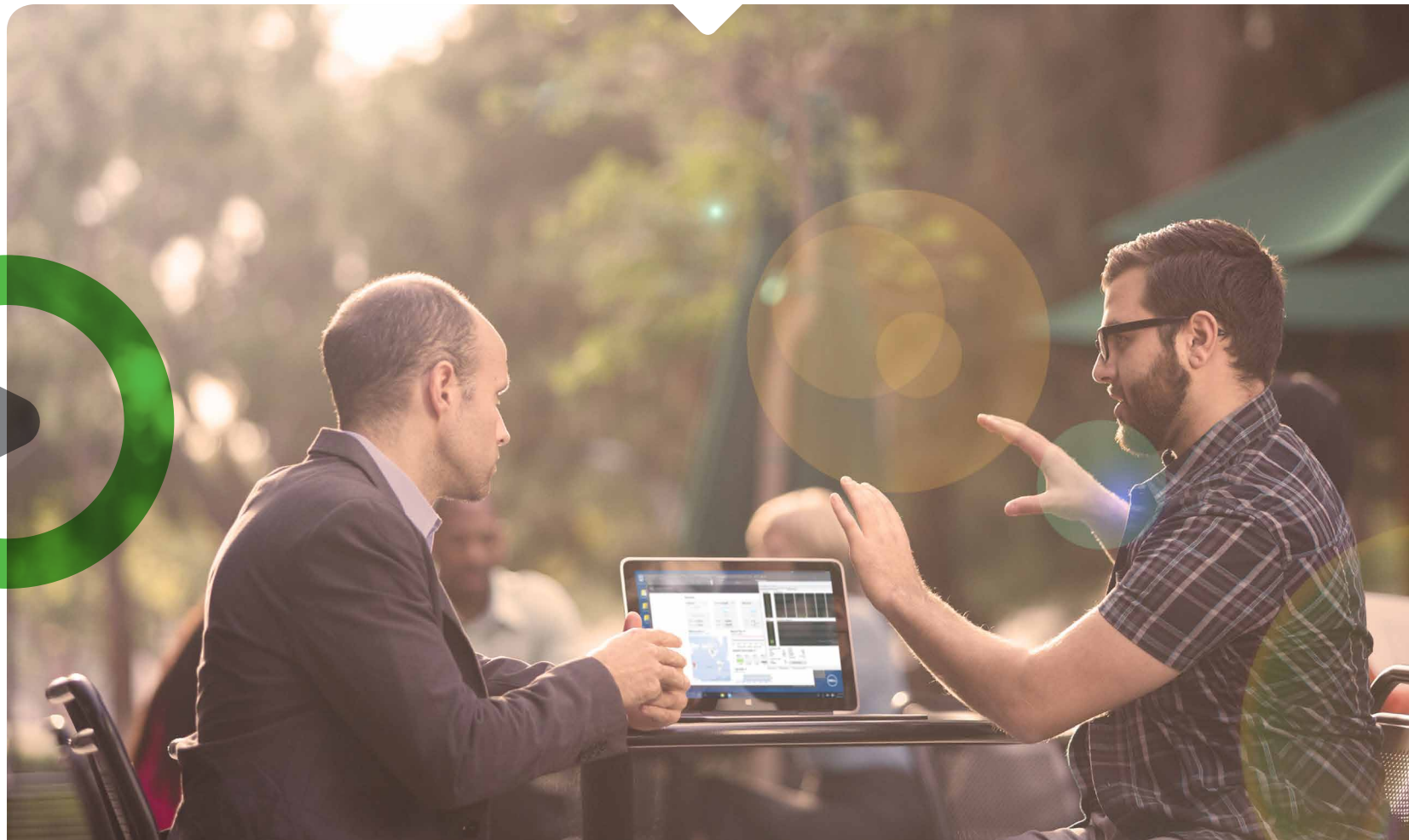



STATISTICA®

Embedded Analytics Empower the Citizen Data Scientist

A guide for analytics teams and line-of-business managers
trying to do more with embedded analytics





With the advent of technologies that connect more people, machines and processes to one another, the importance of extending advanced analytics and machine learning to your users is growing fast. But the effort to derive maximum benefit from those advanced analytics is still limited in most organizations by the human element. Data scientists, seen as the only people sufficiently trained to navigate big data successfully, become the bottleneck.

