DATA SCIENCE TRAINING AT 84.51°

CONTRIBUTORS:
MARK ROEPKE
BRAD BOEHMKE

DATA SCIENCE TRAINING AT 84.51°

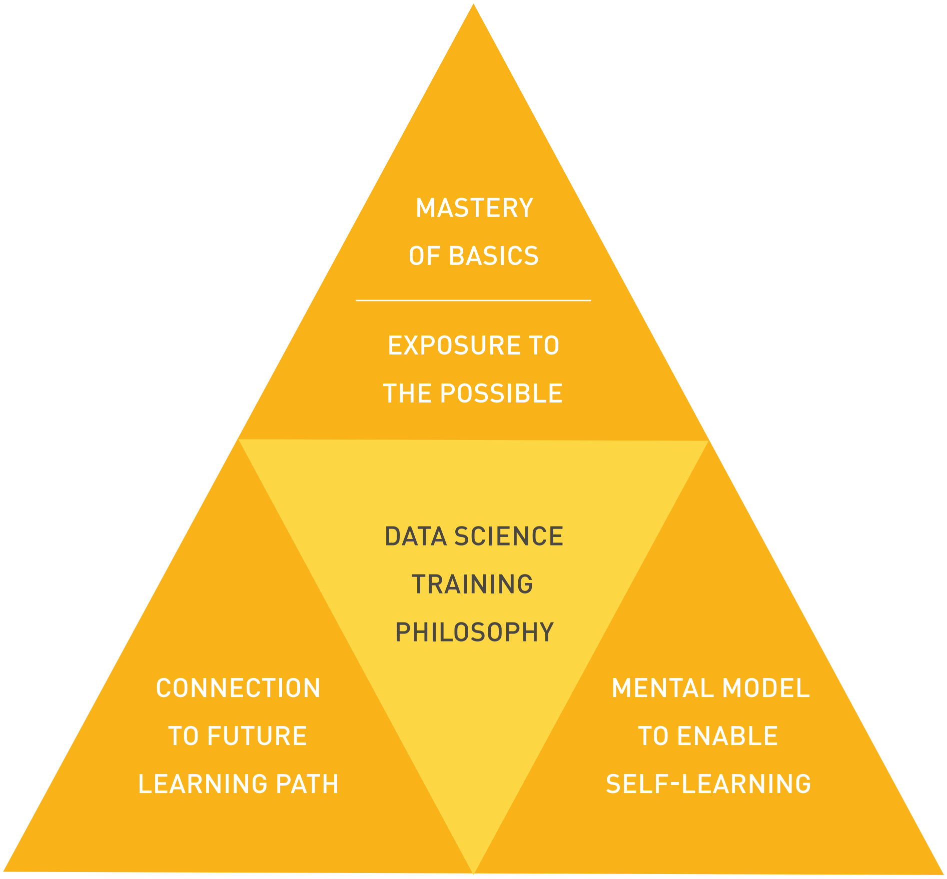
The growing emphasis of data science paired with the quickly-changing nature of methodologies and technologies creates an increase in demand for data science talent. There are many articles and discussions about attracting data science talent to an organization through recruiting techniques and by creating working environments that appeal to data scientists. Both of these are important and effective considerations when creating and managing a data science team. However, continuing to grow your top talent is a whole other challenge. To do so, it is critical for organizations to establish focused and objective-based data science training programs.

TRAINING PHILOSOPHY

At 84.51° we believe internal training can have a significant impact on the depth and breadth of a data scientist's skillset. To do so, our trainings follow an opinionated philosophy: narrowed focus and general objectives for each training.

Our trainings tend to be focused on a specific topic within the realm of data science skills. We find that by narrowing the attention of the data scientist to a specific training topic, the relevance and practicality of the skill can be better communicated. Moreover, the extrinsic cognitive load is reduced leading to better learning outcomes.[1] For example, discussing the idea of containers and operating system virtualization in general can be overwhelming and data scientists often find the importance of such a capability less obvious. These generalized conceptual trainings tend to focus on the *why* of a skillset or capability. However, providing a workshop that discusses the idea and application of dockerizing a Flask or Shiny app narrows the topic and allows us to demonstrate why and how an analyst can apply this capability to scale their current projects.

[1] Ginns, P. (2006). "Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects". Learning and Instruction. 16 (6): 511–525.



Building a training philosophy around these three pillars prepares data scientists to be immediately and increasingly knowledgeable and effective.

