WHITE PAPER

DISRUPTIVE DATA SCIENCE SERIES

Transforming Your Company into a Data Science-Driven Enterprise

BY ANNIKA JIMENEZ

INTRODUCTION

Big Data is the latest technology wave impacting C-Level executives across all areas of business, but amid the hype, there remains confusion about what it all means. The name emphasizes the exponential growth of data volumes worldwide (collectively, 2.5 Exabytes/ day in the latest estimate I saw from IDC), but more nuanced definitions of Big Data incorporate the following key tenets: diversification, low latency, and ubiquity. In the current developmental-phase of Big Data, CIOs are investing in platforms to "manage" Big Data.

But there is an emerging realization across public and private sectors that there must be more to "Big Data" than just data and platform. CIOs must transform these Big Data platforms and the data they house from cost-centers to data-monetization engines. Forrester likens this very transformation to "refining oil", and Pivotal believes data science is at the heart of the new oil rush.

Big Data emphasizes volume, diversification, low latencies, and ubiquity, whereas data science introduces new terms including, predictive modeling, machine learning, parallelized and in-database algorithms, Map Reduce, and model operationalization. Don't worry — I am not going to get bogged down here by the debate on the definition of a data scientist.[1]

Instead, I want to emphasize a more important point regarding this new vernacular: It infers an evolution beyond the traditional rigid output of aggregated data: business intelligence. It is a use-case-driven, iterative, and agile exploration of granular data, with the intent to derive insights and operationalize these insights into down-stream applications.

Examples of data science in action are plentiful and also well documented, both in a recent Harvard Business Review article, and these Pivotal presentations:

Click on following titles to access full information

- Harvard Business Review article
- Pivotal presentations

Considering the growing number of use cases relevant to your enterprise, what has changed is that these are no longer confined to legacy data-driven functions such as marketing or finance.

[1] This is a well-blogged topic. We have our definition. It more or less aligns with other earlier definitions, and emphasizes the combination of programming skills and statistical knowledge. I'll save this debate for a different post.

Instead, they can and should involve nearly every functional organization in the enterprise. Given the numerous opportunities data science offers for virtually every functional organization in every sector, it is now no longer enough to be a "data-driven enterprise." Instead, you must build a data science-driven enterprise, a.k.a. the predictive enterprise.

As I build out Pivotal's Data Science Services team, I'm in a very privileged position to witness, first-hand, the organizational transformations underway across nearly all sectors, private and public. Based on these observations as well as my own personal experience at Yahoo!, I believe that any enterprise initiative to move towards more pervasive utilization of data science is disruptive on multiple levels. Any CIO, CTO, or even CEO considering this strategic shift must place the transformation front and center at the C-level table, or risk significant erosion of comparative advantage to the first movers who are getting it right.

WHY A DATA SCIENCE-POWERED TRANSFORMATION IS DISRUPTIVE

With the very initiation of a data science-powered transformation, the endeavor and whoever is driving it are acknowledging that the status quo for analytics utilization does not deliver against the believed potential value for the given business (however defined). As a result, any individual (wherever they are on the totem pole), technology, and even organization overly associated with legacy or the status quo will find themselves exposed to some degree of uncertainty and possibly even vulnerability. What transpires when an enterprise kicks off such an initiative ranges within two extremes. On one side — the "bad outcome" — the effort yields a hot mess of organizational wrangling over concepts like "data ownership" and "where analytics should live", shortsighted technology investments or the digging-in-of-heels around legacy platforms, and analytical project work-to-nowhere — all with the enterprise's competitive advantage wavering on a precipice.

On the other end of the spectrum — the "good outcome" — the effort can yield a complete rebirth or transformation of the company built upon data-derived innovation, resulting in data science-generated intellectual property and competitive advantage. On this end of the spectrum, I often see an executive alignment on prioritization for data initiatives, a thriving data science culture, a drumbeat focus on smart data instrumentation and data quality processes, and modeling efforts with clearly defined paths to operationalization for top- or bottom-line impacting actions.

