ŷhat

Applied Data Science

Practical guide to building data-driven products beyond analysts' laptops

Presented by

Yhat

http://yhathq.com/

March 2014

What are the byproducts of data science?

What are predictive models?

How can data science improve products?

How do you go from insight to prototype to production application?

This is a white paper about data science teams and how companies apply their insights to the real world. You'll learn how successful data science teams are composed and operate and which tools and technologies they are using.

We discuss the byproducts of data science and their implications beyond analysts' laptops and answer the question of what to do with predictive models once they're built. Lastly, we inspect the post-model-building process to highlight the most common pitfalls we see companies make when applying data science work to live data problems in day-to-day business functions and applications.

Describing data science

In the wake of an increasingly digital economy, businesses are racing to build operational knowledge around the vast sums of data they produce each day. And with data now at the center of almost every business function, developing practices for working with data is critical regardless of your company's size or industry.

"Data science," one of many recently popularized terms floating amidst the myriad of buzzwords and big data hoopla, is a field concerned with the extraction of knowledge from data. Practitioners—aptly named "data scientists"—are those charged with solving complex and sophisticated problems related to data usually employing a highly diversified blend of scientific and technical tools as well as deep business and domain expertise.

The central goal of data science

As is the case with any analytical project, the central goal in data science is to produce practical and actionable insights to improve the business. That is to say, data scientists overcome complexities involved in data to empower businesses to make better operational decisions, optimize processes, and improve products and services used by customers and non-technical employees.

> "What distinguishes data science itself from the tools and techniques is the central goal of deploying effective decision-making models to a production environment. "
>
> – John Mount & Nina Zumel, Practical Data Science with R

Profile of a typical data science project

Project scope and definition

With broad strokes, a data science project begins with some question, need, or goal in mind and with varying degrees of focus. Accordingly, a data scientist's primary task at the start of a new project is to refine the goal and develop concrete project objectives.

Analysts will first conduct a preliminary survey of the data, applying domain knowledge to develop a clear and succinct problem definition to serve as the principal object of study.

EXAMPLE:
Detecting erroneous car listings at Vast.com, the premiere marketplace for data powering vertical search in automotive, travel and real estate.

BACKGROUND:
Vast ingests car-listing data from thousands of suppliers and publishes listings to thousands of marketplaces that trust the data are accurate. The listing data itself is initially created manually by users and is therefore vulnerable to human error.

INITIAL BUSINESS NEED:
Identify inaccurate car listings on the fly and fix them before they reach users.

FOCUSED PROBLEM DEFINITION:
We treat this as a binary classification problem where, given a car listing created by a user, we predict whether the asking price is sensible given the vehicle's characteristics (e.g. make, model, year, odometer reading, etc.).

Identify relevant data sets

With a narrow and expressive definition of the problem, data scientists can begin to evaluate different data sets to identify which variables are likely to be relevant to the problem they are trying to solve. Evaluating which data sets should be used for the project, however, is not an activity performed in isolation. Most companies have numerous data sets, each highly diverse in shape, composition and size. Analysts may or may not be familiar with a particular data source, how to query it, where it comes from, what it describes or even that it exists.

For these reasons, quantitative analysts are usually working in proximity to or in direct collaboration with engineers, marketers, operations teams, product managers, and other stakeholders to gain a robust and intimate understanding of the data sources at their disposal.

Cross-functional collaboration

Collaboration at this stage is not only valuable for identifying which data are relevant to a problem but also for ensuring the ultimate viability of any resulting solution. Hybrid teams composed of stakeholders in

separate functions produce deeper collective understanding of both the problem and the data at the center of any project. Knowing how a data set is created and stored, how often it changes, and its reliability are critical details that can make or break the feasibility of a data product.

