## Welcome to Strategic Data Skills

#### Strategic Data Skills

**Welcome to the Strategic Data Skills.**

Please watch the video below for an Introduction to Strategic Data Skills

#### Course overview

**The course aims to develop greater skills and knowledge in:**

* ​​Identifying how open and shared data infrastructures can improve decision making
* Applying strategic thinking to new data projects
* Open and shared approaches that make it easier to create and build impactful products and services
* Making better decisions with data

This course is made up of 5 modules, each module should take between 2 and 3 hours to complete.

Below is a list of the upcoming topics we will cover:

1. Building a healthy data ecosystem
2. Organising and validating data
3. Introduction to data analysis
4. Telling stories with data
5. Data ethics

### Obtaining your Certificate of Completion

In order to get the most out of the course, we encourage learners to engage with the content in all modules. The course has been built sequentially, and builds upon the experiences of each module.

To obtain your certificate, learners must complete the summative activities. There is not a minimum grade to achieve, but the following activities must be attempted:

1. Module 2: Hands-on: organising and structuring data
2. Module 3: Analysing the performance of the London Fire Brigade
3. Module 4: Telling the story of the London Fire Brigade

There is no limit on how many times you can undertake an activity. You will receive feedback on each activity, either through pre-written outcomes or via the Tutor's AI assistant who will provide personalised feedback.

Once the completion conditions are met, the Course Completion topic will be shown and provide information for learners to access their completion certificate.

## About the ODI

The ODI is a non-profit company, founded in 2012 by Sir Tim Berners-Lee and Sir Nigel Shadbolt. Since 2017, the ODI’s operating costs of around £6m per annum have been met through a range of grants and commercial revenue.

Many of our products and services are available openly and for free, including our reports, tools and webinars.

### What we believe

At the ODI, our mission is **to work with companies and governments to build an open, trustworthy data ecosystem.**

We want a world where data works for everyone. This means getting data to those who need it, particularly in response to the UN Sustainable Development Goals.

To achieve our mission, we:

* work with organisations to help them use better data practices
* work with sectors and regions to ensure data is helping them
* help interpret and apply the latest thinking around data and data infrastructure.

### What are we?

The ODI operates as an institute, and we also undertake commercial work that advances our mission. We:

* work in policy – influencing government decisions towards the kind of data ecosystem we want to create that will be best for society
* work with businesses – helping them to develop ethical, equitable data infrastructure that will be both good for business and their social mission
* work with philanthropic and other grant-giving organisations – to build a data infrastructure that brings benefits to all
* work with governments, research organisations, public bodies and civil society around the world – to ensure that they can benefit from better data infrastructure.

Our work includes a range of activities including conducting applied research, delivering consultancy services and training, and providing free tools and courses. We have an active and engaged membership community, over 17,000 weekly newsletter subscribers, and hosted more than 1,000 attendees at our recent summit. Since 2012, we have told data stories through our Data as Culture arts programme which includes over 100 artworks.

### The work we do

The ODI aims to enable the development of data infrastructure in ways that benefit people, companies, governments and civil society. We focus on increasing **data flows around the data ecosystem**, improving skills and capabilities, and encouraging innovation. We support data flows, focusing our efforts in three broad areas:

1. **Improving the data practices of organisations so that they can build and manage adequate data infrastructure and data use**. Our work on data literacy and data assurance is part of these efforts.
2. **Tackling challenges so that the data ecosystem works better**. Our data institutions work and our work on data ecosystems and innovation are part of this.
3. **Gathering and creating research, evidence and knowledge** about data and the benefits of open, trustworthy data access, to inform companies and policymakers as they create data infrastructure, assets, practices and policies. This is our evidence and foresight work.

We will continue working with our network and partners to advance our work across these areas.

## Using Moodle

#### Welcome

The software we use for our courses is called Moodle, you are using it now, below is a quick guide video explaining how to use Moodle. If you have any questions or get stuck [get in touch with our support team.](mailto:trianing@theodi.org)

#### Accessibility

Our course platform includes accessibility tools for an optimised learning experience. For easy access, click on the user icon in the top-right corner of the platform page and select 'Accessibility.' Here, you can enable a toolbar to personalise settings, such as font colour, spacing, and size to suit your needs. Adjust these at your convenience for a comfortable learning environment.

#### Summary

In this section we have learned:

1. Introducing your learning environment, Moodle
2. How to navigate in Moodle
3. How to use the quizzes in Moodle
4. How to get your certificate in Moodle
5. How to access the various accessibility tools

## Frequently Asked Questions

### Frequently Asked Questions

What is Moodle and how does it work?

Moodle is a Virtual Learning Environment (VLE) that provides a platform for online learning. It offers a range of tools and features to facilitate course delivery, communication, and assessment. Learners can access course materials, submit their work and track their progress through Moodle.

How do I access my self-paced course on Moodle?

To access your self-paced course on Moodle, you will need to follow

Can I study at my own pace in self-paced courses on Moodle?

