Outline

Top of Form



Bottom of Form

**[The Way of the Data Scientist](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/course/0)**

**[Data science as a profession](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/1)**

**[Data science as field knowledge](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/2)**

**[Getting up to speed](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/3)**

**[Interviewing for data science jobs](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/4)**

**[The job search starts now](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/5)**

[Checkpoint 1](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/1)

**Data science as a profession**

**Data is the new (s)oil**

*"Half the money I spend on advertising is wasted; the trouble is I don’t know which half." —John Wanamaker*

Making data-driven decisions in business is not a new idea. For more than a century, companies have employed statistical techniques to control processes and quality in many industries, from manufacturing to agriculture. But what *is* new is the unprecedented scope of the data that can now be gathered.

Take, for example, the above quote by nineteenth-century retailer John Wanamaker. Decades ago, measuring the effectiveness of advertising was extremely difficult, and businesses had a limited number of data-collecting strategies. And the strategies they could use—such as polling a handful of customers or tracking the conversion of a coupon in the mail—were imprecise.

Now, retailers can get a click-by-click history of how users shop. They can directly target certain segments of customers on social media. They can even place sensors inside the store to trace traffic volumes and patterns. This *digitization* of retail has led to a deluge of data, and the breadth and depth of this data are extraordinary. Modern businesses now have accurate information on the effectiveness of all their actions, from what they sell to how and where they advertise.

Data is now what powers most commercial interactions, especially online. From the shows Netflix suggests to viewers, to the posts and advertisements people see on Facebook, to the fare a rider pays their Uber driver, companies of all kinds are leveraging their data to deliver on business objectives.

Data has become so elemental to modern business that in 2006—two years before the term "data scientist" was even coined—British mathematician Clive Humby likened data to oil:

*"Data is the new oil. It’s valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc., to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."*

Humby’s statement gets to the heart of the data science workflow. Because "unrefined data" isn't very useful, it's a data scientist's job to "refine" that data. This work can be messy, unglamorous, and time-consuming. But it's an absolutely indispensable part of contemporary business practices. Without it, companies would have no valuable insights to inform their decisions.

A popular alternative to Humby's analogy comes from British designer David McCandless, who suggests that data is not the new *oil*, but the new *soil*:

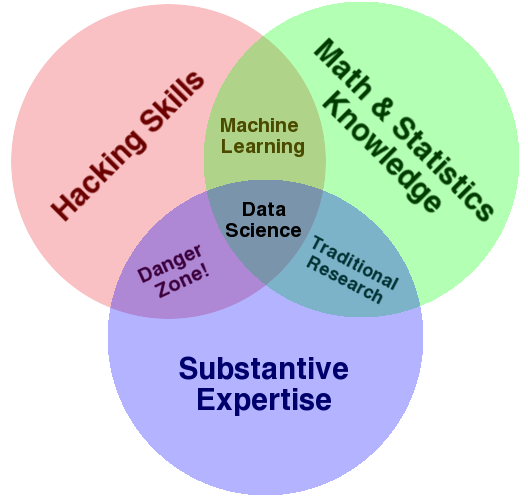
*"Data is a fertile, creative medium … which we can irrigate with networks and connectivity. Data is the new soil."*

For McCandless, data is a necessary ingredient for successful business outcomes. It can, and must, be renewed and repurposed to remain a viable, productive part of company decision-making. McCandless points out that much of the data collected is customer-generated, and that helping those customers—not the data "pipeline" itself—is what makes a business successful.

Both analogies acknowledge that data is essential, helping drive and determine business value across companies, industries, and even the economy at large. And both recognize the time and effort it takes to make data actually mean something. In fact, it often takes a team of experts specializing in different kinds of data science to do this work. Let's take a look at some common specialized roles in data science.

**A data scientist by any other name**

You may remember the image below from a previous checkpoint. Most types of data science fall into one of the three disciplines of the data science Venn diagram:



Some data scientists specialize in hacking. Others focus on math and statistics, and still others work as subject matter experts. What distinguishes someone as a data scientist is that they are still exceptionally proficient in the other disciplines. Engineer Josh Willis captured this quality by defining a data scientist as a "person who is better at statistics than any software engineer and better at software engineering than any statistician."

Not surprisingly, data scientists are often called data scientists. But they can have other job titles, too, which you'll explore below.

**Data engineer, data architect, or machine learning engineer**

These data scientists are not just hackers, but full-fledged software engineers who can handle large amounts of data at scale. They're often experts in data science packages from R and Python. But typically, they also know how to use tools like Java for development, Hadoop for massive-scale data processing, and ETL technologies for data storage and warehousing.