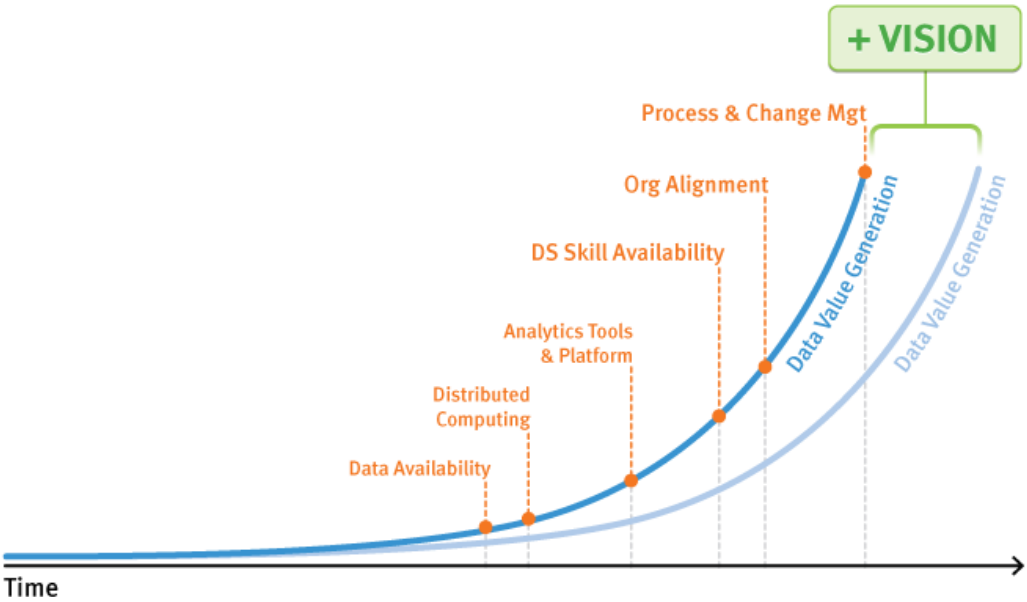
I am frequently asked by our Analytics Labs customers which levers they should control to drive towards the "good outcome" as they embrace data science. The levers are numerous, and each is integral to the success of the effort. I've mentally cultivated my list over an extended period of time, based on my team's data science work with our customers and prospects, my observations of the travails and successes of various Pivotal customers, and my pre-Pivotal days at Yahoo!, where I ran central Insights Services and led globalized data solutions during the company's data "glory days".[2]

Click on following title to access full information

- Analytics Labs customers

TRANSFORMATION CATALYSTS

In this post, I will provide some high-level color on these levers, or "transformation catalysts" as I call them. In subsequent posts I will cycle back to those warranting a deeper dive.



VISION

The shift to a data science-driven enterprise will not necessarily fail without clear established vision, but the presence of a thoughtfully crafted and shared vision can dramatically accelerate

[2] Referencing the time when Yahoo! prioritized data as its core strategic play, hired one of the first "Chief Data Officers", Usama Fayyad, and built a soup-to-nuts central data organization, "Strategic Data Solutions" (2004–2008). The organization was disbanded in 2008 during Yahoo!'s executive revolving door, and the resulting "SDS diaspora" has driven data-savvy executives to Intuit, Salesforce.com, Facebook, Google, and innumerable data start-ups. Many of us are also at Pivotal.

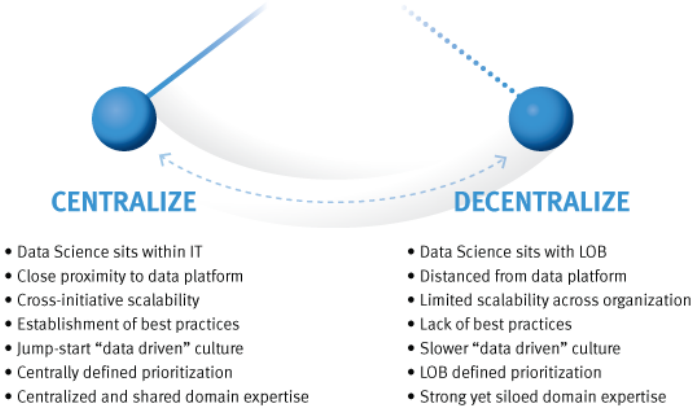
the path to value generation. By "vision" I'm specifically referencing the visualization of a path from current-state of data utilization to a clear understanding of the promise of data science (in a Big Data context,) when applied against key business problems or goals the company is facing. It implies, at least at the outset, a conviction in the financial impact (top- or bottom-line) that data science initiatives can bring, and it usually entails a passionate communication and sharing of these ideas with key influencers (decision-makers and owners of the data science value chain).