This paper examines how the citizen data scientist can use embedded analytics in your software applications, and how the line of business (LOB) can benefit. Powerful software tools like Statistica allow trained data scientists to develop models around advanced analytics, machine learning and algorithmic business, then make them available to the LOB managers and staff who use your applications to make better decisions.



PLENTY OF DATA BUT NOT ENOUGH DATA SCIENTISTS

Any specialty facing a U.S. shortage of 140,000 to 190,000 experts by 2018 has to start coming up with alternatives.¹

The McKinsey Global Institute predicts that shortfall in analytical expertise and a shortfall of another 1.5 million managers and analysts with the skills to interpret and make decisions based on big data. Worse yet, 40 percent of respondents in the McKinsey survey reported that it was difficult to even attract people with analytical skills, let alone to retain them.

In most companies, the initial impulse is to make as much progress on big data as possible with internal resources. It's a grass-roots undertaking that makes sense, but a total of 95 percent of respondents in an Accenture survey found that they had to use one or more types of external help; only five percent were able to meet their needs with internal resources alone,² most likely by lowering expectations and using inadequate tools.

Additionally, skill sets among data scientists vary and are still shaking out. An important factor in qualifying data scientists is their expertise in producing analytics for machines (making recommendations, targeting online ads, trading based on an algorithm) as opposed to analytics for humans (evaluating product quality, understanding customer churn, generating reports).³ Such distinctions about the kind of data scientist best suited to a company split the field of eligible talent even further.

¹ "Deep analytical talent: Where are they now?" *McKinsey Global Institute*, cited in "The Talent Dividend," *MIT Sloan Management Review*, April 2015, <http://sloanreview.mit.edu/projects/analytics-talent-dividend/>

² "Big Success with Big Data," *Accenture*, April 2014, https://www.accenture.com/us-en/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Industries_14/Accenture-Big-Data-POV.pdf

³ Michael Li, "The Question to Ask Before Hiring a Data Scientist," *Harvard Business Review*, August 6, 2014, <https://hbr.org/2014/08/the-question-to-ask-before-hiring-a-data-scientist>

A Gartner survey found that management is skittish about machine learning and data science projects until it sees some track record of successful pilot projects, yet those pilot projects usually require an onboard, experienced data scientist.⁴ The vicious circle hampers the entire organization's attempt to become data-driven.

Given the dearth of data science talent that could take years to go away, and competitive pressure that will never go away, smart **companies are tackling advanced analytics and machine learning by grooming citizen data scientists:** employees with math aptitude and skills that are not being actively tapped for big data projects.

