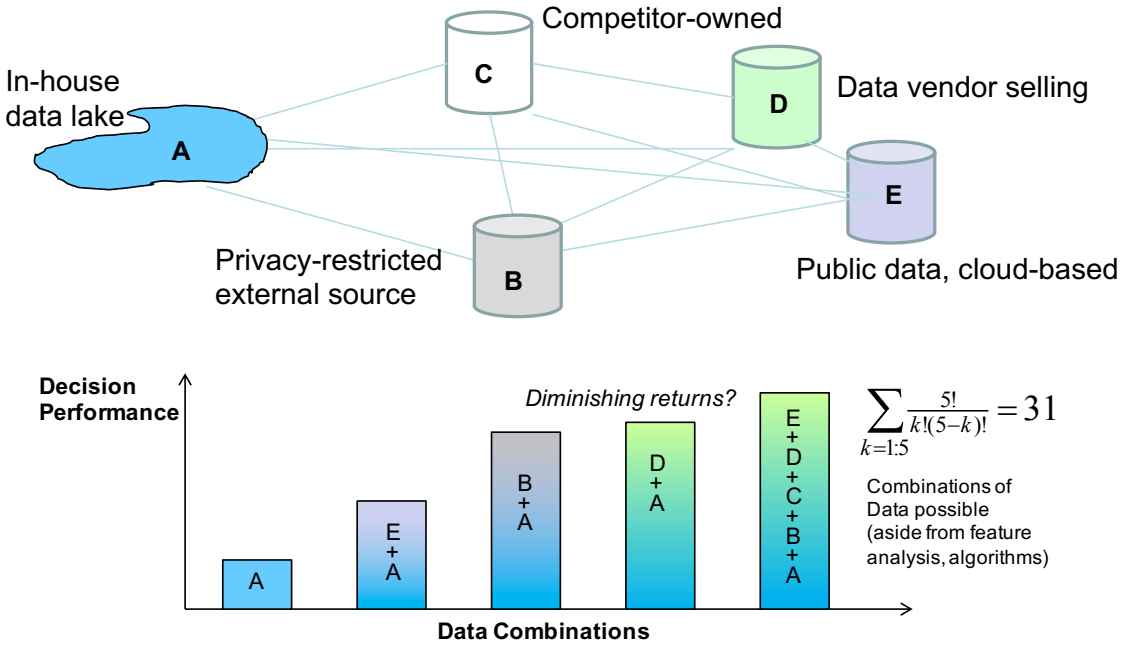


Figure 4-1: What combination of data provides the best value? CA provides a cost effective way to evaluate decision performance versus cost for various combinations of sources, without integrating any raw feature data.

An approach based on CA can help organizations efficiently navigate the data modeling challenge:

1. Efficiently assess the "value of information" to guide data consumers in design of a decision system with the highest performance at the lowest cost and complexity
2. Prioritize sources in terms of decision performance to guide consumers in determining which sources are most worth having, without the cost and time delays of actually having to integrate data to find out; help consumers understand the next most valuable source of data to add, and how much performance gain they are likely to realize for the cost of connecting to it
3. Clarify for consumers the ongoing value of different sources in the face of changing data and the arrival of new source options
4. Provide data producers a means to establish information quality and value with respect to specific business questions

4.3 Future-Scale Your Systems

Organizations are continually motivated to improve internal processes and information systems to improve efficiency and stay competitive. There is constant turnover in hardware platforms for computing, networking, and storage, and in the software applications riding on top. The availability of cloud-based services is creating another dimension of complexity in the decision process about where to store data and where to analyze it, and part of the consideration can be the proprietary nature of the information. It is in an organization's best interest to maintain as much flexibility and lack of dependence on specific systems as possible as it looks to the future. And the ever increasing amount of data coming available brings with it challenges of scale-up in the storage and analysis equation.