With some buy-in established, the vision should lead to strategy definition via the programmatic gathering of ideas across stakeholder communities, and aligning technology and human capital investments with the platform capabilities required to support these. Central to "getting vision" is that it must be elevated beyond any grassroots origination. Many times, a company's data-savvy thinkers who have vision (or at least a passionately held belief in the transformative nature of data science when applied to key business needs) are too buried in the organization. They aren't sufficiently empowered with a mandate to translate this vision to concrete, programmatic strategy. This, then, brings me to the next topic:

ORGANIZATION, ORGANIZATIONAL ALIGNMENT, AND EXECUTIVE SPONSORSHIP

My colleague, Josh Klahr, VP of Products, wrote a nice blog recently about building a data organization, leveraging the lessons learned during our mutual time at Yahoo! As it relates to fully harnessing the promise of data science, my observations on organization extend beyond the functional components he details, to consider fundamental organizational structure.

Within any one company, I've found there is a constant pendulum swinging between the two organizational structures: decentralized, where data scientists and other analytics professionals sit organizationally with the line of business ("LOB"), and centralized, where data scientists organizationally sit with a central data organization. A given single company can frequently swing between the decentralized model and the centralized model. This pendulum pivots along arguments advocating LOB-level domain expertise and LOB-control of project prioritization, versus considerations of the central need to pool and retain scarce talent, encourage cross-LOB knowledge sharing, and establish standards and efficiencies. I am well known to be a "centralist" and have a strong belief that any data science capabilities should sit in the same organization owning the data platform. That organization is usually but not always, IT.



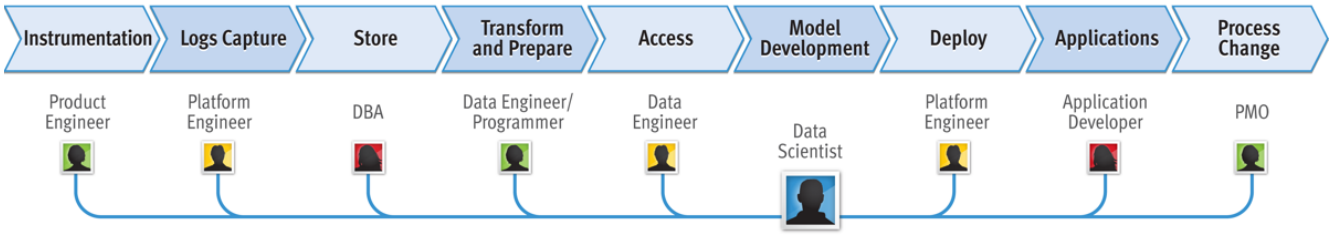
There are numerous reasons why I support this model. At a very pragmatic level, there is an essential collaboration that must exist between the data scientist and all the other owners of the "value chain" of operationalized predictive and machine-learning models.

These owners mostly, but not entirely, sit organizationally together, and the successful productionalization of models is dependent upon strong collaboration across the value-chain points. As I mentioned in my talk at Strata, I believe the data scientist's span of influence is expanding into all these critical points along the chain. To adequately yield the necessary degree of influence to align roadmaps, priorities, and resource allocation, the Data Scientist(s) must sit organizationally in the same world.