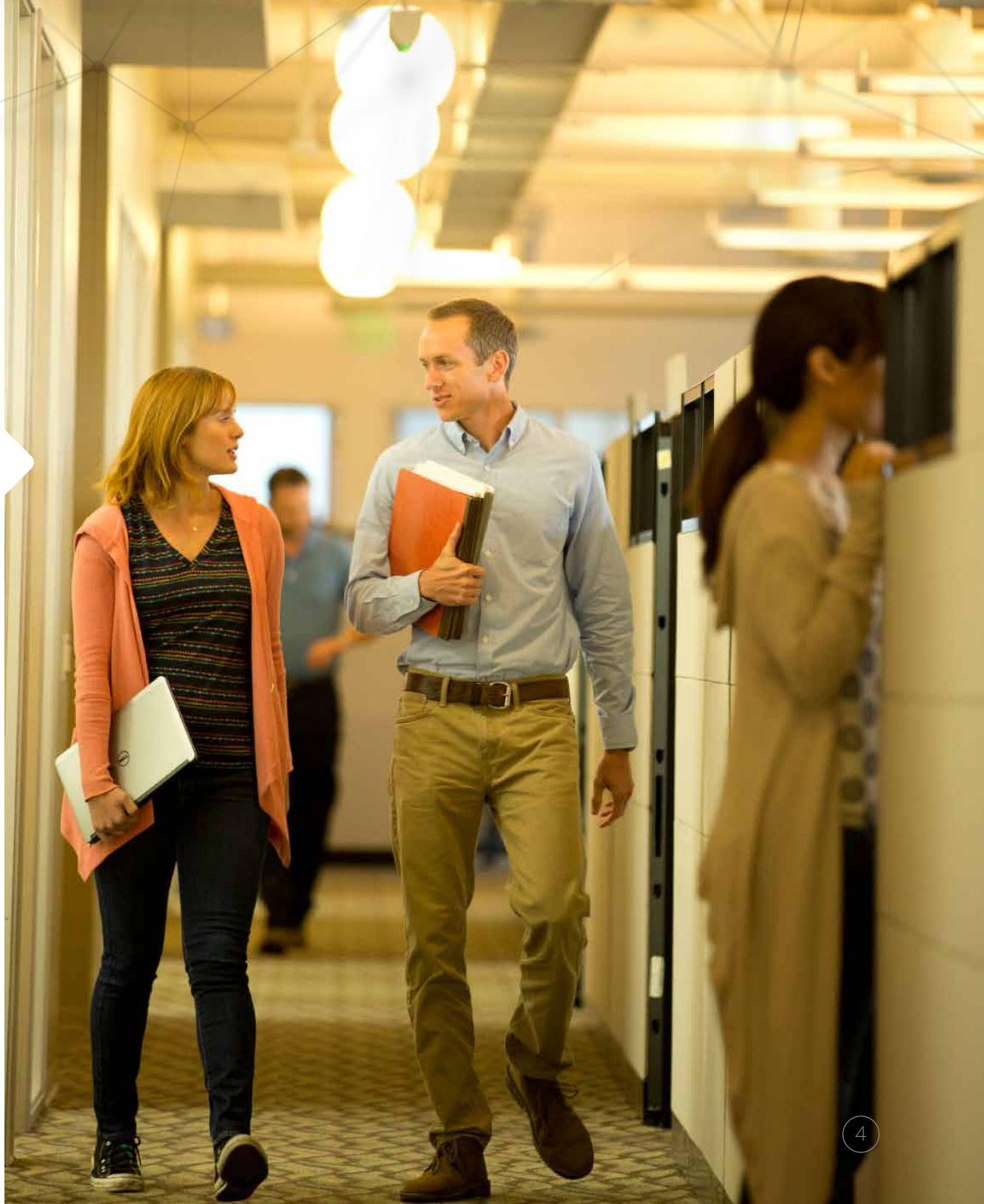
⁴ "Doing Machine Learning Without Hiring Data Scientists," *Gartner*, June 20, 2016, <https://www.gartner.com/doc/3352717/doing-machine-learning-hiring-data>



CITIZEN DATA SCIENCE: PEOPLE, PROCESS AND TECHNOLOGY

In citizen data science, users without formal training in advanced mathematics and statistics can nevertheless use your applications to extract high-value insights from data. They use tools that abstract much of the difficulty from tasks like data preparation and automate much of the work of modeling and detecting patterns in data.

The citizen data scientist is at the heart of getting more out of advanced analytic software without spending extravagantly on data scientists. Smart organizations resolve that squeeze with a combination of people, process and technology.





PEOPLE

Gartner characterizes the citizen data scientist as “a person who creates or generates models that leverage predictive or prescriptive analytics but whose primary job function is outside of the field of statistics and analytics.” That characterization is broad enough to encompass LOB staff, business analysts and employees in business intelligence (BI) and even IT – people who use your applications to work with statistics and analytics, but not as their primary function.⁵

Figure 1 depicts the relationships and characteristics of data scientists, citizen data scientists and LOB users.



Figure 1: Citizen data science empowering more people

5 Cited in “Citizen Data Scientists: 7 Ways To Harness Talent,” *Information Week*, July 24, 2015, <http://www.informationweek.com/big-data/big-data-analytics/citizen-data-scientists-7-ways-to-harness-talent/d/d-id/1321389>

The citizen data scientist plays a valuable role in what analyst Howard Dresner calls “information democracy.” As shown in Figure 2, the percentage of companies that can get by without BI and analytics applications is shrinking, while the percentage of companies with high penetration of BI products is growing. That trend sets the stage for getting information into the hands of all the stakeholders and constituents, instead of just the data scientists and quants. The more eyes on the data, the better.⁶

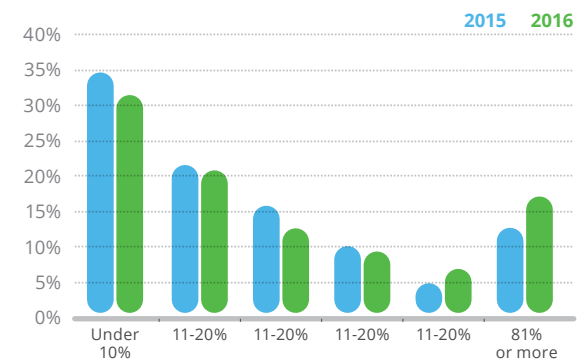


Figure 2: BI penetration, 2015 to 2016

6 “Wisdom of Crowds Business Intelligence Market Study,” *Dresner Advisory Services*, May 2016, cited in webcast: <https://software.dell.com/webcast-ondemand/whats-in-the-dresner-business-intelligence-market-study-analysis-vea8113432/>