The CA framework provides a valuable degree of abstraction, and hence insulation, from the details of the hardware and software in which data in the local silos resides and is mined. This is because the collaborating silos are only loosely coupled through lightweight statistical models, and the details of the implementations are abstracted away and are suppressed "underneath". As is described further in Section 5 on How it Works, the CA software is non-intrusive, in terms of size and compute and communication, and is easily deployed on a variety of hardware and software platforms, making it highly immune to the details of these systems.

An approach based on CA can provide organizations with a predictive analytics capability that is both future-proof in terms of infrastructure evolution and also can scale-up to ever increasing data volumes with the following advantages:

1. Provide an integrated predictive analytics capability without requiring interoperability of systems, and that does not require redesign as the systems underneath are migrated or applications upgraded or replaced
2. Make it straightforward for companies to leverage public-private cloud configurations for business decision making, with an integrated predictive analytics capability that is flexible enough to adapt as data is moved around in hybrid public-private cloud configurations
3. Accelerate the use of information residing in new systems that are acquired through M&A activity, prior to the expense and time of integrating systems, and help to prioritize what information would be most useful to integrate first
4. Continue to extract value from legacy systems, even as those systems are being phased out of primary business operations
5. Reduce Total Cost of Ownership of data infrastructure and analytics applications, business intelligence systems

4.4 Monetize Your Information

A commonly encountered barrier to sharing data is that it is viewed as proprietary, or is owned and controlled by other organizations, or comes with other certain restrictions that limit sharing. This is especially true in a B2B setting where the data that companies collect is a core asset of the business. But there could be power in the analysis if that data could be integrated. This also applies in scientific endeavors, and particularly in sectors like health care and defense.

With the CA approach, the raw data is never communicated, only information derived from that data in the form of abstracted models and privacy-preserving statistics which cannot be interpreted without the context of the entire system computation. And the messages can of course be encrypted or embedded in other transmissions to enhance the security further. This means that the CA framework can be used to broker information between producers and consumers of data in a way that preserves the value of the core raw data asset. For many businesses, this could provide a path to additional revenue streams by selling information products derived from their data, without compromising the data itself.

An approach based on CA can provide organizations a way to share information without compromising data, opening up new revenue opportunities:

1. Partner with other businesses to share information that improves decision processes of mutual interest, and/or buy and sell derived analytic models
2. Gain access to information sources that were historically desired but always unavailable, which when combined with a business' existing information, enable creation of new valuable information products

5. SOLUTION: HOW IT WORKS

The system for Collaborative Analytics described here is a classification-based method (binary) for performing predictive analysis on distributed data while leaving the data in place. The CA capability originates from viewing the problem of performing classification over distributed silos of data as a *distributed optimization* problem. A key aspect of that design process is in deciding what specific information should be shared, and how much of it to share. Note that at one extreme, even centralized optimization algorithms can be implemented as "pseudo" distributed algorithms by decomposing the problems in some fashion and then sharing large amounts of data back and forth between processors in synchronized fashion. Of course, the performance in terms of runtime convergence may be very poor. In designing the CA system, we opted to target the far other extreme of the design space, which was to minimize or totally eliminate as much as possible the sharing of information, and to operate asynchronously, with no notion of a global synchronizing "clock". Our interest in this kind of design was more than just efficiency, but also to provide a mechanism to leverage sources of data as part of a decision process even when the raw data itself cannot be obtained, thereby allowing sharing of information across boundaries, and also to enable operation over the wider Internet.

Secret Sauce

The design of the system rests on three pillars of algorithmic novelty:

- ***Collaborative Decision Engine***: the most challenging problems in distributed decision making are in deciding what should be communicated (message content) and to whom (communication topology), and how to weigh the information appropriately (the information calculus); the method in our system is a principled and quantitative approach that derives from the field of team theoretic optimal statistics, and is based on optimized hedging strategies, not voting schemes; this approach provides superior performance with an internally self-consistent, systematic, and non-ad-hoc approach to processing that is also capable of predicting its own performance versus the best that could be done by centralizing the data