Click on following title to access full information

- Strata

DATA SCIENCE VALUE CHAIN / SPAN OF INFLUENCE



Once a company's data capabilities have evolved to the point organizationally that they have an emerging centralized data team, this typically brings with it more senior-level oversight and sponsorship, which is exactly what will be needed to push key data science initiatives from conceptualization to execution. Remember, the point here is to evolve from current-state, typically highly operationalized data processes supporting BI and other mission critical applications, to initiating proofs-of-concepts ("POCs") around predictive modeling and machine learning. These POCs will start introducing pressure on the value-chain's existing commitments and resourcing obligations, just like it does on the data itself. Driving commitments, adherence to the team's own committed timelines, and keeping the interest and enthusiasm of the executive table all the while, will require the sponsorship and oversight of a peer at said table. Without the right person at the top, willing to go to bat for the collective team and constantly socializing progress, and without the right relationships with key enablers in the wider enterprise, the entire transition itself can fall apart.

Of course there is even more color to the topic of Organization than I have touched upon here. I'll go into more depth in a later post.

DATA AVAILABILITY

The whole dialog begins with data: its assumed availability in some form to down-stream consumers, and in more granular, "raw" forms, to data scientists. Of course, realizing this core requirement is no trivial task. It carries the assumption that valuable data assets have been identified (or newly created), captured, stored in a common store or warehouse, and made available with the right adherence to company privacy and security policies.

Our team typically starts a conversation with a prospect that has:

- Siloed data assets, with valuable data stored in many locations across the prospect's IT footprint
- Immature data policies which either permit improper use of data, or instead overly restrict ease of data access
- A strong belief that there is immense value in applying more sophisticated analytics to raw data.

Getting the uncleansed, scattered assets pulled together and into one place, with usage rights adherent to proper privacy/security policies, are often the first set of tactical hurdles in enabling data science.

In some cases, despite a valuable pre-existing base of data, the actual data being targeted for inclusion in a key project doesn't yet exist. This is where a strong understanding of, appreciation for, and process-definition around data instrumentation become critical. The non-existence of data shouldn't stop efforts in their tracks. Instead, with the right process and the right assessment of value, this lack of data can kick-start an effort to instrument the target capture-vehicle (web logs, IT logs, sensor, transactional systems, etc.) for tracking. With a clear vision translated to strategy and then to specific projects, an understanding of the required data will emerge. For any one company to succeed in transforming itself into a predictive enterprise, coordination between the owners of the data capture — frequently Product Management — and the analytical leadership must be baked into the organization's planning culture.

Whether captured data is pushed to Hadoop, massively parallel processing (MPP) databases, or both, the term "availability" also alludes to an access layer that doesn't overly restrict self-provisioning. SQL, MapReduce, Pig, Hive, etc are all tools enabling access. However, the hybridization of distributed compute paradigms introduces a requirement to reduce the complexity at the access layer so any one data scientist or analyst doesn't have to master so many languages.

Pivotal solves for this issue a variety of ways, most prominently via Chorus, our open source collaboration platform for data science which enables data discovery and self-provisioning. And because we recognize data availability is one of the fundamental enablers of the predictive enterprise, we have additional technology innovations brewing on this front. Stay tuned for upcoming announcements on this.

CHOOSING THE RIGHT DISTRIBUTED COMPUTING PLATFORM

This is Pivotal's raison d'être. Since there are many people blogging about our products and their advantages/disadvantages in supporting Big Data initiatives, I won't go into deep detail on our Hadoop-MPP hybrid stack here. I will simply comment that we see the variety and volume of our customers' data drive their evaluations, and then, either a tepid or passionate embrace of Hadoop. Despite our roots in the massively parallel processing (MPP) PostgreSQL relational database, we firmly believe that Hadoop will be a cornerstone technology for the predictive enterprise.

However, for the iterative, exploratory work typical of model-building, my team continues to push more advanced analytical functions into the Pivotal MPP Database. In most cases, we find this to be the most performant solution across the very large data volumes we are working on, and given the complexity of the functions that are being applied to the data. Our Unified Analytics Platform (UAP) supports this agile utilization of both compute architectures. Along with Chorus and MADLib, UAP delivers a

complete paradigm shift for your data science practitioners that will transform their effectiveness in the organization. The integrated stack dramatically shortens time-to-insight, drives efficiencies in data science, and allows the building and maintenance of much more sophisticated models that rely on increasingly broad sources and structures of data. These benefits are what make "believers" within my data science team, and we are unabashed advocates of the Pivotal UAP.