There are three key objectives set for each training that we offer in an effort to provide data scientists with:

1 *A foundation of practical skills for self-sufficiency while also providing exposure to advanced capabilities to whet the learning appetite.* First and foremost, we want data scientists leaving the workshop having coded a Python class, R package, machine learning algorithm or whatever the focus is. Second, although the workshop examples tend to be simplified, we also ensure the analyst sees a fully implemented example with all the bells and whistles, so they can be motivated by concrete examples of in-production work.

2 *A mental model to enable future self-learning.*[2] Our data scientists have a wide variety of skills and experience. Consequently, it's important that each topic we cover is explicit about where in the data science pipeline the skill fits, which helps to target who will benefit from the training (i.e. is the skill targeted toward an insights data scientists versus a machine learning engineer). We also emphasize the skill level and required prerequisites. This helps to keep our trainings on-pace, but also gives our data scientists an expectation of what skills we believe a novice, intermediate, and expert data scientist should have.

3 *Exposure to additional learning resources for motivation and a directed path to continue the learning journey.* We have found that workshop trainings lasting 1-2 hours are optimal for maintaining attention while still allowing our data scientists to apply the methods being discussed. However, rarely can you go beyond the fundamentals of a skill within that time window. Consequently, each training also includes specially curated resources so that we can provide our data scientists with the best "next steps" to learn more.

[2] Muller, O., Ginat, D., and Haberman, B. (2007). "Pattern-Oriented Instruction and Its Influence on Problem Decomposition and Solution Construction". 2007 Technical Symposium on Computer Science Education (SIGCSE'07).

This research-backed training philosophy, much of which has been motivated by Greg Wilson's Teaching Tech Together[3] and Garrett Grolemund's "Tidyverse Train-the-trainer" workshop[4], enables data scientists by helping them develop foundational skills and by providing the ability and motivation for continued learning.

TRAINING CONTENT

It's very important for data scientists to have a broad skill set. As Andrew Hunt and David Thomas state in *The Pragmatic Programmer:*

> *"You try hard to be familiar with a broad range of technologies and environments, and you work to keep abreast of new developments. Although your current job may require you to be a specialist, you will always be able to move on to new areas and new challenges."*

This logic can be extended to the data scientist as a programmer and a methodologist — there may be a trend toward specialization, but it's important to have the ability to flex to different types of work. As a result, we offer data science training in both a sequential and asynchronous pattern.

SEQUENTIAL TRAINING

Sequential training helps develop a deep-rooted, detailed knowledge within a topic. In alignment with our training philosophy, each training workshop within a topic increasingly teaches the details necessary for a data scientist to be considered an expert.

INTRODUCTORY	INTERMEDIATE	ADVANCED
USER	IMPROVED USER	DEVELOPER
WHEN TO USE	WHY IT WORKS	HOW TO CREATE
HOW TO USE	HOW IT WORKS	HOW TO SCALE
DAY-TO-DAY TASKS	EFFICIENCY	CUSTOM SOLUTIONS
MOTIVATION	LIMITATIONS	EXPERTISE

Sequential data science training programs enable data scientists to accelerate their journey from beginner to expert in specific skill.

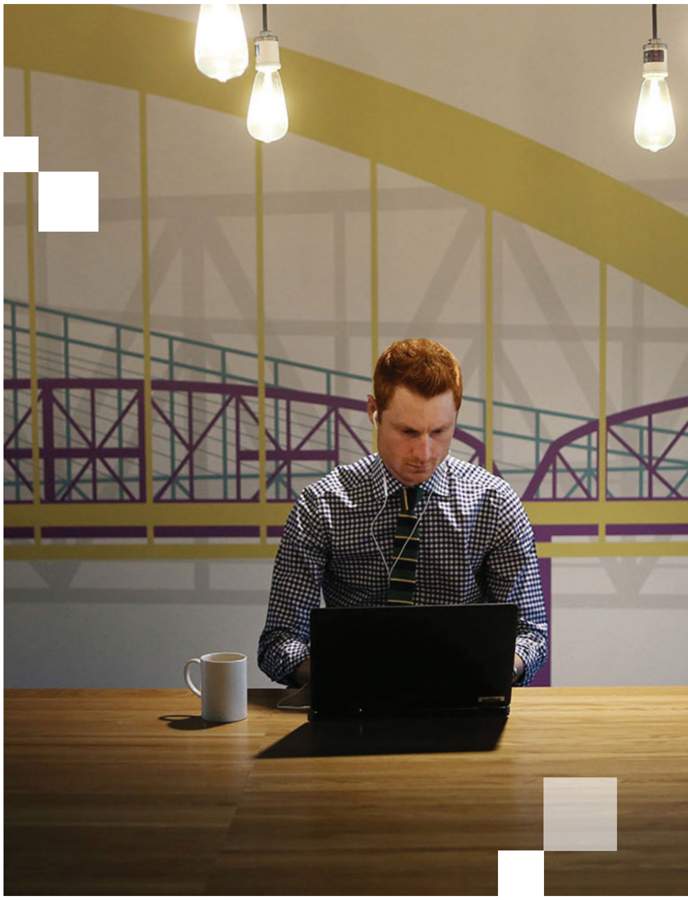
[3] Wilson, G. (2018). Teaching Tech Together. Lulu.com. 978-0-9881137-0-1. http://teachtogether.tech/.
[4] Grolemund, G. (2018). Tidyverse Train-the-trainer. RStudio Conference 2018.