The power is in the distributed algorithms for machine learning and inference

- ***Distributed Machine Learning Algorithm:*** the aggregate correlation structure across the various silos is represented with a graphical model whose structure

and parameters are both identified from data in terms of identifier variables and target variables only, requiring no raw feature level data to be exchanged; the learning algorithm is fully distributed

- ***Hybrid Generative/Discriminative Data Mining Architecture***: data-driven discriminative methods are applied to mine the individual silos and build local models, but those are then integrated or "fused" into an overall global decision model using generative methods which enforce an intelligent degree of sparsity, thereby leveraging the strengths of both categories of approach into a single system; the handling of the interfacing between these classes of methods is performed via a unique methodology

These algorithms have been engineered to create a practical distributed decision system that is robust and highly efficient in learning and predicting from data.

System Implementation

The system is instantiated as a collaborative network of agents that self-organize and automatically design a decision engine to answer user queries. Specifically, the node-agents collaborate to learn a team-distributed predictive model that integrates models learned locally and independently at each data source, each of which can be itself a big data problem. The learning of the global network model is a fully-automated process. The node-agents collaborate to generate a team optimal prediction, with a lower error rate than any of the local models, using the team-distributed model together with predictions from each local model. This is possible because the local analysis agents understand the performance characteristics of the collaborating agents at other sources, and hedge communication appropriately for them within the context of the overall collective. Scalability to data sources comes naturally since local data mining at each additional data source is performed independently and the difficulties of data integration are avoided.

It is important to note that although full scale data integration of disparate data sources is not necessary, data matching to link records across all sources using key or identifier fields at each source is required for two purposes:

1. In learning, matching of target variables across data sources is required to learn a graphical model representing the aggregate correlations between the decision problems at each source.

2. In prediction, the global query needs to be mapped to individual data source queries in terms of the source identifier fields.

The CA system uses a semi-automated, schema matching and mapping based

approach[8] to data matching. Source schemas for the identifiers at each data source are matched and joined to form a global logical data model. Mappings between attributes in this global model and the source schemas are then defined. These mappings form the rules that map the global query to individual data source queries for prediction. The mapping rules only translate to match the format of the data. Matching the identifier data itself, both for linking target variables in learning and retrieving data records at the individual data sources for prediction, uses entity resolution incorporating fuzzy matching[9].

It is assumed with this method that the data is being mined locally in each silo with respect to a question that is at least related to, and could be the same as, the global question being posed of all the data, and that standard data mining performance characterizations are being done to evaluate performance (ROC, confusion matrix, etc)[10]. The individual data sources themselves may be big (volume, velocity, variety), in a way variety of formats (e.g., structured, unstructured text, etc), and can be processed on any kind of hardware (e.g., individual servers, Hadoop clusters, etc.) because the CA system characterizes critical aspects of performance and then abstracts away these implementation details underneath. The system is comprised of a logical overlay network that resides on the physical network connecting the sources, as illustrated in Figure 5-1 below. A logical decision network provides inter-agent peer-peer communications. The analysis solution for each user query takes the form of an agent team that corresponds to a model network. Multiple model networks can be stacked to answer different queries concurrently.

[8] Bellahsene, Zohar, et. al., Schema Matching and Mapping, Berlin: Springer-Verlag, 2011.
[9] Christen, Peter, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Berlin: Springer-Verlag, 2012.
[10] Note: creation of a local analytic model provides the mechanism for private information sharing; however, it is also possible to bring in an actual feature of the data, such as a discrete variable like a geographic region or a seasonality, and use that to select from among multiple CA models