Yes, self-paced courses on Moodle are designed to allow learners to study at their own pace. You have the flexibility to progress through the course materials according to your preferred timeline.

Are there any specific requirements for taking self-paced courses on Moodle?

You need a computer or mobile device with internet access and a compatible web browser to access the Moodle platform.

Can I start a self-paced course anytime, or are there specific start dates?

Self-paced courses allow learners to start anytime they wish.

How long do I have access to a self-paced course on Moodle?

You are licensed to access and engage with the content of this course for 1 year from the point at which ODI launches the product to you.

How do I submit assignments or assessments in self-paced courses on Moodle?

Moodle provides a designated area within each course where you can submit assignments or assessments. You can upload files, enter text, or complete online quizzes depending on the specific requirements set. In some cases, ODI learning will have embedded activity types via our Adapt authoring tool. Submission instructions are typically provided within the course materials/activity.

Can I interact with other learners in self-paced courses on Moodle?

Interaction with other learners in self-paced courses on Moodle can vary depending on the course structure. Some courses may include discussion forums or collaborative activities where you can engage with peers. However, the level of interaction may differ. On some self-paced courses, peer-to-peer interaction is not required. Information about the courses can be found at <https://learning.theodi.org/>

How can I track my progress in a self-paced course on Moodle?

Moodle offers progress tracking features to help you monitor your progress in self-paced courses. You can typically view completed activities, track grades, and check your overall progress through the course dashboard or gradebook.

Are there any support services available for learners in self-paced courses on Moodle?

ODI Learning is always on hand to support you with your learning experience. If your query cannot be answered via the FAQ page, please reach out to training@theodi.org and we will be glad to assist you.

Can I receive a certificate upon completion of a self-paced course on Moodle?

The availability of certificates of completion on our self-paced courses on Moodle are triggered based upon the completion of required activities in your course. The required completion criteria will be explained in the introduction/welcome module of your course.

Are there any additional resources or external tools available for self-paced courses on Moodle?

Additional resources or external tools to support your learning are referenced throughout the course.

Can I retake formative and summative activities?

Absolutely, we encourage learners to reattempt activities as many times as they wish to reinforce learning.

How can I provide feedback or suggestions for improvement in self-paced courses on Moodle?

You can typically provide feedback through the Moodle platform via survey questions we have integrated into your course. Sharing your thoughts can contribute to the enhancement of future course offerings, and we encourage feedback. Should you wish to offer feedback outside this process, please email us at training@theodi.org.

I am visually impaired, how can I adjust the course interface for better accessibility?

You can find accessibility options by clicking on your user icon at the top right corner of the screen and selecting 'Accessibility.' Here, you can customize settings like font color, spacing, and size.

## Support

#### Siobhan Donegan - Course Coordinator and Support

Siobhan is the Learning Coordinator for ODI Learning. She runs the day-to-day platform administration for all the courses and is the main contact for students. Before joining ODI Learning, she spent 4 years at ODI HQ in the Business Support team.

If you have any issues, please contact Siobhan at [training@theodi.org](mailto:training@theodi.org)

## Data we collect about you

We collect data about you when you sign up to complete an ODI course. We do this so we can contact you with details and to help us manage and run the course.

You have rights over what is done with data that is about you. You can contact us at [datacontroller@theodi.org](mailto:datacontroller@theodi.org) if you want to get a copy of that data. You can also ask us to update it, delete it or stop using it.

The data controller for this website is Open Data Institute of 5th Floor, Kings Place, 90 York Way, London N1 9AG.

Please send any questions about this privacy policy to [datacontroller@theodi.org](mailto:datacontroller@theodi.org).

## Module overview

#### Building a healthy data ecosystem

In order to unlock the full potential of data we must understand the data ecosystem and how to support and build a strong and supported data infrastructure.

#### Learning Outcomes

1. The aim of this module is to introduce data, why it exists on a spectrum, and how to build a strong data infrastructure to support unlocking the full value from data.
   * Define what data is, explore how it exists on a spectrum and explain how it used to construct insights and evidence for new concepts and ideas
   * Explore the history of data journalism and storytelling, and contextualise data storytelling and how you use it in your professional practice
   * Examine the concept of ‘data is infrastructure’ and how this analogy enables you to understand the need for positive and sustainable investment in open data practices through the use of real-world examples
   * Explore a number of measures of success for building a healthy data ecosystem and some of the barriers that could prevent the full value of data being realised.
   * Evaluate your own practices and how data is used in your organisation

#### What is data?

Data is a material. On its own it doesn’t mean anything, but when we combine it with our prior experiences and prior knowledge, we are able to construct insights. These insights provide evidence for new hypotheses. From there, we make decisions on how to act on data.

**In this section we will explore the following:**

* What is data, information, knowledge and wisdom?
* How do they differ from one another?
* Why does knowing this difference ensure we can use data appropriately to create value?