For example, consider a new credit-scoring algorithm more accurate than previous methods but that relies on data no longer sold by the credit bureau. Such circumstances are common today given that data sets are so diverse and subject to frequent change. By incorporating interdepartmental expertise in the early stages of model development, companies dramatically reduce the risk of pursuing unanswerable questions and ensure data scientists are focusing attention on the most suitable data sets.

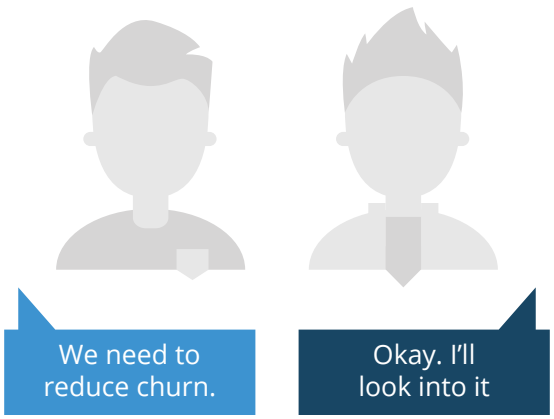
Model-building

After firming up the project's definition and completing a preliminary survey of the data, analysts enter the model-building phase of analytics lifecycle. The notion of "model" is often obscure and can be difficult to define, even for those well versed in data science vocabulary.

A statistical model, in short, is an abstract representation of some relationship between variables in data. In other words, a model describes how one or more random, or independent, variables relate to one or more other dependent variables. A simple linear regression model might, for example, describe the relationship between years of education (X) and personal income (y).

But linear regression is far from the only way to represent the relationships in data, and identifying the right algorithms and machine learning methods for your problem is largely an exploratory exercise. Data scientists apply knowledge of the business and advanced research skills to identify those algorithms and methods most likely to be effective for solving a problem. Many and perhaps most data science studies are bound up with solving some combination of clustering, regression, classification, and/or ranking problems. And within each of these categories are numerous algorithms that may or may not be suitable for tackling a given

Lots of conversations like this:



problem.

To that end, the model-building phase is characterized by rigorous testing of different algorithms and methods drawing from one or more of these problem classes (i.e. clustering, regression, classification, and ranking) with the ultimate goal being to identify the "best" way to model some underlying business phenomenon. "Best," importantly, will take on a different meaning depending on the problem, the data, and the situational nuances tied to the project. For example, the "best" way to model the quality of the Netflix recommendation system is very different from the "best" way to model the quality of a credit-scoring algorithm.

Actionable data science & applications in operations

When a data science project progresses beyond the model-building phase, the core question is how best to take advantage of the insights produced. This is a critical junction and one ultimately determines the practical ROI your data science investment.

Extracting value from data is like any other value chain. Companies expend resources to convert raw material—in this case data—into valuable products and services suitable for the market.

As is the case with any value chain, a product gains value as it progresses from one lifecycle stage to the next. Therefore, the manner in which activities in the chain are carried out is important as it often impacts the system's value.

Consider the product recommendations example again—our goal is to increase average order size for shoppers on our website by recommending other products users will find relevant.

Data science lifecycle steps:

1. Refine the problem definition
2. Survey the raw material and evaluate which data to include in the model
3. Rigorously test modeling techniques
4. Identify a winning modeling strategy for implementation
5. Integrate recommendations into the website to influence customers

Common sense indicates that progressing through step four without achieving step five falls short of the objective. But, sadly, this is a common scenario among companies developing data science capabilities. Similarly, it is often the case that hypotheses are disproved only after companies have invested

> A data product provides actionable information without exposing decision makers to the underlying data or analytics.
>
> Examples include: Movie Recommendations, Weather Forecasts, Stock Market Predictions, Production Process Improvements, Health Diagnosis, Flu Trend Predictions, Targeted Advertising.
>
> – Mark Herman, et al., Field Guide to Data Science

substantial time and effort engineering large-scale analytics implementations for models which later prove to be suboptimal or entirely invalid.