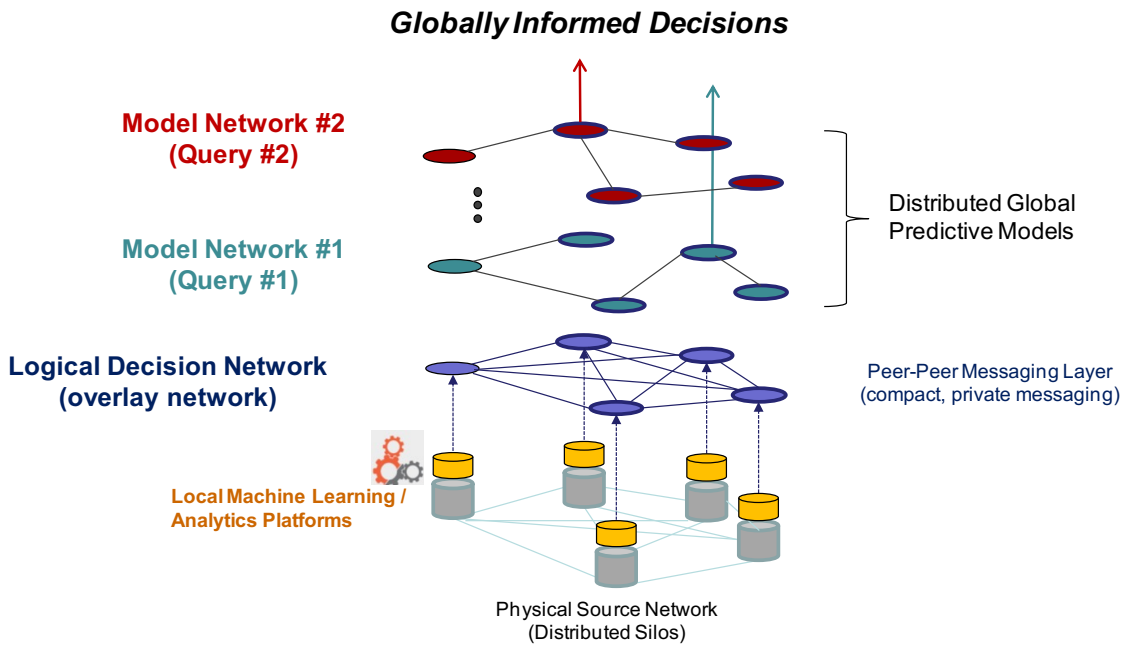


Figure 5-1: CA operates as a layered system of agent teams, where a layer corresponds to a question, and the layers may be segmented into applications that make use of the same data concurrently.

The system automatically adapts and configures itself to new questions and data sources. Due to the efficiency of the distributed learning algorithm, this adaptation is orders of magnitude faster than learning the local models, once the local models have been learned and deployed. This ability to adapt the system rapidly also contributes to its fault tolerance and insensitivity to missing data since the system can reconfigure itself around node/network faults and missing data at any data source, re-optimizing on the fly. The learning of the global predictive model requires exchanging only a modest amount of training data, due to its coarser representation. This data only involves identifier and binary outcome variables, no raw feature data. So the training data would be very light, and something of the form “Susie 0, Bob 0, Joe 1, …”.

CA node agents interact with the existing data mining tools in use, such as R, SAS, scala, Python, etc. Figure 5-2 illustrates a networked collection of silo’ed sources in which local data mining and machine learning is being performed independently, using local toolsets and heterogeneous methodologies. To deploy CA at a local site, an agent is installed that is associated with the source (it need not physically reside at that source). The agent is connected to local data mining through an interface to the machine learning / model construction software, shown in Figure 5-2 as a plug-in to the local machine learning toolset. The agent also interfaces to the local scoring engine through an adapter that supports prediction from newly arriving data. The agents are servents and collaborate with nodes at other sources.