These roles focus on making data work in production. This work includes troubleshooting models that have gone stale, maintaining software logs, and even looking out for security threats.

**Statistician, mathematician, or operations researcher**

Although these roles still exist in many organizations, they could be considered the closest thing to "data scientists" before data scientists really existed. Experts in these positions must perform rigorous quantitative analysis in a business context. Proprietary languages such as SAS, SPSS, and MATLAB are more common in these roles and departments.

Titles like mathematician and operations researcher are especially common in resource-intensive industries, like transportation and manufacturing. These roles are primarily rooted in the mathematical and statistical foundations of data analysis. But professionals in these positions might not be as involved with the *productionization* of machine learning models.

**Data analyst or business intelligence analyst**

While trained in the foundations of data science, individuals in these roles are likely to serve as the vital "go-betweens"—the expert liaisons that communicate with both the data science team and nontechnical audiences and executives. Professionals in these roles are comfortable retrieving data from databases and data warehouses to conduct small-scale analysis in tools like Excel, Tableau, and Python.

Long-term data projects for data analysts might involve maintaining a dashboard and communicating with executives on key performance indicator (KPI) results and strategies. These data scientists could even be versed in user experience (UX) and design methodologies, and they may have experience working with HTML and JavaScript.

**Where do all the data scientists go?**

There is no central organizing committee or licensing board of data scientists, so there’s no official voice of the data science labor market. In fact, "data scientist" is not even listed in the US Department of Labor’s [**Occupational Outlook Handbook**](https://www.bls.gov/ooh/home.htm) because it’s such a new, and ever-evolving, profession. What the Department of Labor *can* tell us, though, is that the employment of computer and information technology occupations is projected to [**grow 13%**](https://www.bls.gov/ooh/computer-and-information-technology/home.htm) from now through 2026. This is a faster growth rate than that of any other profession. So, where do all the data scientists go? Simply put, they go *everywhere*, from the public sector to nonprofits to the private sector.

**Public sector**

Many governments are undertaking [**open data initiatives**](https://www.state.gov/open-government-initiative/), which hire data scientists to collect, analyze, and report on data. In the United States, this work is spearheaded at [**data.gov**](https://www.data.gov/), which is also an excellent source of datasets. DJ Patil—whose name may sound familiar, as he helped coin the term "data scientist"—is also known for being the first [**chief data scientist**](https://www.wired.com/2015/02/white-house-names-dj-patil-first-us-chief-data-scientist/) of the United States. Patil's appointment symbolized the integration of data science into the public sector.

**Nonprofits**

There is a growing movement to make nonprofit organizations more data-driven and digitally literate. Although many nonprofits lack the resources to take on large-scale data projects, these organizations have strong community ties, access to robust data, and important, value-driven missions. Some organizations, like **[DataKind](https://www.datakind.org/about" \t "_blank)**, have become "nonprofits for nonprofits" by helping other mission-driven institutions grow their data science capacities. The Data Science for Social Good Initiative even has a paid [**fellowship program**](http://www.dssgfellowship.org/) to train data scientists to work with nonprofits.

**Private sector**

Data scientists work in all industries of the private sector, from retail to education to agriculture. However, they are most likely to work in knowledge-intensive industries like healthcare, finance, technology, and energy.

**Tech**

The UK network [**Tech Nation**](https://technation.io/news/tech-company-definition/) defines a *tech company* as a "business that provides a digital technical service/product/platform/hardware, or heavily relies on it, as its primary revenue source." For example, Uber offers transportation services, much like a taxi company. But the key differentiator is Uber's technical platform. Lyft, Airbnb, and many other companies operate the same way, using a technical platform to provide even a simple, but in-demand, service.

By their nature, tech companies are highly digital organizations, and this generates a lot of data. This is where the data scientist comes in. Data scientists at tech companies use data to learn about customers, meet business objectives, and validate companies' offerings.

**Finance**

Data science [**has revolutionized**](https://www.investopedia.com/articles/active-trading/040915/how-big-data-has-changed-finance.asp) the finance industry, and data scientists play a valuable role in decision-making around trading and investments. Data scientists in finance also work in more internal-facing roles at banks, monitoring accounts for fraud or risk of default. They may also analyze social media activity and customer reviews to uncover insights about customer experience. And of course, there are the *quants*, or quantitative analysts, who perform rigorous analyses of financial markets and securities.

