whitehat analytics

Harnessing the potential of big data

How to use data science to transform your business

Finn Wheatley
Director of Data Science, Whitehat Analytics

About the author

Finn Wheatley is a specialist in building production-grade, enterprise-scale strategic data science capability. He has more than a decade of experience working with and engaging staff at all levels in data-intensive environments. He holds an MSc in Computer Science from University College London, an undergraduate degree from King's College London and worked for several years as an investment analyst in the hedge fund industry. He has experience across multiple industries, including consumer-facing, public sector and financial services. He has helped establish and scale data science teams from the ground up in several large organisations, including the Department for Work and Pensions and EDF Energy. He combines business awareness and commercial expertise with a clear vision of the role of data in a technology-driven organisation.

HOW TO BUILD A DATA SCIENCE FUNCTION IN YOUR BUSINESS

Many organisations understand the potential in exploiting the data they hold to unlock operational efficiencies, create new products and improve customer experiences. It is all too common, however, to see business leaders fall into the trap of investing in one-off solutions, which are often driven by technology rather than goals, come with a large price tag and deliver mixed results. To ensure a positive outcome that endures in the long-term, it is important to set realistic goals and establish models of success from the outset.

Building a data science function from scratch is not a short-term undertaking. The most effective approach is to invest in creating a strategic data science capability within your business. Rather than using data analytics piecemeal, a successful data science strategy achieves a cultural change, moving an organisation towards using analytical insights day-to-day to generate sustained business value.

Setting up a data science capability within any business involves time and effort – and the specialist help of experts. We recommend starting with one small project. If you set out to solve one incremental problem, it will be more straightforward to prove its worth, as well as secure buy-in for the approach from the top.

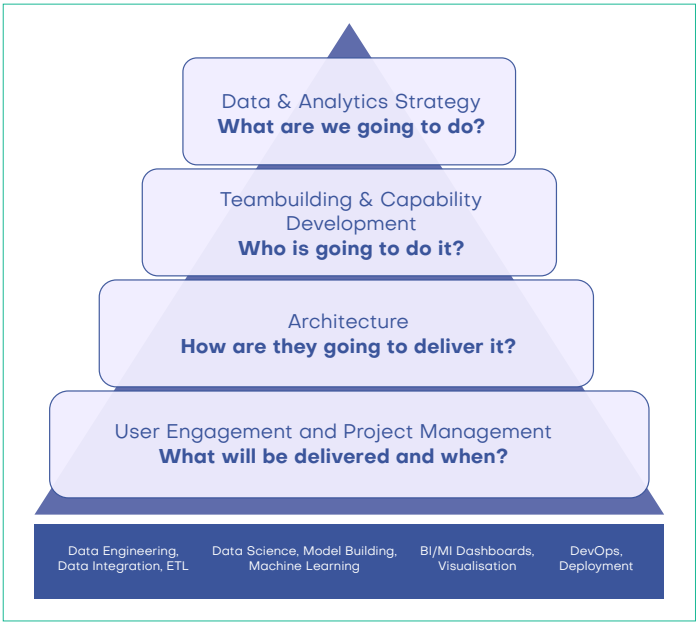

Figure 1 The major elements to consider when establishing a data science capability

1. GETTING STARTED

1.1 Securing buy-in within your organisation

Any change to an organisation must be managed carefully and setting up a data science team is no exception. Changes to established procedures, technologies, and skill requirements can all cause concern, particularly among frontline staff, where there may be a perception that automation could threaten, rather than enhance, their jobs.

Being data-driven does not, however, mean being controlled by technology and it is important to emphasise to all staff that data science, and related technologies such as automation and robotics, will enable them to focus on the most productive parts of their jobs – the elements where their human skills are most needed.

1.2 Selecting your data science team model

Broadly, there are two models that a business can adopt. The most common is the hub or centre of excellence model. The central data science hub will contain all the organisation's data science skills and resources, as well as the appropriate project management, business analysis and user-focused functions to create and deliver projects.

This group may be housed in a quasi-independent experimental, R&D-focused ideas lab or contained within a broader business unit. In contrast, a seed-based team or virtual team involves inserting senior data scientists into external development teams to provide data science expertise in areas such as model building.