As an example, one of our core data science programming technologies, R, follows a sequential pattern:

1. The introductory courses focus on the fundamentals of data wrangling with R while exposing data scientists to areas of data visualization and working with text data. Introductory courses such as these tend to focus on teaching our data scientists how and when to use a tool or methodology to complete their day-to-day tasks.

2. The intermediate courses are more detailed and programming-centric by teaching how to work with data more efficiently and create more robust programming applications. Intermediate courses tend to dive more into the technical details behind the tool or methodology to better understand how it works and where particular constraints and limitations may exist.

3. The advanced courses focus on expert areas that are still core to R like package development, unit testing, and functional versus object-oriented programming. Advanced courses like these are designed to move our data scientists from simply a user of a tool or methodology to a developer so they can help craft internal custom solutions that scale across the enterprise.



We also follow sequential trainings for other core tools and methodology spaces like Python and machine learning, and we're in the process of building sequential Spark trainings. Each of our sequential trainings provides a similar transition for our data scientists where our first goal is to get them using the tool and methodology for work tasks, then we work on getting them more foundational knowledge of inner workings of the application. Finally, for those interested, we guide them towards developer-level knowledge.

ASYNCHRONOUS TRAINING

As evidenced by the rise of the specialization, data scientists do not need to be an expert in everything. However, our training program aims to provide opportunities for data scientists to learn new material outside of their specialization. This can empower data scientists to expand the possible in their current workstreams while also enabling flexibility of data scientists across workstreams.

To help expand the breadth of our data scientists' knowledge, we offer monthly asynchronous trainings on topics that may not apply to all data science specialties but are still key to many data scientists being effective. For example, workshops on Docker, Shiny, Git, and GitHub help to develop a broader knowledge on technology tools. While workshops on writing good functions, natural language processing, interpretable machine learning, and hyperparameter search approaches help to develop broader knowledge across the methodology space.

Moreover, these breadth-stretching trainings are all provided by internal data scientists, which means they are typically created in a way that allows our data scientists to see the relevancy and practicality of these new topics in their day-to-day practices.

TRAINING DELIVERY

When it comes to the actual material and format of the training, we have five core principles we like to follow:

1 **Training is a code-based living product**
Nearly all training materials are developed and treated as code-based products by using slides and notebooks via RMarkdown and Jupyter and hosted in a centralized "Training" GitHub organization. This has helped to increase the reproducibility and adaptability of our trainings, while also providing organization and issue tracking via GitHub for the training maintainers.

2 **Accompanying scripts and notebooks**
Our trainings emphasize application and participation; consequently, nearly all trainings include code-based resources (i.e. scripts, notebooks) that accompany the delivered training content. These resources are forked by the attendees and help keep them engaged with pre-organized code, exercises and notes, and allow them to add their own information they find useful. This helps to reduce intrinsic cognitive load by highlighting the key areas of the training and keeps the data scientists engaged by focusing on application rather than note taking.

3 **Hands-on work**
Exercises help develop data scientists' foundational skills and understanding within a topic by progressively building confidence, reinforcing key concepts, and challenging them to self-teach something new. We emphasize a wide variety of exercises to include fill-in-the-blank, minimal-fix, pick-the-right-argument, code-from-scratch, etc. Teaching Tech Together (Wilson, 2018)[5] offers a wide variety of exercise types that we've found very useful. Case studies can achieve similar goals with less guidance to reinforce concepts and connect learnings to tasks similar to data scientists' daily projects.