Why building data driven products is hard

Losing sight of non-technical objectives

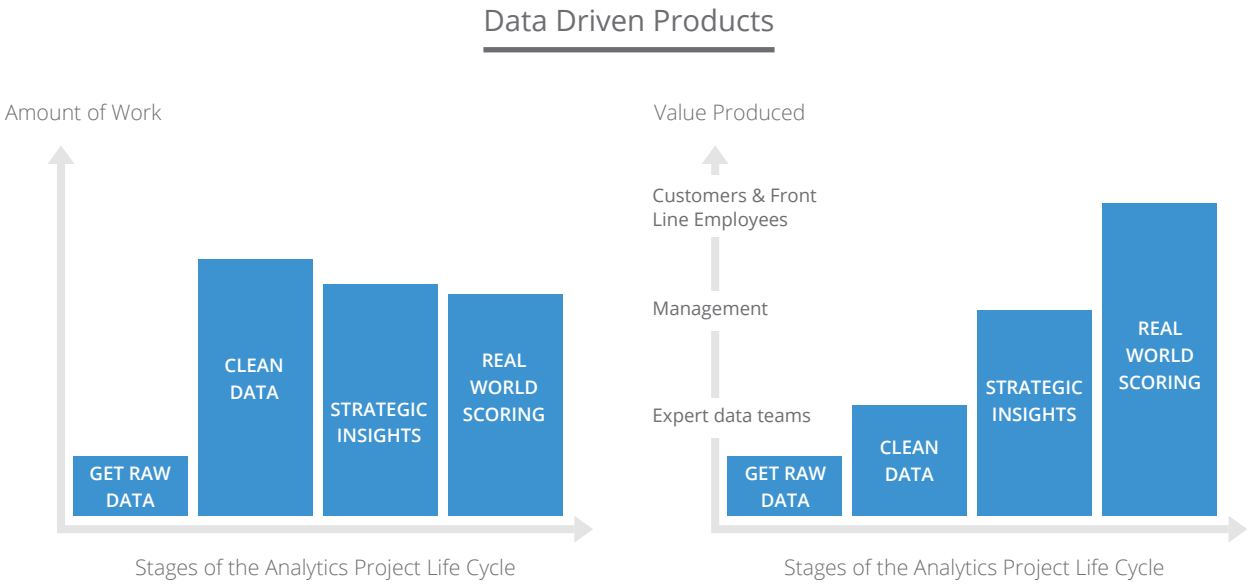
One issue is that the effort expended during each lifecycle stage is not necessarily equal to the value created. Cleaning data requires considerable effort from data engineers, data warehouse experts, and data scientists, but produces no significant utility for the customer directly.

To be clear, cleaning data is valuable and often necessary. It is an activity that makes modeling and application building easier and in some cases possible. For the most part, however, the early activities of a data science project rarely produce noticeable differences in the quality of your product at least as far as the customer is concerned.

For this reason, successful data science teams are often intensely focused on concrete, non-technical business objectives. By keeping an eye on success criteria expressly dealing with an end-user or front-line employee, teams are less likely to declare a project completed prematurely.

Using the wrong tool for the job

Another obstacle in data-driven product building is rooted in tools. Data scientists use incredibly different tools than those used by application developers responsible