Click on following title to access full information

- MADLib

All of this said, and beyond the Pivotal UAP pitch, a more nuanced take-away from my list is that the intelligent selection of the right distributed computing platform is not the only critical catalyst to transformation. It is clearly a key enabler given the paradigm shift it yields, but it does not ipso facto yield a predictive enterprise, given the need to focus on all the other catalysts I discuss here.

NEXT-GENERATION TOOLKITS FOR ADVANCED ANALYTICS

In the world of predictive analytics, my team distinguishes between "Big Data" toolkits and "Small Data" toolkits. The key distinction is whether the analytical functions for descriptive statistics, classification, regression, clustering, and other common machine learning techniques, are written to be applied against very large volumes of data while leveraging some form of parallelism or distribution of compute power. With "Small Data" toolkits, the application of the function or algorithm often still happens at the desktop, against heavily sampled data, which implies limited computational power and performance, and lots of data movement (typically touching at least one other person beyond the data scientist). Tools with origins in this category include: R, Stata, Matlab/Octave, SAS, SPSS, etc.

Many of the purveyors of these toolkits are actively developing "Big Data" versions, which would better leverage the computational horsepower of distributed computing platforms. Big Data analytical toolkits, then, might also include R (i.e., Revolution R, Radoop, and even R when rendered through a Procedural Language,) and SAS (High-Performance Analytics), in addition to those which were born with a focus on parallel algorithms: Mahout (open source algorithm library for Hadoop), MADLib (open source algorithm library for MPP and PosgreSQL), and GUI-based Big Data predictive modeling packages, such as Alpine Data Labs.

Click on following titles to access full information

- MADLib
- R when rendered through a Procedural Language
- Mahout
- Alpine Data Labs

These Big Data toolkits are cornerstone-enablers of the predictive enterprise, delivering on the promised paradigm shift I mentioned in the previous section. Investing in a Big Data platform that continues to rely on "Small Data" modeling toolkits is not delivering the power of Big Data innovations to those "monetization architects" (aka Data Scientists) most able to create value.

My team actively works in an agile fashion across many Big Data toolkits. We may, within a single project, work with PL/R, MADLib, and Alpine — choosing algorithms and functions depending on how they were written, their performance within Pivotal's UAP, and the quality of the output. Or, we may, if the right algorithm isn't currently available, author a new version for MADlib. I believe this sub-toolkit agility is the future of advanced analytics, where data scientists will assess the same algorithm from many toolkits and select the most suitable one depending on the data, the use-case, and the underlying platform and performance requirements. Tools like Pivotal's Chorus and open-source algorithm libraries and the efforts of our partners such as Kaggle, will enable this algorithm-level evaluation and selection. All of this will put additional pressure on the analytical walled gardens that we have in the market today.

In this scenario, where data scientists are evaluating functions and algorithms below the toolkit level, the HR implications are clear. Any organization working to become a predictive enterprise will need to cultivate or aggressively recruit the right skills to enable this sophisticated tool selection (not to mention drive the design and development of the actual model itself). This then, leads me to:

DATA SCIENCE SKILLS & TRAINING EXISTING TEAM MEMBERS

Beyond the core question of what levers control the transition to a predictive enterprise, the next most frequent questions I get from customers and prospects are "What is a data scientist?" and "How do I build a data science team?" I'll reserve the answers to these questions for a separate blog post, but I will say now that I am one of those people who think slick UIs and visualization tools will never "toolify" away the need for the data scientist's skills. The blood and guts of predictive modeling must remain firmly in the hands of well-prepared data science practitioners.

As an initial source of talent, you will want to devise a continuing education plan for your pre-existing employees who might be ready to deepen their analytical skills. Pivotal was first-to-market to respond to this need with a data science certification

curriculum. However, to rapidly build-out a team of solid data scientists (the number depending on the projects envisioned and knowledge required,) and to establish the right analytical leadership, you will likely need to recruit new skills into your organization. Your ability to lure strong data science talent in this white-hot market for data scientists depends entirely on your ability to build the most attractive analytics playground for your candidates.