**Health care**

Data science is used in every area of health care, from forecasting emergency room visits to detecting anomalies in insurance claims. Data science has also directly affected patient care: multiple reports state that algorithms have [**made better diagnoses**](https://www.med.stanford.edu/news/all-news/2017/11/algorithm-can-diagnose-pneumonia-better-than-radiologists.html) than doctors. Hospitals may hire data scientists in a variety of capacities as they work to integrate highly predictive data science models with human medical expertise.

**Energy**

Sensors and the [**Internet of Things**](https://www.merriam-webster.com/dictionary/Internet%20of%20Things) have provided the energy sector with massive amounts of data. Armed with that data, data scientists in this industry often work on forecasting consumption, identifying locations of potential energy sources, and preventing outages.

**Startups**

Startups exist in all of the above industries, but they often come from tech. Working as a data scientist at a startup comes with unique challenges and rewards. While large companies likely have large data science teams with an established data infrastructure, startups are often "scrappier" when it comes to data science resources. In a startup context, the distinction between data science and data infrastructure is often blurred. You might move from being a "hacker" to being an "amateur developer" in this data science environment.

A data scientist's ultimate objective at a startup is to validate who the customer is and what they value. The quicker a startup can learn and iterate on this value proposition, the higher the likelihood of success. Startups are working under the gun, facing short timelines and high pressure to deliver value. On top of that, data can be hard to come by at a small company, which requires further innovation on the part of the data scientist. For those reasons, data science work at a startup can be exciting and rewarding, but it can also be extremely challenging.

**The many hats of a data scientist**

The specific duties of a data scientist vary across companies and industries. But on a fundamental level, data scientists work at the intersection of research, development, and storytelling. They're also privy to highly sensitive information, which means managing security and ethical considerations is another important element of the data scientist's job.

**As a practitioner**

As subject matter experts, data scientists work across a business to identify its pain points and opportunities for improvement. However, they must be careful not to look at the data outside the context of the business. This work could involve staying abreast of industry news, building relationships with professionals outside of data science, and learning about the business processes that actually generate the data that they analyze.

In entrepreneur Steve Blank’s famous words, this role of the job encourages data scientists to "[**get out of the building**](https://steveblank.com/2010/03/11/teaching-entrepreneurship-%E2%80%93-by-getting-out-of-the-building/)" and become more than just technicians. They should be business practitioners who "live and breathe" their industry and understand how to build data science projects on top of this core knowledge.

Additionally, data scientists often play the role of internal consultant or program evaluator for business projects. By honing their skills in project and client management, data scientists can better manage competing demands on their time and prevent never-ending projects.

**As a researcher**

These duties fall largely in line with what is known as the *data science workflow*. With business objectives and standards in mind, the data scientist conducts rigorous quantitative analysis using a variety of data sources. When most people—including data scientists themselves—think about what data scientists actually do, this is likely what they picture.

**As a storyteller**

With a productive and successful data science model in the works, the data scientist acts as a storyteller and presenter across their organization. They use their research to craft narratives, design compelling presentations, and deliver relevant information to a variety of expert and nonexpert audiences.

Data scientists may need to share insights with executives and the C-suite, who are likely more interested in concrete, quantifiable objectives and than the technical details of how those objectives were achieved. Technical audiences, on the other hand, may crave this information. Data scientists may find that sharing these more granular details with other experts and their fellow data scientists may be beneficial, as data scientists often critique and learn from each other’s work. Finally, some organizations may involve data scientists in public-facing communications as well. After all, the data scientist has garnered a legendary public status, bolstered by headlines like "[**the sexiest job of the 21st century**](https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century)."

**As a guardian of ethical standards**

Since its emergence, the data science community has not established many central organizing principles. This has its benefits, including mass democratization: people of all backgrounds can get involved in data science. However, it has also resulted in the absence of some ethical self-policing, which exist in traditional professions like medicine and law. As the data science profession matures and the world becomes increasingly more digital, many data scientists recognize the need for clearer, more explicit standards. The data science community is learning that they must make a concerted effort to consider the ethical dimensions of their projects.

Data scientists often work with human-generated data, including sensitive details about people and their lives. Before even collecting this data, data scientists should have a method for obtaining realistic consent from participants and keeping their data secure. Data scientists with an academic background may see how using an [**institutional review board**](https://www.niehs.nih.gov/about/boards/irb/index.cfm) is a similar approach.