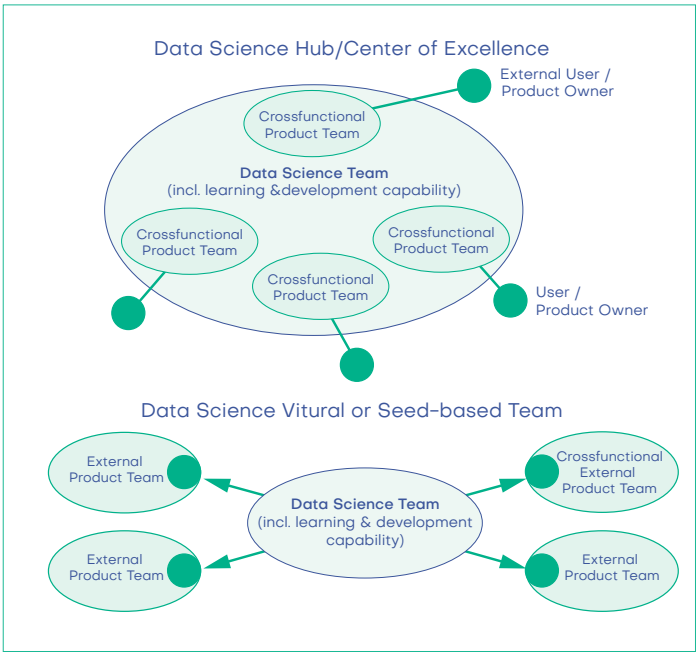

Figure 2 An illustration of the 'hub' versus 'seed' models for a data science team.

There are advantages and disadvantages to both, but since the seed configuration is difficult to effectively use without a framework of experienced data scientists, we typically recommend the hub model.

1.3 Building your first data science project

It is important to first prove the value of data science to your organisation. The key objective is to find a problem where you can build a firm business case that identifies measurable and sustained business value with as much certainty and as little complexity as possible. We recommend starting with one definable project, which should have five key characteristics, illustrated below.

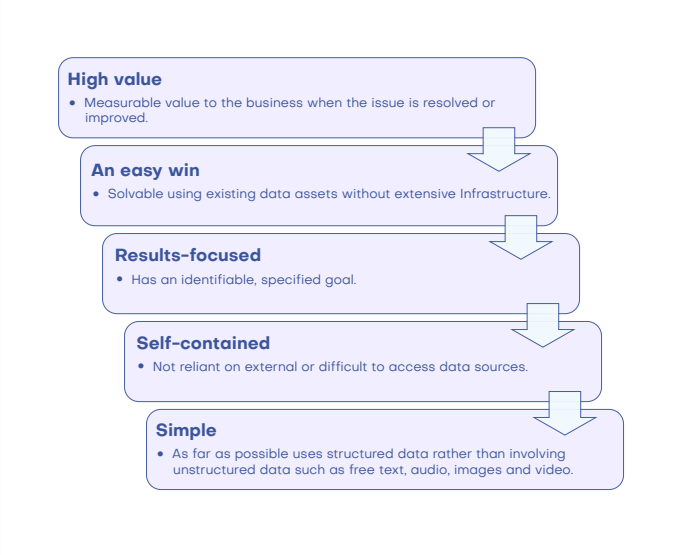

Figure 3 The five factors to consider when choosing a first data science project

The best place to start is with an incremental problem, where a solution is already in use, and the objective is to improve results. This makes it more straightforward to benchmark your progress. Depending how developed data use is within your company, this could be as simple as building a useable, interactive dashboard designed to replace an outdated mishmash of spreadsheets and presentations.

Alternatively, it could involve developing and incrementally improving the performance of an existing model. Ask yourself some key questions. For example, if your sales department currently forecasts using linear regressions and simple seasonal adjustments, can you do any better with a machine-learning driven approach? Or if your website currently uses a simple rules-based approach to upsell or cross-sell potential clients, could you improve results with a recommendation system?

1.4 Start with your customer data

A good starting point is your customer data, typically a large dataset and one which is often well structured but frequently under-utilised. This lends itself well to solving incremental problems, such as cluster analysis for customer segmentation, classification of high value customers for marketing purposes, customer churn modelling, and developing recommendation engines. These are all well understood problems which makes them a great starting point for finding a solution – allowing the team to compare different approaches and choose the one that works best. Figure 4 lays out some of these ideas.