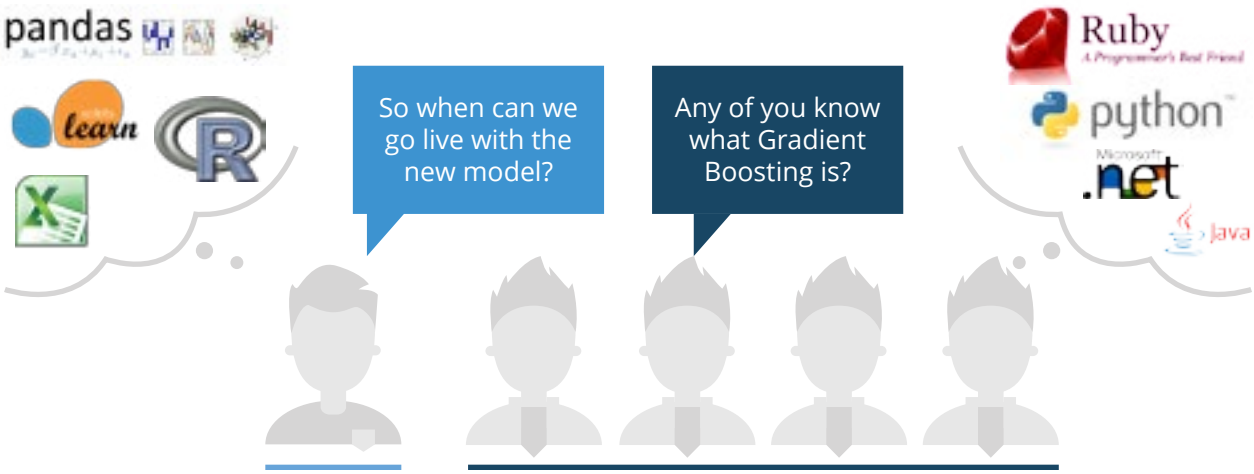
for building customer-facing apps. Although both groups use code to solve problems, the byproducts of data science are not necessarily compatible with other systems or vice versa. This is an important distinction and one that is often underestimated or overlooked by managers albeit inadvertently.

Scientific programming languages like R and Python are ideal for cleaning, exploring, and modeling data which explains the overwhelming popularity of these tools among data scientists. Likewise, application developers gravitate towards tools designed specifically for building scalable web and mobile apps and creating compelling user experiences—think .NET, Ruby on Rails, Node.js, or JVM based web and mobile frameworks.

Analysts use tools like R and Python to perform advanced statistical procedures that may take many tens or hundreds of lines of code in another language. For example, R's `glm` function, a command that fits a generalized linear model capable of predicting a True/False outcome, has no native equivalent in Java or Ruby. As a consequence, data science prototypes built using this type of built-in scientific functionality must be ported to another environment via a notoriously difficult, time consuming, and error-prone process in order to be useful in production.

This operating paradigm is especially alarming considering the inherent exploratory nature of data science in the first place. Conceiving hypotheses, designing tests and measuring results are core to the data science job function and primary reasons these professionals are in such demand. Operational practices that inhibit data scientists from deploying and running experiments are, by logical extension, counterproductive and should be avoided.

Now what?



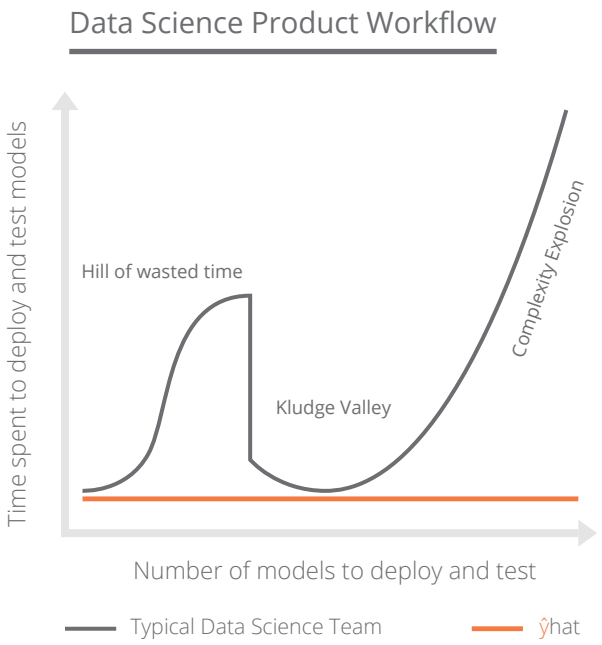
A more efficient means of shipping data products

The value companies get from data science is governed largely by the pace at which teams can conceive and test new ideas and bring data-driven products to market. At Yhat, we are re-envisioning how data scientists apply their work to real world problems to help companies move from proof of concept to data-driven product faster and more efficiently than ever before.