Data scientists must also be aware of the bias or [**redlining**](https://en.wikipedia.org/wiki/Redlining) that their algorithms may produce. As algorithms learn to improve prediction, they may employ unacceptable tactics, such as discriminating on the basis of race or gender. This dangerous phenomenon, known as "machine learning bias," was discussed in a popular book, [**Weapons of Math Destruction**](https://en.wikipedia.org/wiki/Weapons_of_Math_Destruction), by American statistician Cathy O’Neil.

By employing highly networked, sensitive data, data scientists must also consider the risks of a malicious attack on their models. Not only might *black hat hackers*, or hackers with malicious intent, breach customer data, they might even inject faulty data, wreaking havoc on a data science model.

For these reasons and more, the data science community is getting serious about ethics. The current consensus is that instead of developing the data science equivalent of the Hippocratic Oath, each organization or project should publish and adhere to a unique [**checklist**](https://www.oreilly.com/ideas/of-oaths-and-checklists) of principles and values. And now, setting and maintaining these ethical standards is an important duty of data scientists across industries.

**Checkpoint**

This checkpoint will be autograded. Once you complete the assignment below, you will see a button allowing you to submit your answers and move on to the next checkpoint.

Your work

**05.21.20**

**Completed**

Next checkpoint

How would you rate this content?

Thanks! Glad to hear it. Change

[**Report a typo or other issue**](javascript:void(0);)

[**Go to Overview**](https://www.thinkful.com/students/dashboard/me/)

Outline

Top of Form



Bottom of Form

**[The Way of the Data Scientist](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/course/0)**

**[Data science as a profession](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/1)**

**[Data science as field knowledge](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/2)**

**[Getting up to speed](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/3)**

**[Interviewing for data science jobs](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/4)**

**[The job search starts now](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/5)**

[Checkpoint 2](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/2)

# Data science as field knowledge

## What is machine learning?

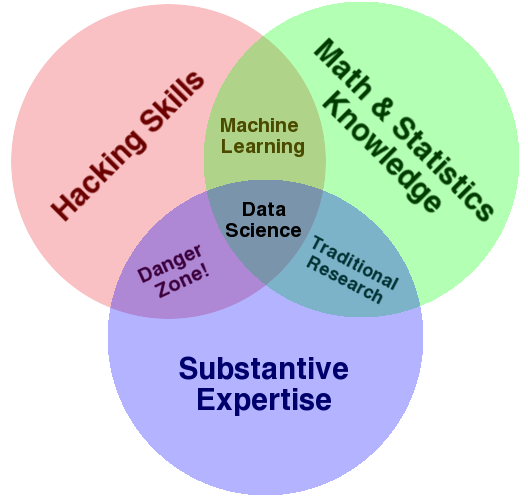
Unlike traditional rule-based statistical analysis, machine learning aims to build systems that can automatically learn and improve from experience without being explicitly programmed. Learning machines are able to adapt to changes in information.

Machine learning includes both supervised and unsupervised techniques. The core distinction is whether there's a known variable. A model is considered supervised if there is a known target variable, because it "trains" the data to learn how to predict it. But a model is unsupervised if it has no known target variable and because it's free to make connections between data points without having to consider an outside target variable.

## Machine learning and data science

Machine learning is often the headline-generating part of data science. But machine learning is not the only thing that data scientists do. In fact, even when data scientists are doing machine learning, they're not doing only machine learning.

To make sense of the relationship between data science and machine learning, you'll revisit the data science Venn diagram. Focus on the relationship between machine learning and data science.



Based on this diagram, you can see that machine learning is a subdiscipline of data science (or that data science includes elements of machine learning). How do these relationships work?