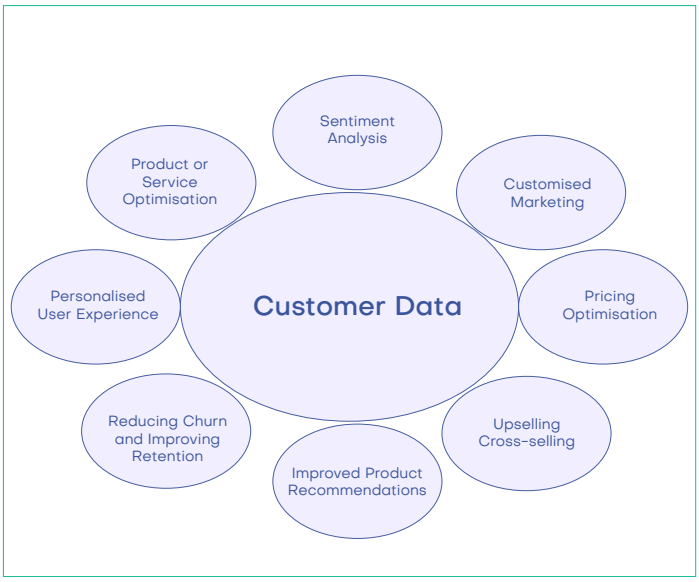

Figure 4 Some ideas of potential business problems that can be derived from customer data.

1.5 From concept to production: creating your first prototype

Your first prototype data product gives you a proof of concept to test the pipeline, develop your team's skills and generate buy-in from your stakeholders. This self-contained product will incorporate all the major elements such as data pre-processing, modelling and an appropriate method of surfacing the results to users, such as visualisation. The focus in the first six to twelve months should be on developing technical and professional capability within the team.

2. BUILDING YOUR DATA SCIENCE TEAM

2.1 Optimising ways of working

We always recommend Agile methods when delivering data products.

There are two critical, non-negotiable components of Agile which make it so successful

- Iterative delivery
- Staying close to users.

When developing software, the largest source of uncertainty is the user and their changing priorities. Agile developed as a method of handling this change, to ensure the development team was always working on whatever was most important to the user at the time and could shift focus quickly if needed.

2.2 Fail fast and learn fast

Agile methods work well in data science because data products have an additional layer of uncertainty compared with regular software. In data science, it is often unknown at the beginning of the project whether the problem can be solved using the proposed approach. It is

common to conclude after a large amount of research that additional data is required, often requiring additional processing. It may be discovered that the marginal gain in a certain project from a data science approach is not worth the additional work, and the project is cancelled.

For this reason, rapid development of proofs of concept is important. This means that if the project fails, it does so rapidly, allowing the team to quickly conclude whether the proposed approach is feasible and will generate the projected results.

Small, agile data science teams can research possible use cases for the company's data and develop prototypes accordingly. Once the concept has been proven, a separate digital team will build and maintain the data product.

2.3 Creating an effective data science team

An effective data science team is made up of people from across the business with different skillsets. As well as data scientists, the team will need:

- Data engineer, often called an ETL developer
- Data visualisation developer
- DevOps engineer, sometimes called DataOps or MLOps in the context of data science
- Software engineer (full stack)
- Data architect
- Data analyst, for specific domain knowledge of the data.

A balanced data science team of eight members might contain one of each of the above, plus a business analyst and a product/delivery manager. With just one or two data scientists in the team, it shows the importance of the skills mix.

2.4 Match your team to your business

Every data science team must be customised to its organisation. Companies with a large demand for dashboards might add members who specialise in designing visualisations. Those with especially complex data requirements – such as handling large amounts of streaming or unstructured data, or with complex data matching and entity resolution needs – should include a larger number of data engineers.

Data architects do not usually take a permanent position in the data science team. This role is critical during the first 6–12 months to specify, design and build a data platform. The platform is usually then maintained and developed by the DevOps engineers. For especially large or complex architectures, it can sometimes be worth employing dedicated platform engineers, with specific experience in creating cloud robust infrastructure on your chosen cloud provider.