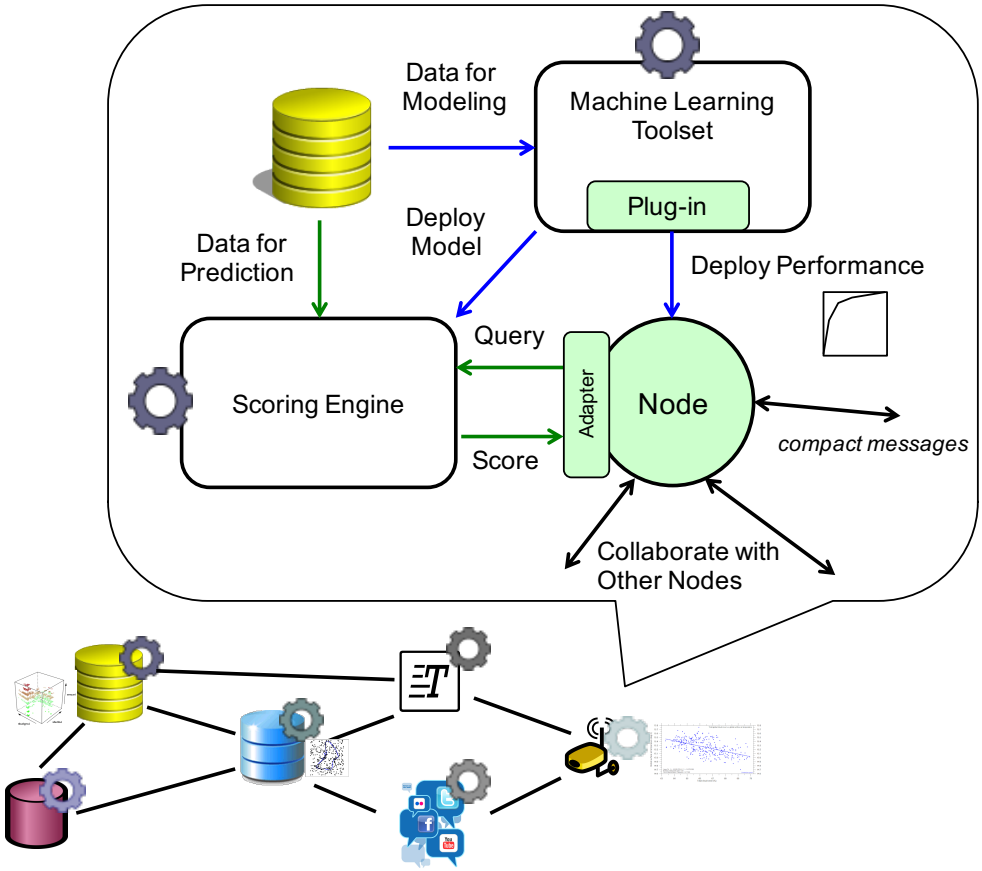


Figure 5-2: CA agents are easily deployed & low impact to existing systems. Prism software components are shown in green. Learning related interfaces are shown with blue arrows, prediction related interfaces with green arrows.

The system exposes web services for clients who can connect to the system using browsers, as illustrated in Figure 5-3. Web clients are used by business analysts to pose queries, construct decision networks, and make predictions from new data. Multiple users can be mining the data concurrently and with different objectives because of the layered partitioning of the system, as was previously shown in Figure 5-1.