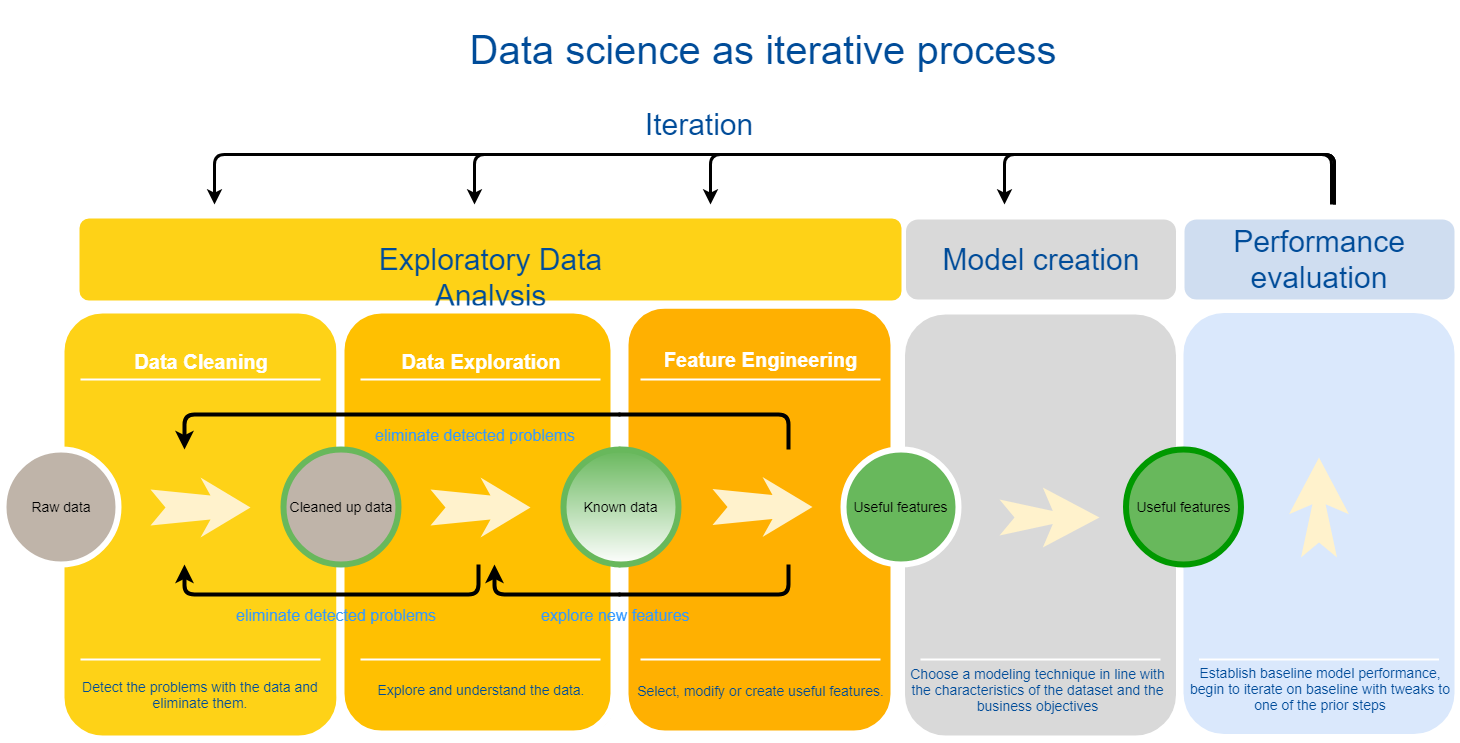
Typically, data science projects start with a single, small-scale analysis, usually motivated by a known business problem or objective. Through an iterative process that you'll learn more about, data scientists learn about the data. It's perfectly acceptable to call this style of analysis "data science." But it's important to note that this is not machine learning. Some data sources, such as customer surveys or clinical trial results, are inherently small datasets that can't become productionized machine-learning models. This is because these sets are small and fairly fixed; they aren't regularly receiving new data.

Machine learning often comes into play when a data source grows at regular intervals and on a large scale. Big, ever-changing data sources like these are overwhelming, and data scientists can't work on them directly. A machine-learning algorithm allows the system to learn from the data without human intervention. However, this can come at the expense of interpretability, or how easily a human can understand or predict a decision or result. Furthermore, there may be bias.

## The data science workflow

As you're beginning to see, data science is not a one-size-fits-all exercise. Data science is a messy, unpredictable practice, and no two projects are the same. However, data scientists still need some guidelines to help them structure and plan their work.

Below is a basic blueprint of a data science project. This is not an entirely linear process. As you'll see, this is a highly iterative blueprint, and data science projects often involve a nonlinear workflow. Many of the steps outlined below overflow and expand into one another, and sometimes it's necessary to jump across or redo steps, depending on the outcome of a particular stage of the process.



### Establish the research question and business needs

Data scientists are not simply statisticians or machine-learning engineers—they also have a strong understanding of business objectives. Even before starting a data science project, data scientists and business stakeholders should explicitly agree on the criteria that they'll use to evaluate whether the project is a success. By clearly articulating the business contexts, research questions, and main objectives for the project, the team can ensure that the results of the data science project have value and can be implemented.

Although this step technically lies outside the core data science workflow, no data science project can be successful without these goals, criteria, and contexts to guide the process.