Another common role is the machine learning (ML) engineer. A hybrid of data scientist, data engineer and software engineer, the ML engineer specialises in writing production quality data science models. Most companies do not need a ML engineer during the initial phase of team building, but it may prove more efficient to add one at a later stage if the data scientists are working inefficiently or are stretched too thin.

2.5 Find a balance of skills and experience – and then develop both

As important as building a team with a range of skillsets is finding the right balance of experience – and the effectiveness of a data science team depends on its quality rather than quantity. It is worth keeping the ratio of juniors to experienced team members low, as their speed of development is heavily influenced by the amount of interaction with more experienced colleagues. It is worth investing heavily in learning and development, especially in the first six to twelve months. Consider establishing a training budget, and encourage each function to cross-pollinate by sharing their expertise with the wider team. Consider establishing 'side projects' to allow team members to try out different techniques and skills. You can tilt your learning and development toward areas you think might be of use to the team in the future. Figure 5 offers some ideas.

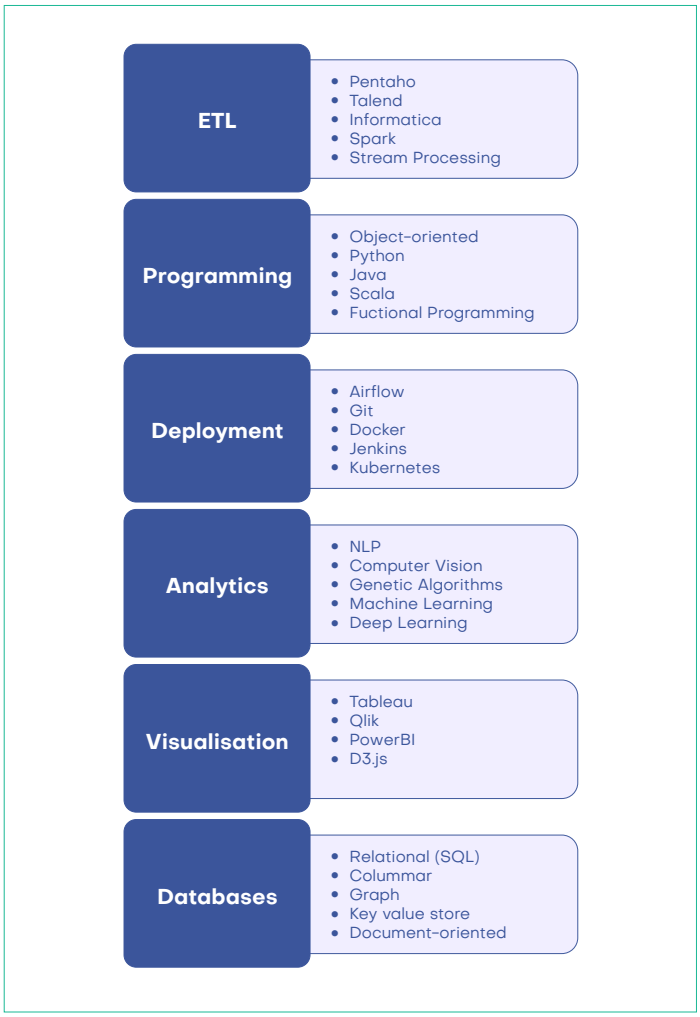

Figure 5 Some potential areas of focus for a data science team's learning and development plan

3. INFRASTRUCTURE AND ARCHITECTURE

3.1 Critical success factors

The infrastructure is a key success factor in building a data science function. By making it easy for your team to explore data and then build and deploy data products, you can substantially increase productivity. Your infrastructure will need to support:

- ETL pipelines (data ingestion, data quality validation)
- Data processing (data asset creation and curation)
- Data exploration and research
- Software development
- Data visualisation
- Test and deployment
- Enforcing prescribed compliance, audit and governance standards.

A number of these tasks can be fulfilled by utilising a few mature, well-established components or platforms.

3.2 Building a data platform

The base component of a data platform is the data lake. Data lakes can be thought of as a data warehouse for unstructured data, or a data warehouse without (most of) the rules.