The People Factor and the citizen data scientist

Robert Lake,
Senior Data Scientist,
Cisco⁷



"We run a team of 11 data scientists, plus myself, working directly for the COO. That's four senior data scientists doing the main work, backed up by seven others learning as they go along. We can take undergrads and develop them to the level of a data scientist. The advantage of working with such a mixed group is that it helps me find people easier.

"We do these projects rapidly. When I took over this team, we were solving things in about nine months minimum; now the challenge I put to the team is to solve them in one day. We've been able to do that with Statistica and other tools, but the way we train or work together helps us get to a solution rapidly for a customer, instead of working for months and popping out an answer at the end.

"We cover everything you can think of, which is another reason we develop our own people: we need them focused on our business. A straightforward example is customer churn from services around particular products. We can find patterns to certain products and their maturity life cycle. We've been able to show some of our product groups that for a certain customer profile they'll have, say, two years to sell services to them, and then it's important to get them onto another product and be ready to move to sell it.

"The way I lead my team is to ask, 'What decision is the customer trying to make?' By going backwards from that decision, and understanding the information and insights our customers need, our interaction starts with the business leaders. They come up with the first insight, which is strange because we're supposed to be the ones helping them. But as we come back and start to discover data related to the decision, that's where we find additional insights. 'Here is the insight you're looking for,' we tell them, 'but we've also found this.'

"If you start with the decision and work backwards, you find others that are related, and that means you provide continuous value."

⁷ Excerpts from "Three big (data) things you need to know," #ThinkChat video, November 2015, <https://software.dell.com/video/three-big-data-things-you-need-to-know8107113/>



PROCESS

The mechanical process by which data scientists and citizen data scientists make better use of data and analytics is underpinned by a deeper question about the organization as a whole: Does it have processes for sharing anything?

This is not always a given in companies that have grown quickly, have grown through mergers and acquisitions or have begun to shrink. If the culture has never embraced or fostered the notion of transparency and sharing, then whatever process the company may put in place to use software to publish analytical models and the data they harvest is unlikely to succeed.

Figure 3 illustrates survey responses on the topic of sharing information and insights:⁸

⁸ Dresner Advisory Services, May 2016.

- Closed-loop processes indicate that the organization has already thought about and implemented a process for sharing insights.
- Where ad hoc sharing prevails, analysts make information and insights available as they present themselves. Despite the lack of formal process, impromptu sharing takes place.
- The parochial attitude dictates one-off sharing among a select group of people.
- Plenty of companies don't share BI or insights at all. That can be a function of either the culture or the immaturity of the data science initiative in the company.