### Acquire, explore, and clean the data

Like money, good data doesn't just grow on trees. Data scientists have to work for it! Revisiting the "data is the new oil" analogy, you'll see that this step is like drilling for the raw material.

One hallmark of data science is the wide variety of data sources that are used for analysis. Data might be collected from the web, gathered from a traditional scientific experiment, or captured from another activity or channel. Even when the data comes canned—in other words, from a flat file or database—data scientists often need to blend this data with other data sources to uncover value and meaning. For example, a data scientist at a hospital may want to combine the results of a patient survey with the socioeconomic factors determined by each patient's zip code. This could lead to some interesting public health insights.

Once they've acquired data, data scientists need to get a sense of the overall "look and feel" of it. They ask many questions to better understand the scale, dimensions, and complexity of the data they're working with. How big is it? Are there any anomalies in the data? Were these anomalies caused by human error, or are they truly outliers? Is the data complete, or are there missing values? Are variables categorical or continuous?

This process of "getting to know" the data is called exploratory data analysis (EDA). EDA employs a variety of techniques to reveal information about the data, and many of these techniques use data visualization. Rigorous statistical techniques are also used to clean the data of potentially bias-inducing data. The statistical concept of bias is important here; it's the tendency of a sampling method to over- or under-estimate the value of an underlying "true" population parameter.

### Complete feature engineering

If data is like oil, then the feature-engineering step is like the oil-refining process. This is how data scientists transform potential variables into features. Features are designed to make the variables suitable for the kinds of models data scientists will be building.

Feature engineering covers many kinds of tasks, and the process can involve different activities. For instance, depending on the size of the data and the intended outcome, data scientists may need to recategorize or recode. Depending on the assumptions of the model they plan to use, it might make sense to logarithmically transform variables. And when working with text, they may need to convert all the words to lowercase and fix misspelled words.

These examples barely scratch the surface of feature engineering. But every feature-engineering technique has the same underlying objective: to build features that are more efficient, more robust, and more interpretable in a model.

### Create a model

Many people believe that model creation is "what data scientists do all day." And although this activity is important, in reality, building and executing a model is only a small part of the workflow.

A common mistake of inexperienced data scientists is to build the most complex model, with the most data possible, on the first attempt. This method poses a couple of issues. First, the more complex the model is, the harder it is to troubleshoot when something isn't working. Second, when a data scientist uses more data at the outset, they run the risk of building a faulty model—one that has based its results on spurious patterns in the given dataset. In other words, the model can't "generalize" to other data points because it has been designed to model the so-called "training" set, in particular.

This problem is known as overfitting, and it's something data scientists constantly look out for. To make sure that their models don't overfit their data, data scientists separate their datasets into training, testing, and validation datasets. This redundancy ensures that the model can hold up to changing datasets.

For these reasons, data scientists consider their first model a rough draft—or, in [**lean startup**](http://theleanstartup.com/) terms, a [**minimum viable product**](https://www.techopedia.com/definition/27809/minimum-viable-product-mvp). The first model should capture just enough relationships in the data to be useful, without becoming overly complicated. You don't want your first model to be too difficult to troubleshoot or overfit the data.

### Evaluate the model's performance

At this point in the data science project, the results are finally in, and the model is ready to be evaluated. So, how does the model perform?

As mentioned previously, there is no one criterion by which to evaluate a model. Data scientists and their stakeholders use the business objectives, evaluation criteria, and standards of success that they established earlier in the process to measure the effectiveness of a given model. This way, there's no "cherry-picking the results" to draw an impressive conclusion.

A model is rarely, if ever, a slam dunk on the first try. When evaluating a model's performance, the data scientist gathers clues on what might be done differently in the next round. They might decide to train the model on more data, include new features, adjust the current features, or try another model entirely. Going back to the drawing board is certainly not a failure, and it happens frequently. There is no one-and-done in data science.

Even a successful data science model needs to be validated. Remember the warning by John Mount and Nina Zumel:

*"The worst possible modeling outcome is not failing to find a good model. The worst possible modeling outcome is thinking you have a good model when you don't."*

In supervised machine learning, the goal is to design an algorithm that will develop better predictions when it's presented with more data. That's why it's so important to test that the model can perform well when it's provided with data that it has never encountered before.

### Iterate

After moving through each step, the data scientist should have a model and a baseline performance of the model. Now it's time for them to iterate by trying different approaches to the model.