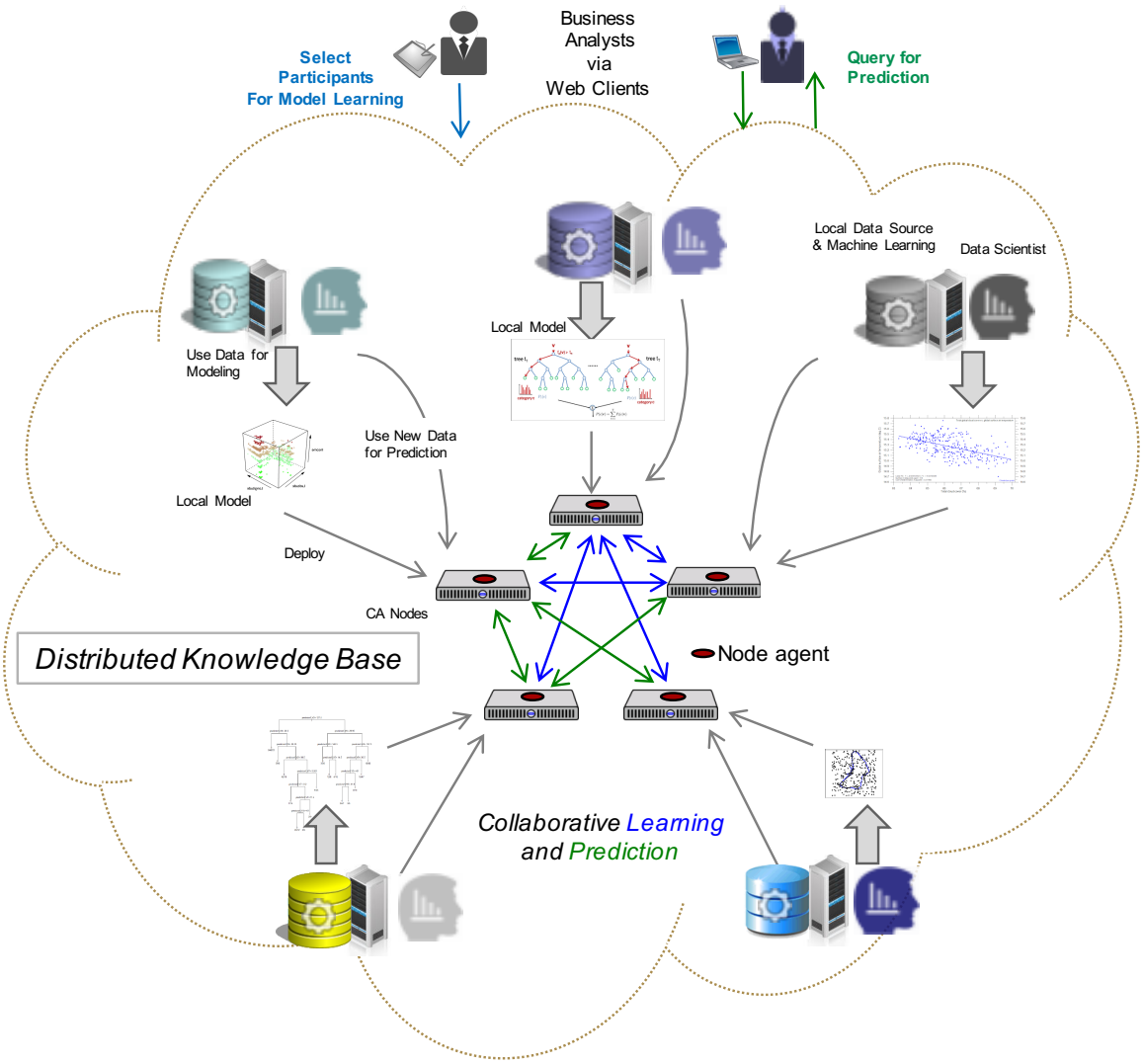


Figure 5-3: The CA system, depicted here for five silos[11], provides web-based access of distributed heterogeneous data stores to business analysts. To users the network of sources appears as a single system, and users do not require knowledge of where the data used to answer their queries even resides. Multiple users may be running multiple different queries on the same data concurrently.

[11] Node agents can reside entirely within local data mining infrastucture. They are broken out separately in the graphic for clarity.

A key point is that CA does not change the existing data mining workflow at the local sources. It operates within the context of a standard analytics workflow of Learn-Evaluate-Deploy-Validate, and does not require that the local analysts do anything different than they would already be doing to evaluate the quality of their models. CA sits on top as an overlay, and augments local workflows by connecting them to other workflows happening at other sites. The data analysts at each data source can continue to use their existing machine learning tools and methods to learn the local model, in whatever way gives the best performance, providing minimal disruption to ongoing operations and low cost to adoption. It is straightforward and non-disruptive to try out.

CA is Easy to Try

- *Sits on top*
- *Lightweight*
- *Does not impact local workflows*

6. IMPACT

It is likely that important opportunities for economic and scientific progress will remain bound up in the various scattered and unwieldy datasets that are being collected, while waiting for a cost effective and efficient means for exploiting the information captured in all that data. And it seems clear that the default approach of pushing hard on the "business as usual" conventional technologies that require data integration and centralized machine learning is unlikely to scale to have sufficient runway to address the overall long term need. With approaches that involve sharing significant amounts of data across the network, and moving it to centralizing processing, data aggregation and semantic unification become fundamental pinch points, and exacerbate challenges related to privacy of information. The ability to share information efficiently and confidently is recognized as the core challenge to successfully unlocking business value in high impact data-intensive application spaces such as IoT[12].