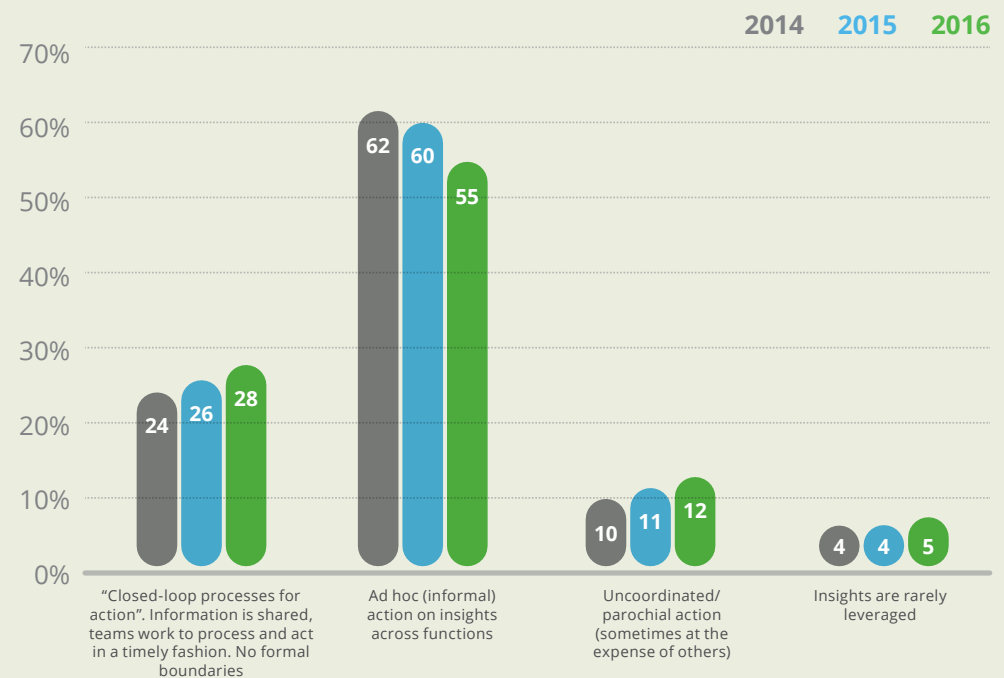


Figure 3: Processes for transparency and sharing, 2014 to 2016

Once the citizen data scientists have stepped forward and the data scientists have qualified them, the process of carving up work begins.

The goal of engaging citizen data scientists is not to replace data scientists but to complement them with a set of power users who can use your applications to pick up where scientists leave off and fill in any skill gaps. Given that the optimal use of big data requires knowledge of coding, statistics, machine learning, database management, visualization techniques and industry-specific knowledge, the best way to pull it off is by combining multiple skill sets. At the very least, citizen data scientists offer the greatest value in the area of LOB knowledge, something it would be inefficient for a data scientist to stop and learn to any useful degree.

Once a process is in place, the traditional barriers that data scientists face to buy-in – both upstream to management and downstream to staff – begin to fall as information democracy puts more data into more hands. Beyond arriving at insights that boost revenue or lower costs in the short run, the promise of data science lies in applying those insights in ways that beneficially shape the company's direction in the long run. The smoothest way there is by linking the efforts of trained data scientists and citizen data scientists.⁹

In practice, it makes sense for data scientists to stick to the work of advanced analytics and statistics for which they are trained, creating workflows for data preparation and modeling. When those workflows are ready to test or take into production, the data scientists use your analytics software to push them to the citizen data scientists, who run them and ensure they work as designed. In time, the citizen data scientists can assume greater responsibility, using your application to modify workflows and create their own.

Of course, that process requires the technology component of an advanced analytics platform.

⁹ Luc Burgelman, "The Rise of the Citizen Data Scientist," NGData Blog, June 17, 2016, <http://www.ngdata.com/the-rise-of-the-citizen-data-scientist/>





How Process Relates to the Citizen Data Scientist

Tim Alosi,

Director of Data Analytics,
Sanofi¹⁰



SANOFI

¹⁰ Excerpts from "Automating a single source of truth in manufacturing," #ThinkChat video, January 2016, <https://www.youtube.com/watch?v=1uQzRhoHjY>

"Sanofi has 107 manufacturing facilities with a broad range of systems and manufacturing processes. Our users spend up to 40 percent of their time just managing, preparing and extracting data. Then, because we're a pharmaceutical company, they also spend a lot of time managing compliance. It's relatively simple reporting, but it's required by regulatory bodies to ensure the quality and efficacy of our products.

"Our initial focus is to ensure we can bring in the data and provide a single source of truth to the user, make sure they can trust it and make sure it's in the right context for them to use it. Next, we want to automate the simple analytics, the stuff they have to do day after day. We want to take away all the non-value-add tasks and free them up to create products using advanced analytics tools.

"We're trying to ensure that we're not leading with a solution looking for a problem. We spend a lot of time in data preparation, but also a lot of time enabling the consumers of the data. When we have specific issues like a yield being impacted or some repeating problem, our users ask us to educate them on technologies like machine learning and cloud-based analytics.

"Last year we connected three of our small biologics plants into our central system. From those three plants we have 60,000 unique tags, or sensors, sending data into a central historian as rapidly as once per second. We think we'll grow from 60,000 to half a million tags over the next few years."



TECHNOLOGY

Most analysts reach reflexively for a spreadsheet program to crunch numbers and arrive at insights. The intuitive, trusted, row-and-column format makes immediate sense and is infinitely flexible. However, **spreadsheet software eventually runs out of gas, either in collaborating, sharing, combining disparate data sets, performing advanced analytics or executing repeatable workflows.**

Data scientists know that it is futile to impose raw math and statistics on people who are not adept at them. The goal is to get an analytics platform into the hands of people who can build the models for use all around the organization. Every analytics platform claims ease of use, but that is not enough. It must be sufficiently powerful to meet the needs of data scientists yet easy enough for LOB staff to use.