Click on following titles to access full information

- Data science certification curriculum
- White-hot market for data scientists

The best way to do this is to not just offer strong market-based compensation packages, but also offer the best possible environment for the top talent to practice their craft. This includes:

- Cultivating the richest data in your sector
- Maintaining an executive spotlight on their work and a constant drum-beat of executive sponsorship for data science initiatives
- Providing state-of-the-art data platforms and analytical tools
- Creating opportunities to share what they are doing with broader audiences
- Forging strong paths for model operationalization to deliver the real monetary impact

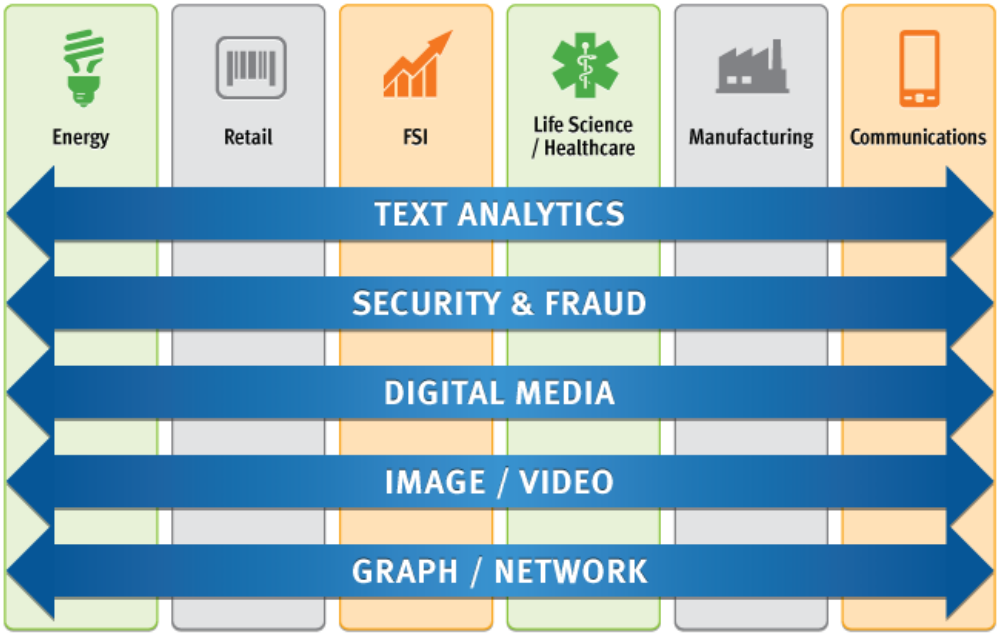
If you pull these things together, as WalMart has done, and its competitors Target and Sears are now also doing in the retail space, you will become a contender in the recruiting battles against well-known competitors such as Google and Facebook.

Lastly on this point: your data scientists will become increasingly valuable with every project. Your job, Mr. or Ms. Creator of the Predictive Enterprise, will increasingly be less about recruiting strong talent (once you've achieved the target team size that embodies the right analytical knowledge), and increasingly more about retaining your data science talent.

DEFINED PATHS TO OPERATIONALIZATION

At the end of the day, any strategic shift to predictive analytics will need to prove significant impact to business-specific performance indicators, whatever these are – e.g., revenues, costs, new customers, transaction values, customer engagement, hospital readmissions, intruders detected, etc. The pressure to deliver "wins" to the business will begin almost immediately once a concerted effort gets underway.

I have repeatedly found that all sorts of analytics resources can be thrown at a business question only to have the results bubbled up to a few key insights that languish without any meaningful action-enablement. When I see this happening, it's often a yellow flag that the company hasn't moved beyond its earlier BI-centricity, and is still defining the goal of any one effort as "informing decision-making". Informing decision-making is good, in the sense that it hopes to arm executives with the right data and insights to help them maybe make the right decisions, but it clearly doesn't scale. Anyone on the hook to prove his/her value in decision-support knows that "informing decision-making" is not measurable. On the other hand, having a clearly defined and resourced path to operationalize predictive models into production environments for down-stream application integration ensures both automated "actioning" against events of interest, and the ability to measure the impact on the target KPI.