Data lakes are permissive rather than prescriptive – they accept many data types and file formats. One important factor to consider is the importance of metadata in a data lake. Because there are no restrictions or rules on data types, it is critical that metadata are scrupulously logged on ingest to a data lake, together with the master data and reference data standards. A data lake without this metadata is known as a data swamp.

The specifics of data lake architecture depend on the exact scale and nature of the company's data and hosting solution. Divergences in lake vs. warehouse architectures largely stem from increases in the speed, scale and diversity of the data, as well as a desire to take advantage of the flexibility of the cloud and the different open source tools available.

3.3 Advantages of the Cloud

Your infrastructure will need to be both flexible and scalable. Data science workloads are typically unevenly distributed and the most computationally intensive task is usually to train your models.

The training process may account for the least time, at a few hours per day, but most of the cost. The cloud can effectively and efficiently process these workloads, while in on-premise systems you may have to decide between taking longer to train your models or allowing resources to go unused during low demand periods. On-premise systems can be inflexible and adding additional resources is a laborious task that requires adding extra servers.

For all these reasons, we generally advise our clients to opt for the cloud. The big three cloud vendors – Amazon Web Services, Google Cloud Platform and Microsoft Azure, are all well-developed. Other vendors such as Oracle and IBM can be selected, especially if there is a pre-existing relationship, although they are noticeably less mature.

3.4 Using managed services to jump-start capability

Using managed services from your cloud provider supports rapid set up of a prototype data lake. Configurability and scalability are greatly enhanced by the fact that the different components are designed to be easy to deploy and integrate on your cloud platform of choice.

This is usually by far the fastest way to build capability and can substantially reduce maintenance time, since the service provider will handle many common issues. Integrating all the different components in common ETL or a data science platform such as Hadoop is not always straightforward but, by using a mature cloud platform, this can be handled by the cloud provider.

3.5 Productionising the project

Deployment and maintenance can often be neglected in a data science development pipeline. It is critically important to consider how your product will be rolled out and maintained from the outset, rather than scramble to find a deployment solution after development.

The choice of infrastructure and the knowledge of the user can both affect deployment options. Deploying on cloud gives a large advantage: a simple dashboard that can be surfaced to senior managers makes an attractive first project. Getting your products in front of decision makers and clearly demonstrating their value goes a long way towards gaining top level buy-in from the outset.

A sophisticated web application designed to be deployed to thousands of call centre staff and save millions may sound great, but if those staff operate on an inflexible legacy infrastructure, it could also cost millions to modify it to suit your needs. Or if service centre agents have to juggle half a dozen windows to access different source systems to answer simple customer requests, is the new system really helping? When it comes to deployment, it always pays to ask the simple questions early – and double check the answers.

4. THE DEVIL IS IN THE DATA

4.1 The role of the data engineer

Data engineers build pipelines that move and manipulate data from source systems to the data lake, creating data assets from the raw data: .

- Building data ingest pipelines
- Data quality and validation
- Data matching or data integration
- Data asset creation and maintenance.

Organisations with low quality or diverse data sources and formats, especially involving unstructured data such as free text, video, images, audio, IoT and streaming data, will have to spend more time on data engineering than those with more structured data sets..

4.2 Understanding data assets

A data asset is a key idea in data science. It is a special dataset that is essential for your production data analytics process that must meet a number of qualifying criteria:

1. It is subject to a data quality and validation process to ensure integrity
2. It is typically subject to a further level of processing or aggregation
3. It meets a clear business need
4. It is usually a composite, drawn from at least two or more data sources
5. It is live data, which is updated on a frequent, usually daily, basis
6. It must be a single source of truth – i.e. each data element must be generated in a uniform way.

A simple example of a data asset might be a table that contains each customer's current address and contact details. An example of a more dynamic data asset might be a table listing each customer's most common recent purchases. A retailer could require this information for use cases ranging from management dashboards tracking how purchases change over time, to analytics predicting what different customers buy at different times of year.

4.3 Creating data layers

Data science modelling combines all the data assets to generate new ones, resulting in layers of data assets, each in a more highly aggregated and processed form. An example might be a customer master data asset, which integrates all of the data a company holds on each of its customers into a standardised or aggregated form. This could include customer demographic data, the sequence of customer interactions with the company (such as orders, payments, weblog data), contact information and natural language data.