To support the process described above, the Statistica platform meets several important criteria.



AUTOMATED DATA PREPARATION

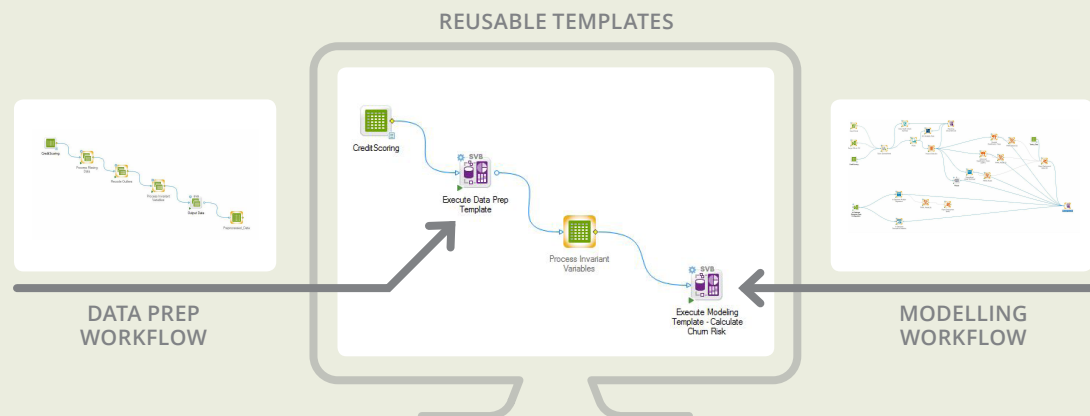
Data preparation can be cumbersome and daunting to the uninitiated, so strong analytics software automatically cleans and prepares data for analysis by checking its health first. As shown in Figure 4, wizards and guided workflows in Statistica play a big role in automating data acquisition and preparation as citizen data scientists bring together a wide range of data sets, evaluate factors like missing values, invariance and outliers, then accept or reject recommendations from the analytics platform.



REUSABLE WORKFLOWS

Meanwhile, data scientists still operate at the steep end of the math-and-statistics spectrum, so they look for the flexibility to code in their languages of choice – typically, R and Python – rather than in a proprietary programming language.

In the ideal division of labor depicted in Figure 5, data scientists create advanced analytic workflows for data preparation and modeling, then make those workflows available through Statistica to the rest of the organization as reusable templates. That abstracts the complexity of advanced analytics while making useful results accessible within your application to LOB users, who can pick up where trained scientists have left off and apply the analytics to business problems.





ANALYTICS SHARING ANYWHERE

Data scientists create analytical models, train them, score them, prepare them to go into production and export them as code. The business then wants to transport them from the analytics platform to a data center, a transactional environment, an IoT device or the cloud – to any number of places, anywhere the business needs them.

When analytics are running in diverse sites and regions, it's an important option to be able to continually revisit and modify the code in the models and templates.

Statistica's Native Distributed Analytics Architecture (NDAA) is designed for sending models to far-flung internal sites, customers and partners.

NDAA provides complete, mature model management for collaboration and sharing of analytics anywhere. By exporting their models as Java, PMML, C, G++ and SQL, data scientists can send their math to the data at all sites instead of hauling the data back to the math.

Figure 6 depicts an example of sites in Oklahoma, Brazil, Taiwan and California. With Statistica running in each location, the model management of NDAA distributes and shares analytical models across all these environments.





COLLECTIVE INTELLIGENCE

Analytical models consist of software code, in the same way that mobile apps do. So why not buy and sell analytical models and algorithms the same way as mobile apps?

Algorithm marketplaces have sprung up to fulfill this need, and they are capturing attention.¹¹ They are also providing a marketplace for external expertise: even an in-house cluster of data scientists will occasionally encounter analytical questions it cannot answer on its own.

Consider three representative sources of models, insight and perspective for data scientists:

- **Apervita** is an analytics community designed for capturing and sharing knowledge among health professionals and enterprises.
- **Algorithmia** has published an API that exposes the collective knowledge of algorithm developers around the globe.
- **Expert Models** hosts models and data processing programs online, and makes it easy to wrap these programs in an appropriate user interface.

¹¹ Alexander Linden, "Algorithm Marketplaces Are Bringing the App Economy to Analytics," *Gartner Inc.*, October 2015, <https://www.gartner.com/doc/3150917/algorithm-marketplaces-bringing-app-economy>

The openness of these data markets and communities is what makes them an ideal conduit for collective intelligence. Figure 7 shows an example combining that wealth of knowledge with the Statistica platform for creating and executing models.