In our engagements with customers, the "path to operationalization", which I consider the "last mile of data science," includes the following elements:

- Production, likely low-latency, data scoring against the resulting model,
- Storage and delivery of scored output to downstream applications,
- Definition of response actions and development of these within the application environment, or the development of a new application to enable automated actioning,

- Development of reporting layer to track impact, and
- (More rarely) creation of the closed-loop, channeling action-response data back into the model to power self-learning capabilities, or adaptive models.

This work can be easy or hard depending on your pre-existing investments in enterprise systems and in-house development capabilities. The key to a successful transformation to the predictive enterprise is that all of these elements and their respective requirements are actively understood, levels-of-effort are determined, and resourcing is established. Without definition of the path to operationalization and committed resourcing for execution, there will be no value delivered back to the business.

PROCESS & CHANGE MANAGEMENT

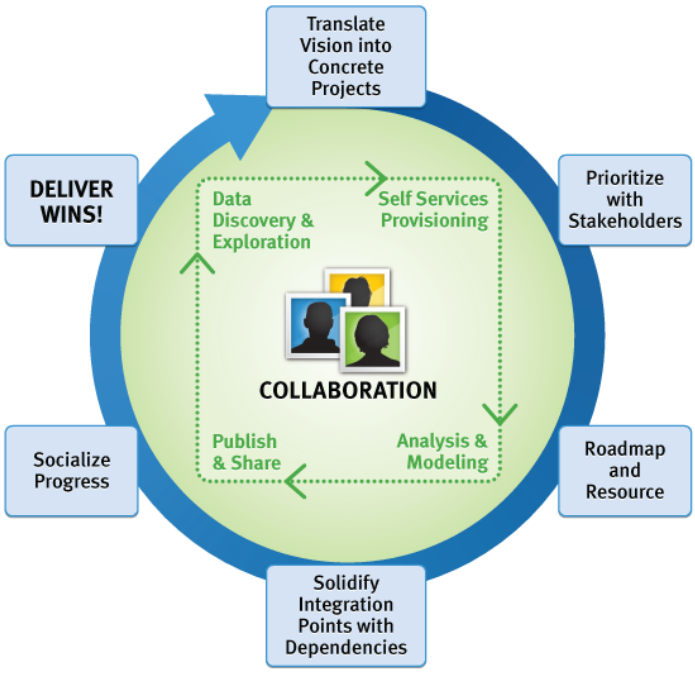
The requirements around defining the path to operationalization for any one model immediately make me think about process definition and change management. If there is one silent killer of data science initiatives, it's the lack of defined process and program management to drive project conceptualization to committed resourcing and then to execution. Because the value-chain around data science is so extended across so many disparate owners (aka "influencers of success"), the lack of defined process will make coordination and commitments, and ultimately the delivery of "wins" exceedingly difficult.

As I wrote earlier, it is common to lack the right data to support a defined data science project. In many cases the data is there for the pickin' (likely in some upstream product), but the product owners haven't instrumented the product for the proper data capture. The seemingly basic process of:

- getting in front of these product managers to drive commitments to define and then deploy the right form of instrumentation, and then
- initiating data generation, capturing, and storing of the new data for use in a predictive modeling exercise, can be extended and timely depending on the data requirements, product-engineering's existing roadmaps and resource commitments, and the agility of the data platform to absorb new data. Imagine having to repeat this basic process repeatedly across many product owners, re-defining a new process each time, and you can see how delay-prone key data initiatives can become simply due to poor process definition.

Beyond instrumentation, other critical planning junctions frequently driven by data science and involving multi-team coordination are:

- Analytics roadmapping
- SLA changes to drive lower latencies on data delivery
- ETL changes
- Data quality initiatives
- Data policy changes
- Model operationalization



Without a defined process for expediting coordination along the key influencers of the data science value chain, there are many opportunities for projects to break down.

Given the importance of this area, the creator of the Data Science Team must also think about new roles beyond the data scientist. These include positions like project managers and engagement managers to coordinate planning, gather requirements, socialize plans, drive execution against committed timelines, and communicate progress to internal stakeholders. These are all critical tasks that should not be owned by the data scientist.