In many ways, iteration in data science follows the [**agile methodology**](https://agilemanifesto.org/principles.html) of software development; in other words, there's a preference for quick, incremental pushes in improvement. Making small tweaks to the model at each iteration makes it easier to diagnose what's working, and what isn't working, in the model. Other improvements in the model can come from incorporating insights from the business domain and subject matter experts. Ultimately, iteration is often more about fine-tuning the model than reinventing it entirely.

Over time, each incremental improvement in a model's performance comes with additional effort. And at some point, it becomes necessary to drop the perfectionism and ship the data science project. This often means putting the model into production, which means the model is incorporated into the day-to-day operations of the business. This is likely what the model was built to do in the first place! The final model could be any number of things—from a product recommendation algorithm that shows Amazon customers various items they might want to buy, to a tool that physicians use to interpret whether a tumor is malignant. Once in production, the model will be a business or research tool, working to meet the objectives that informed its creation.

## Checkpoint

This checkpoint will be autograded. Once you complete the assignment below, you will see a button allowing you to submit your answers and move on to the next checkpoint.

Please click **Start** when you are ready to begin the activi

Outline

Top of Form



Bottom of Form

**[The Way of the Data Scientist](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/course/0)**

**[Data science as a profession](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/1)**

**[Data science as field knowledge](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/2)**

**[Getting up to speed](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/3)**

**[Interviewing for data science jobs](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/4)**

**[The job search starts now](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/5)**

[Checkpoint 3](https://courses.thinkful.com/dsbc-way-of-data-scientist-v1/checkpoint/3)

# Getting up to speed

In this checkpoint, you’ll learn about key institutions in data science as you take your first steps into the job search.

As a “[**digitally native**](https://www.techopedia.com/definition/28094/digital-native)” profession, the data science community is particularly vocal online. Data scientists have also built a strong network on the local and personal level. A benefit of getting involved with your local network is that your next job is likely to be based locally, so getting to know those already in the industry will go a long way towards being a competitive candidate. In this way, networking bleeds into applying and interviewing for jobs, so you should start to have an idea of the kind of company you’d like to work for as a data scientist.

# Grab your seat at the data table

The data science community are a welcoming and often opinionated group. By immersing yourself in popular data science resources, you’ll slowly start to “speak the language” of data science.

## Social media

One easy way to add some data science to your life is by following data scientists in your social media feeds. Many of the best and brightest data scientists post regularly on Twitter, using hashtags such as #pydata (for Python), #machinelearning, #datascience and #dataviz (for data visualization). For suggestions on particular data scientists to follow, check out [**this Twitter list**](https://twitter.com/reshamas/lists/notable-data-scientists/members?lang=en). Of particular interest for aspiring data scientists is Renee Teate’s account, [**@BecomingDataSci**](https://twitter.com/BecomingDataSci).

LinkedIn also has a vibrant data science community. Some data science LinkedIn groups boast tens of thousands of members, including [**Data Science Central**](https://www.linkedin.com/groups/35222/) and [**Data Mining, Statistics, Big Data, Data Visualization, and Data Science**](https://www.linkedin.com/groups/152247/). A further list of LinkedIn profiles to follow is available [**at KDnuggets**](https://www.kdnuggets.com/2016/09/top-big-data-science-leaders-linkedin.html), a site which we’ll talk about in a bit.

## Websites

Data scientists have their own veritable “blogosphere” and more of websites, broadly falling into three categories: sites for collaboration, getting help, and sharing knowledge.

**Collaboration.** Perhaps the sine qua non of data science websites is GitHub. This will be the source of your data science portfolio and the host of code-based projects ranging from open-source software development to Fortune 50 model deployment. GitHub serves as a place for data scientists to showcase their previous projects, and collaborate on new ones.

If GitHub is the platform for collaboration, then consider [**Kaggle**](https://www.kaggle.com/) the platform for competition. While hosting a vital collection of datasets and an active message board, Kaggle is best known for its data science competitions, in which teams of data scientists compete for the best model. Many of these competitions are hosted by real corporations, using real data — and with real rewards.

For example, Netflix offered a million-dollar reward for the algorithm that was best at predicting movie recommendations — not a bad incentive to practice data science! To make things even better, the results of each competition are available after the contest, so data scientists can learn from and compare the results of other contestants. For example, the results of the winning Netflix algorithm are [**here**](https://www.kaggle.com/netflix-inc/netflix-prize-data).