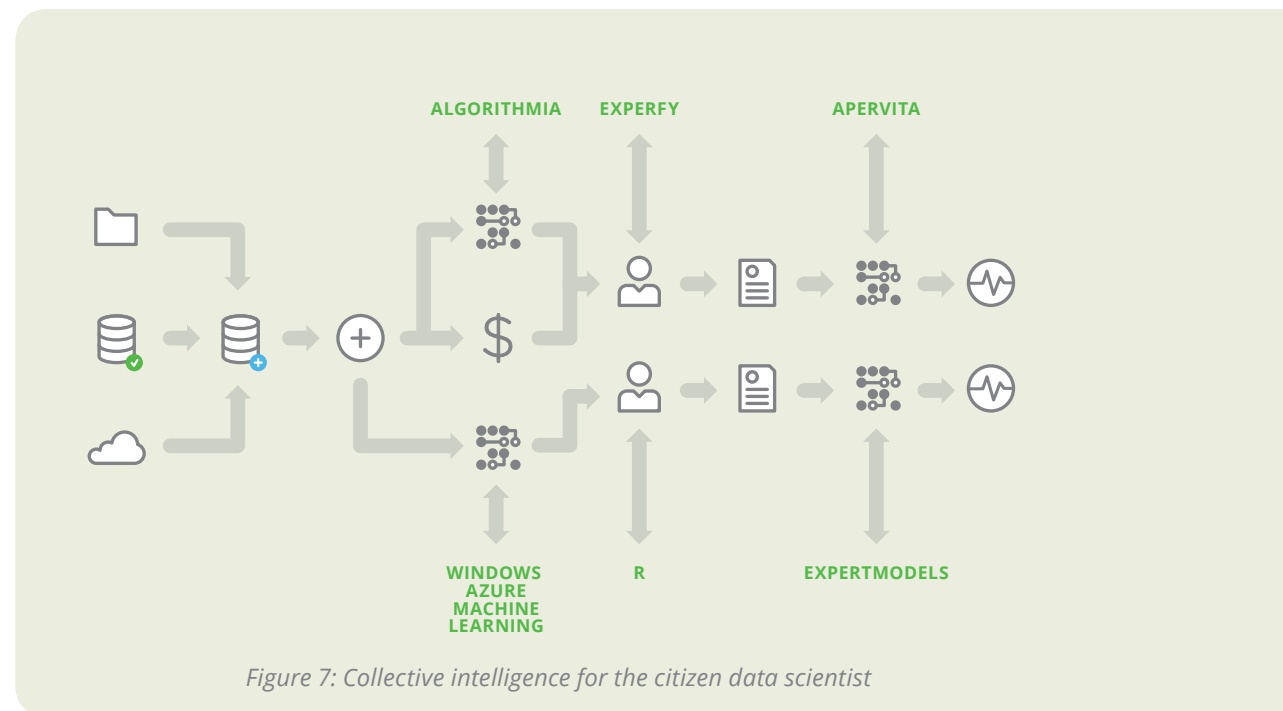


Figure 7: Collective intelligence for the citizen data scientist

Collective intelligence is one way to narrow the skills gap between data scientists and citizen data scientists. Few organizations will be able to hire all the internal talent they want, so most will want to take advantage of talent they don't have in house. Conversely, data scientists will want to make their analytical models available in the market and turn them into a source of revenue. Workflow in Statistica accommodates models from different vendors in different locations and in different cloud environments.



The Technology Perspective on the citizen data scientist

Jim Hare,
Research Director at Gartner Inc.¹²

Gartner®

Recent research by Gartner underscores the growing role of algorithms and analytic models in decision making. Calling this “algorithmic business,” Gartner believes that half of large organizations will apply proprietary algorithms for competitive advantage by 2018.

In fact, smart companies are moving from mere measurement of activity to advanced analysis. By developing algorithms that take into account their own business practices, companies have the chance to derive more insight from data and become less dependent on intuition.

But with the wider application of advanced analytics comes a new burden: The organizations that use them must also prove that their algorithms are trustworthy. Data, analytics, trust and business relationships are interdependent, so any company that sets store by and predicates decisions on its analytics must also govern their impact.