4 **Domain relevancy**
Data, exercises, case studies, and explanations within each training should be relevant to data scientists' daily projects to help the learner interpret topic importance and minimize the extrinsic cognitive load of learning a new data subject. Consequently, we try to place all training topics in the context of data science at 84.51°.

5 **Condensed takeaway**
Aggregated snippets of code and accompanying brief explanations for common tasks within the topic of the training enable data scientists to reinforce their learnings when utilizing their new skills on real projects.

Delivering trainings that adhere to these principles has resulted in better learning outcomes, as evidenced by self-reported post-training surveys given to data scientists.

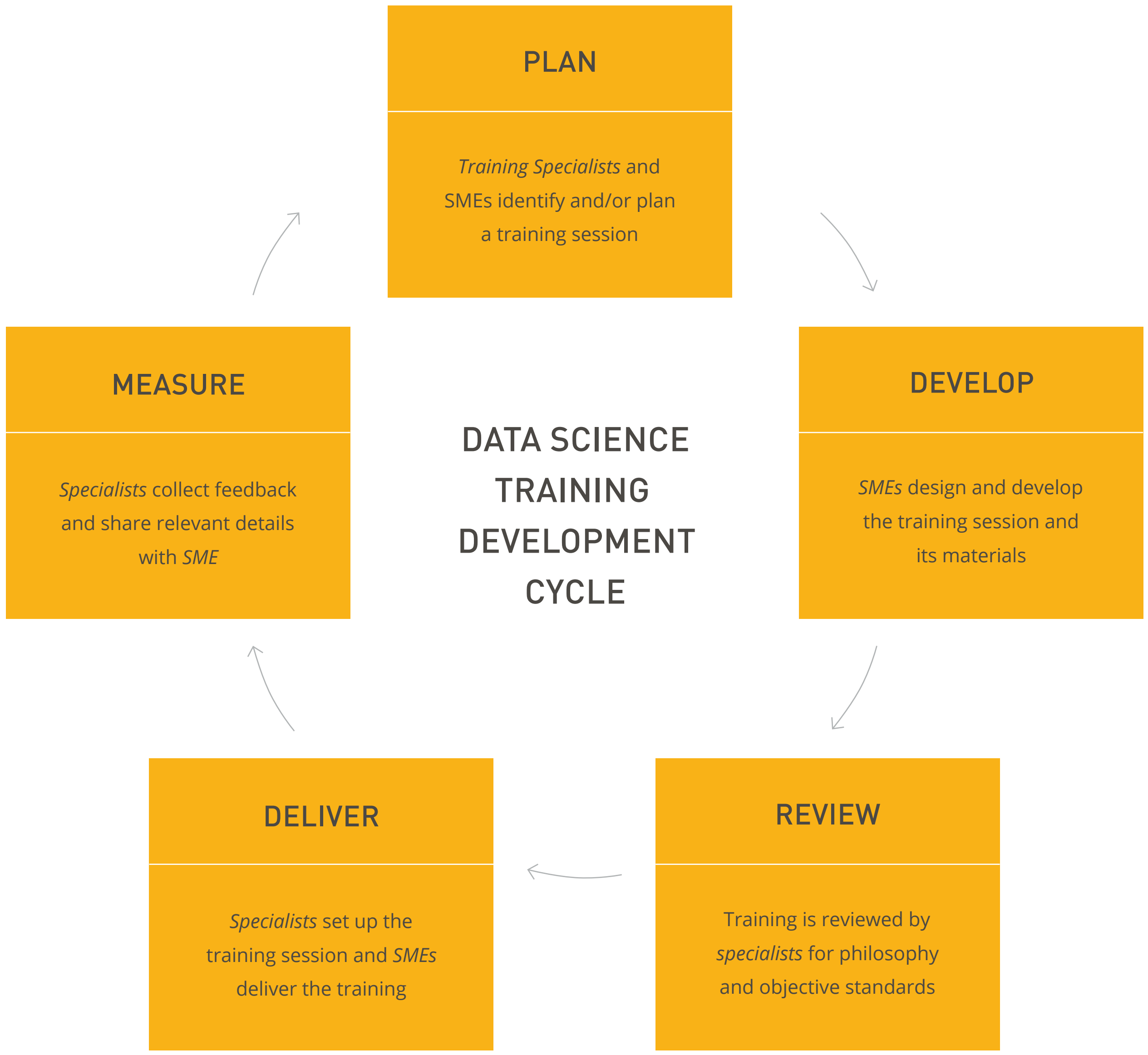
TRAINING DEVELOPMENT

The execution of the preceding ideas is challenging without time and people dedicated to design, develop, and deliver trainings effectively. We've traditionally treated training as side-of-desk work, but multiple teams have recently been established to coordinate, facilitate, and execute the training of data scientists. The result is a more organized training process, more data scientists becoming trainers, and better training performance.