In this paper, we have presented a radical alternative to conventional predictive analytics, one that pushes toward the other extreme of fully distributed analysis frameworks. In particular, we have described an approach to obtaining a globally-informed predictive model across distributed silos of data, without sharing any raw data. ***This capability is the only one of its kind***. The approach represents the most extreme of all possible implementations along a continuum exploiting the tradeoff between communication cost and performance. We have targeted a system design that requires sharing the minimal possible amount of information and that can operate asynchronously. It embodies an "anti-data integration" paradigm.

[12] "Data Sharing and Analytics Drive Success with IoT: Creating business value with the Internet of Things", MIT Sloan Management Review, Report Fall 2016.

This unique capability is instantiated into a software platform for collaborative analytics that overcomes the barriers that have hampered previous efforts to exploit high variety data at scale:

- Learning and prediction are distributed organically with the data, providing *de facto scalability* to data sources, and addressing head-on the exponential growth in cost and complexity for integrating data and learning large-scale centralized decision models
- Efficient distributed learning algorithm allows *rapid exploration* of different combinations of data sources to determine which sources are most important for analysis, and prioritize which ones to include.
- The analytics cycle time and time-to-results are *vastly accelerated*.
- Global prediction model is *learned incrementally* as data sources are added, updated, and/or removed, and while the system continues to operate, providing continual adaptation to new data and resiliency to changes in availability.
- System communications employ compact messages that obfuscate the underlying data, *preserving privacy*.
- There is minimal impact to existing predictive analytics workflow, lifecycle, and investment providing *low cost to adoption.*
- Prediction decisions are made in *real-time* and optimized on-the-fly to make use of whatever data is available.

These benefits have been realized in a software platform with multi-application reach, and that has a well-described and efficient deployment strategy as described.

Now invest with confidence to realize the full value of high variety distributed data At Scale

New Market Mechanisms for Information Exchange

New approaches can unlock new opportunities.

Our world is being transformed by the growth of a data economy in which there are suppliers and consumers (demanders), and a desire for information to become more "commoditized". But viewing information as a traditional good that is bought and sold for profit is somewhat problematic, particularly in view of the fact that the more sensitive, least known, and most timely information often holds the most value to know. *The most significant barriers to trading in information have always been i) establishing its value, and ii) defining control and ownership.*

Collaborative Analytics provide a mechanism to enable profitable and confident trading in information because the value of adding data to a decision process is readily quantified, and because data can be used while leaving it in place and protecting its

content. The right kind of incentives exist for both consumers and producers in the approach, and consumers can pay for just what they use in a transactional way. Overcoming these barriers clears new avenues toward commoditized information and marketplaces.

The big idea here is **Collaborative Analytics**, and new opportunities with Big Data are now possible as a result.

About the Authors

Alan Chao and John Wissinger were once graduate school officemates at MIT's Laboratory for Information and Decision Systems (LIDS), and have since sustained a twenty-five year history of professional collaboration. They are engineers with a shared passion for distributed algorithms and architecting mathematically intensive distributed systems. They founded Prism Informatix to offer a new way of managing complexity and realizing value from distributed Big Data, with an aim to advance endeavors of economic and societal consequence. Alan and John always appreciate discussions on how Collaborative Analytics could address the requirements of your application.