(As a side note, and related to the earlier topic on Organization, Organizational Alignment, and Executive Sponsorship, this is also very frequently why many organizationally complex businesses move to new C-level roles like "Chief Data Officer" or "Chief Analytics Officer". An executive champion, with the right relationships with his/her peers, will not-so-surprisingly "grease the wheels" to drive momentum, coordination, prioritization, and resourcing commitments)

WRAPPING UP

So there you have them; my "transformation catalysts" to yield a predictive enterprise. When taken in full, you should appreciate why I call this transformation "disruptive". I'm sure many of you

will have strong opinions on what's missing or overly emphasized, and I welcome any debate or constructive input. After all, given the newness of the transformations underway at this moment, this is an ongoing learning process for us all. Over the course of the next few months, I will review a few of these catalysts in a bit more depth than what I have room for here today. Those topics warranting more discussion include Vision, Organization, and Data Science Skills.

And, lastly, you will start hearing more directly from my data science team on the innovative work we are doing across sectors, the tools we are building and using, performance benchmarks of the varying analytical toolkits, and more. Stay tuned here for more disruptive data science from Pivotal!

ABOUT ANNIKA JIMENEZ

Annika is a seasoned leader of analytics initiatives, coming to Pivotal after over six years in data leadership roles at Yahoo! At Pivotal since April 2011, she has built the "Data Science Dream Team" – an industry-leading group of Data Scientists – representing a rich combination of vertical domain and horizontal analytical expertise – who are facilitating Data Science-driven transformations for Pivotal customers. During her time at Yahoo!, she led Audience and International data solutions for Yahoo!'s central data organization, Strategic Data Solutions, and led Insights Services – comprised of a team of 40 researchers covering Web analytics, satisfaction/brand health metrics, and audience/ad measurement. Annika is a recognized evangelist for "applied data" and well known for her acute focus on action-enablement.

ABOUT PIVOTAL

Pivotal is building a new platform for a new era, setting the standard for Enterprise Platform- as-a-Service (PaaS). The company's mission is to enable customers to build a new class of applications, leveraging big and fast data, doing all of this with the power of cloud independence.

Uniting selected technology, people and programs from EMC and VMware, the following products and services are now part of Pivotal: Greenplum®, Cloud Foundry, Spring, Cetas, Pivotal Labs®, GemFire® and other products from the VMware vFabric™ Suite.

Powered by new data fabrics, Pivotal One is a complete, next generation Enterprise Platform-as-a-Service that makes it possible, for the first time, for the employees of the enterprise to rapidly create consumer-grade applications. To create powerful experiences that serve a consumer in the context of who they are, where they are, and what they are doing in the moment. To store, manage and deliver value from fast, massive data sets. To build, deploy and scale at an unprecedented pace.

Pivotal One integrates these sophisticated new data fabrics with modern programming frameworks and cloud independence in a single, united platform.

LEARN MORE

To learn more about our products, services and solutions, visit us at **pivotal.io**.

Pivotal

Pivotal, committed to open source and open standards, is a leading provider of application and data infrastructure software, agile development services, and data science consulting. Pivotal's revolutionary Enterprise PaaS product, powered by Cloud Foundry, will be available in Q4 2013.

Pivotal CF is a complete, next generation Enterprise Platform-as-a-Service that makes it possible, for the first time, for the employees of the enterprise to rapidly create modern applications. To create powerful experiences that serve their customers in the context of who they are, where they are, and what they are doing in the moment. To store, manage and deliver value from fast, massive data sets. To build, deploy and scale at an unprecedented pace.

Uniting selected technology, people and programs from EMC and VMware, the following products and services are now part of Pivotal: Greenplum®, Cloud Foundry, Spring, Cetas, Pivotal Labs®, GemFire® and other products from the VMware vFabric™ Suite.

Pivotal 3495 Deer Creek Road Palo Alto, CA 94304 **pivotal.io**

Pivotal is a registered trademark or trademark of Pivotal Software, Inc. in the United States and other countries. All other trademarks used herein are the property of their respective owners. © Copyright 2014 Pivotal Software, Inc. All rights reserved. Published in the USA. PVTL-WP-211-09/13