4.4 Using your data assets

This customer master data asset could be used to answer almost any question asked about any customer, from the marketing techniques they are most likely to respond to, to predicting how to tailor customer service to reduce the risk of a complaint.

Data assets are not static. Data engineers must be able to update them when new requirements emerge. Most of a data engineer's time can be spent building data assets as the value added can be substantial, avoiding the need for data scientists to manipulate messy data every time they work on a new product. Thinking in terms of reusable data assets can substantially increase your data science team's productivity and deliver consistent and easily comparable results.

4.5 Data modelling and machine learning

The core process of data science involves exploring and integrating different sources into a dataset and building machine learning models on top. Because this process is, and should be, truly exploratory, it will involve frequent sharp reversals and sudden stops.

By providing your data scientists with the data, architecture and support required, this core investigation and modelling process can be one of the most rewarding elements of a data science project – and is where data scientists really come into their own.

We typically advise using tools that the data scientists are most comfortable with – with a standard Python-based data science toolset being our preferred choice, because of its maturity and depth of open source community. An alternative could be Java, or one of a set of one-stop data science studios which can be beneficial for small or more inexperienced teams as they can show results rapidly.

There is always a trade-off between standardisation and flexibility, and our experience is that specifying a standardised data science stack is usually the best approach.

5. USER EXPERIENCE: MAKE IT SIMPLE

Returning the results of your model in usable form is one of the most important factors that distinguishes data science from analysis or statistics, and we always advise investing time in defining a clear approach, especially where the dataset is large, complex and multidimensional.

Sometimes, the data scientist will be simply delivering the results of a model to an API for other development teams to use as they wish. Often, however, the data science team must return results in the form of visualisations such as graphs, maps and charts, as well as text. Closing the loop ensures that the potential benefits of your model are realised and show the business value the data science project delivers.

It is critical to consider your user and their level of subject knowledge. It is also important to consider the diversity of the user base. If user needs are diverse and distinct, consider building multiple dashboards for each user type and always use the simplest tool that can do the job well.

6. DON'T GO IT ALONE

Setting up a data science capability within a business offers the potential to deliver substantial operational efficiencies. Embedding data science throughout the business has the goal of delivering next level organisational performance together with a strong return on investment. As a business, it is critical to look at the big picture from the outset and focus on the end-goal rather than get immersed in the detail, which is often technology-related and can detract from the project objectives.

Establishing a data science function is a strategic undertaking that should be viewed as a long-term investment, which will start to show results within the first six months to a year.

Enlisting the help of specialists in the field will drive a results-focused approach, help you to upskill your teams, advise on the right technology, and keep an essential eye on the detail – allowing you to focus on the business outcomes you need.

About Whitehat Analytics

We enable companies to be powered by their data.

To prepare themselves for the information age, companies with legacy technologies, processes and skillsets must undertake a business transformation to become data-led organisations.

Our mission is to help companies undertake that journey.

Our goal is to help companies to structure themselves around an empirical, data-driven business model, where decisions and process are analytical and outcome-focused.

By developing a data-led culture that runs through the entire organisation, we allow our clients to unlock the power of data and thrive in a more competitive, data-led global marketplace.

We drive cultural change: we embed analytical capability deep within the organization, via a structured program of engagement, education, and enablement. All parts of the organization are empowered to use data to do their jobs better.

We deliver technical excellence: we work on the bleeding edge of data science, integrating large, complex data sets with the latest tools, and using the most advanced analytical techniques such as neural networks to allow our customers to see further and react faster.

We are compliance-driven: we know your data must be secure, it must be accounted for, and the quality must be verified. Your businesses reputation relies on your customers trusting you and your processes, so we build in governance, audit, compliance, and quality validation at every stage of the process. We are ISO 9001 and ISO 27001 certified – so we understand the compliance process first-hand.

We are outcome-focused: we believe that data science must be deployed to drive business value. That's why we ensure that we deliver products and services that are usable, scalable and robust, to change business outcomes on the frontline.

whitehat analytics

20-22 Wenlock Road
London
N1 7GU
United Kingdom

+44 (0) 203 603 2773
contact@whitehatanalytics.com
www.whitehatanalytics.co.uk