A NEW WAY TO SHARE THE WORLD'S INFORMATION

*** Contact us to learn more ***

Prism Informatix, Inc.
1 Watermill Place, #223
Arlington, MA 02476

contact@prisminformatix.com
www.prisminformatix.com
(520)-991-8381

Appendix: Terminology Backgrounder

Analytics can be distinguished into the categories of “descriptive” and “predictive”. The descriptive category involves assembling historical data together, and computing/extracting statistics from it or using unsupervised machine learning algorithms like clustering, in order to make assessments and support visualization and interpretation, generally in semi-automated fashion. Data from diverse sources might be analyzed together by performing independent analysis on the sources, and then combining them through distributed/federated database techniques such as Google Big Query, Netezza, etc., or even combining “by eye” on a dashboard. **Predictive analytics** are techniques which use features extracted from historical data that have been understood with respect to an outcome of interest, such as whether a loan default occurred, an investment paid off, etc., to support the supervised ***learning*** of a statistical model that is then deployed to make ***predictions*** on newly arriving instances of data, in automated fashion. The most common method to create a predictive model from many sources of data is to combine that data together into a single master training set for learning a centralized model.

Predictive analytics are performed using two principal methods: regression, which produces a continuous-valued score as an output, and **classification**, which produces a discrete class label. For instance, a regression model might predict the time until a device is likely to fail, while a classification method would predict whether or not the device was going to fail (yes/no). The method described in this paper is a classification-based method, although it may be used to augment regression methods.

We are concerned here with the particular challenges associated with **“high variety”** aspect of big data that is distributed across multiple silos. What we mean by “high variety” is that the data is distributed by feature types, as opposed to samples of the same kind of data. From the conventional data mining perspective, samples or instances of records typically correspond to “rows” in a table, and features of the record correspond to “columns”. For example, the rows might be indexed by an identifier like customer or product name, and the columns might include information such as how much money that customer has spent in the last year, whether her account is paid up, what products she bought, whether she responded to a marketing campaign, the service history on those products, her customer satisfaction rating, etc. We are treating the particular class of problems in which the various features of “customer” residing in the columns are actually distributed across multiple silos, e.g., are maintained in different business systems such as service records, product configurations, marketing data, and financial data might be.

It is not a priori obvious that more data is necessarily better, in terms of leading to better analysis results. However, the desire to use more disparate sources of data, more

targeted (identifier specific) data, more timely data, and to incorporate more features extracted from that data, is motivated by the fact that predictive models produced by machine learning often improve, meaning they become more accurate, when trained with a wider variety of complementary and up-to-date sources. In other words, the quality of models often improves when more "columns" of data are used to learn the model. Hence the interest in high variety and complementary source data.

A **data lake** is a storage repository that can be scaled to hold a large amount of data including structured, semi-structured, and unstructured data in its native format. The data is simply pooled until it is needed, at which time the structure and requirements are imposed.

In terms of business process surrounding analytics, many companies start with a process of defining requirements and questions they want to answer, then identifying the available data, cleaning it up and assembling it, either by integrating it, or creating federated systems that provide virtual integrated views, just so that they can take a look at what they have. The next step is most often doing some descriptive analytics analysis to figure out how to make use of it. After that, the feature extraction and machine learning of a predictive analytics model may come as a downstream process, as it requires more sophisticated capabilities to design and deploy.

The Collaborative Analytics (CA) system is designed around **natively distributed algorithms** for learning and prediction. By "distributed algorithms", what we mean is that the data is processed in place, and the allowed communication is restricted to be compact, and can tolerate intermittent availability and timing delays, which are manifestations of the "loose coupling" often used to describe distributed systems. We also mean that only local state information is maintained at the various nodes in our system and a full global state estimate across all the data is never explicitly constructed. So our use of the term "distributed" should be understood to be in stark contrast to the use of the term to refer to distributing computations such as large-scale optimizations that have been mapped onto parallel computing clusters such as Spark/Hadoop, which often involve the shuffling of large amounts of data between processors as well as global synchronization. To distinguish the two, we would refer to those systems as expediting the computation for a natively centralized algorithm by parallelizing it through distributing data and reassembling partial results (so-called "scatter-gather" operations).

When we use the term **"real-time"** in this paper, it is with respect to analysis latency, and means the use of newly arriving data to provide dynamic analysis output, with near to zero latency from the time that data is available for use, or in practice, within seconds or less.