Through our cloud-based and on-premise applications, analysts deploy and update predictive routines using tools they know and love without impacting application developers or production systems. The platform is designed not only to play-nice with the leading tools and workflows but to naturally extend them to provide a highly streamlined process for moving from insight to prototype to full-scale production application efficiently.

The platform includes native packages for R and Python, a simple and elegant command line interface for deploying, testing, and managing models in production, and an administration and reporting web app for monitoring and tracking model performance and usage. Models that would otherwise need to be ported manually from analysts' laptops into production are instead made instantly accessible via several universally understood and standards-compliant protocols. Invoke models in real-time via REST and low latency streaming APIs or score data offline via on-demand or scheduled batch jobs.

By reducing or, in some cases eliminating, the handoff from data scientist to engineer, we help analysts package and ship their insights as ready-to-use products and avoid long implementation periods. Through Yhat, you can quickly deploy analytical prototypes, update and improve your recommendation systems, make credit policies changes, try different learning methods, or incorporate new data sets into workflows at an unheard of pace and without impacting your development roadmap.



Works Cited

Chandrasekaran, Swami. "Curriculum via Metro Map." Pragmatic Perspectives. Swami Chandrasekaran, 8 July 2013. Web. 12 Mar. 2014. <http://nirvacana.com/thoughts/wp-content/uploads/2013/07/RoadToDataScientist1.png>. A pragmatic visual inspired by transportation metro maps to depict the learning path to becoming a data scientist. Each area / domain is represented as a "metro line", with the stations depicting the topics you must learn / master / understand in a progressive fashion.

Harris, Jeanne G., et al., The Team Solution to the Data Scientist Shortage. Chicago: Accenture Institute for High Performance, n.d. Web. 12 Mar. 2014. <http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Team-Solution-Data-Scientist-Shortage.pdf> Today's data scientist shortage won't go away soon, yet companies need them more than ever to leverage the value of big data. The solution is to build teams of data scientists instead of seeking soloists.

Herman, Mark, et al., comps. The Field Guide to Data Science. Strategy and Technology Consulting Firm | Booz Allen Hamilton. Booz Allen Hamilton, n.d. Web. 12 Mar. 2014. <http://www.boozallen.com/media/file/The-Field-Guide-to-Data-Science.pdf>. Our goal in creating the Field Guide to Data Science is to capture what we have learned and to share it broadly. We want this effort to help drive forward the science and art of Data Science.

Mount, John, and Nina Zumel. Practical Data Science With R. 7th ed. Greenwich: Manning Pubns, n.d. Web. 12 Mar. 2014. <http://www.manning.com/zumel/PracticalDataScienceR_MEAP_ch01.pdf>. This book combines technical content with practical, down to earth, advice on how to practice the craft of data science.

Porter, Michael E. Competitive Strategy. New York: Free Press, 1998. Print.

Witten, Ian H., and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. N.p.: n.p., 2005. The Morgan Kaufmann Series in Data Management Systems 2. Manning Publications Company. Web. 12 Mar. 2014. <http://home.ustc.edu.cn/~hshl05/addition/Practical Machine Learning Tools and Techniques.pdf>.

About

Yhat, Inc. provides the leading data science operations platform for turning analytics projects into products. We improve data science operations by connecting people, processes, and systems into easily managed documentation, applications, and reports.

Through Yhat, data scientists train, deploy and update models; change features and algorithms in production without downtime; configure recurring jobs to score data and more. The platform tracks how your predictive models perform over time to provide non-technical stakeholders new levels of transparency into data science efficacy and ROI.

Contact Us

http://yhathq.com
info@yhathq.com
(917) 719-5959