ORGANIZED DEVELOPMENT

We have started to treat our training as organized series rather than individual one-off training sessions. Even our asynchronous training workshops are organized through the Data Science Training Series. These teams work with the general data science community and the power users to identify where knowledge gaps exist, which allows the teams to manage the direction and development of future training sessions.

[5] Wilson, G. (2018). Teaching Tech Together. Lulu.com. 978-0-9881137-0-1. http://teachtogether.tech/.



DATA SCIENCE TRAINING DEVELOPMENT CYCLE

PLAN

Training Specialists and SMEs identify and/or plan a training session

DEVELOP

SMEs design and develop the training session and its materials

REVIEW

Training is reviewed by *specialists* for philosophy and objective standards

DELIVER

Specialists set up the training session and *SMEs* deliver the training

MEASURE

Specialists collect feedback and share relevant details with *SME*

An interactive data science training development cycle enables the creation of standardized SME-developed trainings with the ability to improve over time.

When a new training session within a series is being planned, a subject matter expert in the field works with training specialists to coordinate and create the session. The subject matter expert brings the knowledge so that optimal and practical knowledge is curated, while the training specialists help to ensure the training development process follows a common path. We rely heavily on GitHub projects and custom issue trackers to ensure common milestones are reached in the training development process (i.e. content reviews ensure training follows our general philosophy and format requirements) and the training specialists also relieve the trainer of the training logistics (i.e. reserving space, marketing and communication of the training). The result is often a well-informed, well-attended, and effective training workshop.

DEVELOPING TRAINERS

Historically, trainings were developed and delivered purely by subject matter experts on the topic. However, recently, we have established a mentorship model that allows those that are less-experienced to increase their understanding by preparing and delivering a training topic themselves.[6] This model is focused on our introductory courses (i.e. Intro to R, Intro to Python) where content has already been developed but just needs refinement prior to each delivery and the junior data scientist has already gone through the course.

In this model, a subject matter expert works with the junior data scientist to review and refine the existing content. The actual delivery of the training session is often a combination of the two in a capacity determined comfortable by the novice. For the novice, this helps to strengthen the mental model of the topic and the actual creation of the material develops practical skills in the topic. Moreover, preparing and delivering training can help junior data scientists build skills that are important to data science but frequently overlooked: communication, teaching, planning, etc.

For the subject matter expert, having a junior data scientist involved in the content building process can help identify where certain assumptions in prerequisite knowledge exist or where certain parts of the training may not connect with more junior data scientists. Moreover, students are more likely to approach relative peers following a training than an expert that may be viewed as less accessible. This allows the collaborative peer learning environment to continue well into the future.[7]

MEASURING PERFORMANCE

Lastly, we have been more deliberate about our training feedback mechanisms. Semi-annual surveys for all data scientists help identify perceived strengths and weaknesses of our data scientists, which can help direct future training needs. Moreover, surveys are performed for every training session to identify:

- *how relevant the training is to their daily work,*
- *the quality and clarity of the material,*
- *was the content an adequate mix of presentation and application,*
- *how clear the main takeaways were,*
- *and how often they believe the training should be delivered (i.e. quarterly, annually, on-demand).*

This feedback helps guide content revisions and delivery approaches, and inform future training development processes and training schedules.

THE FUTURE OF TRAINING

The above philosophies and strategies were developed using a combination of experience and academic research with the goal to create an effective and approachable training program for the current data scientists at 84.51°. This approach has significantly increased the variety, effectiveness, and attendance of our data science training. As the data scientist profile continues to change the specific trainings offer will also change, but the general philosophy should remain the same.

[6] Boud, D. (2001). "Introduction: Making the Move to Peer Learning". Peer Learning in Higher Education: Learning From & With Each Other. London: Kogan Page Ltd, 1–17.

[7] Boud, D. (2001). "Introduction: Making the Move to Peer Learning". Peer Learning in Higher Education: Learning From & With Each Other. London: Kogan Page Ltd, 1–17.