In Gartner’s opinion, the bright future for advanced analytics and algorithms lies in the ability to share them. Algorithm marketplaces are a first iteration that, once joined to Platform-as-a-Service environments, will build a bridge between consumers who want data at the atomic level, and providers who want to control licenses and own the integration. For now, placing controls on the algorithmic data processing seems the best way to accommodate all needs.

¹² “Gartner Says More Than Half of Large Organizations Will Compete Using Advanced Analytics and Proprietary Algorithms by 2018,” *Gartner Inc.*, January 28, 2016, <http://www.gartner.com/newsroom/id/3192717>



PROTECTION FOR INTELLECTUAL PROPERTY (IP)

A completed workspace represents long hours of work and many steps for data acquisition, preparation, modeling and deployment. The result of all of that effort could be considered IP. You or your customer may want to allow use of a workspace while still protecting the IP behind it.

Starting with version 13.2, Statistica includes a feature for publishing protected workspaces. When you publish a workspace, Statistica saves it in a format that will run on systems where you want to protect your IP, but the format prevents users from seeing or modifying the contents. **The system can access the published workspace and provide inputs and outputs as defined by the workspace creator**, but otherwise the workspace remains a black box.

Thus, you and your customers can create data preparation or analytic workflows and obscure the details to protect your IP.



CONCLUSION

As more of your customers address the constant stream of data arising from daily operations, they realize that they are constrained by their data science resources. They can either compete in the limited pool of expensive data scientists or supplement the one(s) they have by building out their own staff of citizen data scientists.

As the user experience of advanced analytics improves, citizen data scientists can use your application to perform high-end analytics even though they lack the formal statistical training of data scientists. More information democracy, greater impact on the bottom line and better-organized use of big data are a few of the principal advantages.

The citizen data scientist is the organization's best chance to groom scarce modeling and analytical skills that will allow them to meet urgent business demands and turn data into action. The advanced analytics platform provided in Statistica arms your application or service with the high-end tools your power users need for data preparation, automated workflows, distributed analytics and collective intelligence while equipping citizen data scientists with mid-level tools for their growing array of big data responsibilities.





ABOUT STATISTICA

Statistica's big data, advanced analytics and IoT offerings provide you endless possibilities to innovate your enterprise. Whether it's uncovering the genetic basis of a disease, reducing hospital readmissions, mitigating financial risk, or ensuring procedural validation, Statistica enables organizations to transform in new and exciting ways. By embedding analytics everywhere and empowering a wider community of citizen data scientists, you'll accelerate innovation, improve patient experiences, and streamline your enterprise for the future.

If you have any questions regarding your potential use of this material, contact:

Statistica

2300 East 14th Street
Tulsa, Oklahoma, 74104

statistica.io

Refer to our Web site for regional
and international office information.

© 2016 Quest Software, Inc. ALL RIGHTS RESERVED. This document contains proprietary information protected by copyright. No part of this document may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording for any purpose without the written permission of Quest Software, Inc. ("Quest Software").

Statistica and the Quest Software logos and products — as identified in this document — are registered trademarks of Quest Software, Inc. in the U.S.A. and/or other countries. All other trademarks and registered trademarks are property of their respective owners.

The information in this document is provided in connection with Quest Software products. No license, express or implied, by estoppel or otherwise, to any intellectual property right is granted by this document or in connection with the sale of Quest Software products. EXCEPT AS SET FORTH IN Quest Software'S TERMS AND CONDITIONS AS SPECIFIED IN THE LICENSE AGREEMENT FOR THIS PRODUCT, Quest Software ASSUMES NO LIABILITY WHATSOEVER AND DISCLAIMS ANY EXPRESS, IMPLIED OR STATUTORY WARRANTY RELATING TO ITS PRODUCTS INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NON-INFRINGEMENT. IN NO EVENT SHALL Quest Software BE LIABLE FOR ANY DIRECT, INDIRECT, CONSEQUENTIAL, PUNITIVE, SPECIAL OR INCIDENTAL DAMAGES (INCLUDING, WITHOUT LIMITATION, DAMAGES FOR LOSS OF PROFITS, BUSINESS INTERRUPTION OR LOSS OF INFORMATION) ARISING OUT OF THE USE OR INABILITY TO USE THIS DOCUMENT, EVEN IF Quest Software HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Quest Software makes no representations or warranties with respect to the accuracy or completeness of the contents of this document and reserves the right to make changes to specifications and product descriptions at any time without notice. Quest Software does not make any commitment to update the information contained in this document.