**Getting help.** Knowing where to turn when stumped is essential for data scientists, and usually the place to go is **[StackOverflow](https://stackoverflow.com/" \t "_blank)**. This is a question-and-answer board that uses up-voting, reputation scores, and other features to validate the best answer to any question. That said, it’s not a good idea to post a question on StackOverflow without doing extensive research and preparing a reproducible example first (See StackOverflow’s guide to [**writing a good question**](https://stackoverflow.com/help/how-to-ask) for more information). Forum members demand high standards from posters, which can easily blur into haughtiness and plain rudeness. This can be a jarring, unwelcoming experience for new coders, and **[StackOverflow itself](https://stackoverflow.blog/2018/04/26/stack-overflow-isnt-very-welcoming-its-time-for-that-to-change/" \t "_blank)** has acknowledged the problem.

Reddit also provides an active platform for getting help with data science, with subreddits like [**Data Science**](https://www.reddit.com/r/datascience/), [**Learn Machine Learning**](https://www.reddit.com/r/learnmachinelearning/)and [**Learn Python**](https://www.reddit.com/r/learnpython/).

**Sharing knowledge.** Many sites serve both to keep data scientists informed as well as to provide them a place to share their knowledge with the community. **[KDnuggets](https://www.kdnuggets.com/" \t "_blank)** is a leading site in the former, and [**Data Science Central**](https://www.datasciencecentral.com/) in the latter. Check out a more extensive list of educational resources with Thinkful’s [**Tech Industry Media Resources List**](https://docs.google.com/document/d/1lfK2M3fCUzgBKpMD9fFIbkecpAPx3sMk-lUTm4xAfaE/edit).

Of particular note in the data science “blogosphere” is the Medium account, [**Towards Data Science**](https://towardsdatascience.com/). This has become a premier source for data scientists to learn and share about their field, almost to the extent of what GitHub has become to collaboration and StackOverflow to getting help.

## Local interest groups

Throughout this program, you’ll be “geeking out” on difficult, technical subjects. It can be a huge relief to find professionals who get paid well to do what you are spending so many painful hours on now. Here are common ways to meet data scientists in your area:

**Meetups.** Check [**meetup.com**](https://www.meetup.com/) for data-related meetups in your area. These usually take the format of hour-long presentations (including question-and-answer), possibly followed by socializing and networking. Some meetups are platform-specific; for example, an R or Hadoop Users Group. Others feature a variety of topics in data science. Meetups give you a glimpse of what professionals in your area are doing with data science, and it strengthens your burgeoning data science network.

**Kaggle workgroups.** This is a sub-set of data science meetups that takes a different format. Rather than a formal presentation, these groups will casually meet to discuss and work on a Kaggle problem. This is an opportunity for you to work on data science on a team, in real life. Plus, maybe you’ll get a cut of a million-dollar prize!

**Hackathons.** These are often weekend-long sprint-like competitions with the goal to build, or “hack” some digital product. While most hackathons are based around general software development, there is a growing movement of “data hackathons,” when teams compete to build the best model in a short period of time. Pro tip: while round-the-clock participation at hackathons are a great experience, it’s also okay to just show up at key events, such as kick-off and final demos, to meet and network with other hackers.

## Who does the data life best?

When you begin to survey the data science profession’s landscape, you’ll start to hear some companies mentioned repeatedly: Google, Airbnb, Facebook, etc. This is the “Ivy League” of data science teams, and while getting into them may be a dream for some, it’s a stretch for most. That’s okay because becoming IBM’s next AI engineer is not the best choice for everyone.

In fact, starting in data science at a less-established company may allow more room for ownership and growth. Or perhaps you’re more interested in a data science role with high social impact. Start to think about what might be a good fit for you as a data scientist, given your wants, needs, and circumstances. There’s no one “dream job” for a data scientist!

## Checkpoint

This checkpoint will be autograded. Once you complete the assignment below, you will see a button allowing you to submit your answers and move on to the next checkpoint.

Please click **Start** when you are ready to begin the activity.

<https://twitter.com/jeremyphoward>

I'd love to work for 'Tasty Trades' because they're inspirational. I know the owner - Tom Sosnoff and I always listen to their podcast about financial markets. They have built one of the best trading platforms "Thinkorswim", they're great at connecting people from all over the world. They have the latest technology, there's an amazing professional ladder to grow, where I can learn from the best. They have an amazing team of data scientists. They specialize in trading options, which is fascinating and difficult to do. I'd love to be in their progressive team. Tom pushes everyone for growth. Data scientists get to research the global market, prices, different industries: their researches let the company make big profitable investments in new found companies, which returns over 100% over time. I'd love to work there and absorb every single thing they teach me.