#### The data spectrum

Data exists on a spectrum, from closed, to shared, to open. In this section, we explore these three types of data, how to go about classifying data and the role of a balancing test in making the best decision.

During this section consider what data you use in your professional practice: is it closed, shared or open? Are there benefits to increasing the openness of the data you use, or do the risks outweigh the benefits? Would it be inappropriate to share the data beyond the named users?

**In this section, we explore:**

* open, shared and closed data
* the ODI Data Spectrum
* why we need open, shared and closed data
* how to classify your data performing a balancing test

#### Introduction to data storytelling

In order to make data more understandable, we need to be able to present findings and insight from data. To further drive positive change, we need to be able to clearly communicate to senior level decision-makers. A key part of this is being able to tell powerful stories.

No field is more experienced at finding and telling stories than journalism, and likewise no field is better at using data than data science. Here, we look at what both fields can learn from each other, and what you can learn from them, in order to find and tell compelling stories with data.

**In this section we will explore the following:**

* How data has changed the way we find and tell stories
* The history of data journalism and storytelling
* How you use stories in your professional practice to communicate and generate buy-in

#### Data is infrastructure

To truly harness the potential of data, we should regard it as a form of infrastructure. It holds equal importance to our road, electricity, and water networks, serving as a fundamental resource from which we can all benefit.

Data infrastructure is made up of data assets, standards, technologies, policies and the organisations that steward and contribute to them. At the ODI, we use the analogy that 'Data is infrastructure' - by that, we mean it is not a commodity, like oil. Instead, it is a necessary framework that facilitates the creation and delivery of services and products across all sectors and industries. In order for the value of data to be realised, it needs resources, investment and management.

**This section will introduce:**

* The concept of ‘data is infrastructure'
* What is open data?
* What makes data open and why do we need open data?

#### Measuring success

In this module we have explored what data is, and what value it has. We’ve examined the history and application of data journalism through the use of storytelling using Florence Nightingale and John Snow as historical examples to demonstrate how effectively sharing data and insights helps individuals and organisations to make better, more informed decisions.

One way of maximising the value of data is by opening it up and investing in data as infrastructure. In doing so, we must understand the limitations of what we can and can’t share, but also the most appropriate ways to share it. We can use the Data Spectrum as a guide to understand what constitutes Open Data, and the Theory of Change to explore the risks associated within our professional practice.

**In this section, you will explore:**

* How we measure the success of data and open data
* Your own practices and how data is used in your organisation

#### Barriers to success

At the ODI, we have a theory of change. This theory promotes collaboration and trust between organisations, governments and individuals. We want those who steward data and those who create information from this data to act in ways that lead to the best social and economic outcomes for everyone. Our theory of change explores the balance of trust and openness to create positive impact and the dystopias that could become barriers if not addressed.

As you are developing your understanding of the theory of change, consider what data you create and steward. Think about how you build trust within your professional practice, and how open you and your organisation are. How does this effect the positive outcomes you deliver?

This section explores:

* The ODI’s Theory of Change
* How value is created from data
* What happens when we hoard data
* What happens when we are fearful of data

## Reflecting on your practice

### Reflecting on Module 1: Building a healthy data ecosystem

To support your learning we recommend taking the opportunity to reflect on the module and examine how the lessons covered can be applied to your own professional practice.

**To aid the process, we've provided a SDS Module 1 Reflective Template (.docx).**

The purpose of this document is to give you an opportunity to explore how your learning can apply to your professional practice. It is not an assessed piece of work. Instead, it is intended to provide a discussion framework for you and your line manager, mentor or colleagues.

We recommend that the reflections you make during this course are used as part of a portfolio of learning and development, used in conjunction with your own professional development frameworks.

## Module summary

### Summary: Building a healthy data ecosystem

In order to unlock the full potential of data we must understand the data ecosystem and how to support a build a strong and supported data infrastructure. Open data plays a key role in this ecosystem. Open data maximises innovation and opportunities for unexpected benefit  data that anyone can access, use and share. Open data is a raw material for the digital age but, unlike coal, timber or diamonds, it can be used by anyone and everyone at the same time. Any restrictions imposed on the use of open data will limit its potential for creating new value. Open data can bring diverse benefits to governments, businesses and individuals. It has the power to help improve services, grow economies and protect our planet.

**In this module we explored:**

* Data, what it is, why it exists on a spectrum and how it is used to construct insights and evidence for new concepts and ideas
* Data journalism and storytelling, and contextualise data storytelling and how you use it in your professional practice ‘
* Data as infrastructure and how this analogy enables you to understand the need for positive and sustainable investment in data
* Measures and barriers to success and a theory of change for data
* Finally, you have reflected on your own professional practice and how you tell stories and create value

There are many places we could go from here, a strong data infrastructure needs to be supported with:

* A clear strategy, governance and ethical framework
* Experts who maintain the infrastructure for the benefit of everyone
* A community of users and innovators who build on top of the data infrastructure

The ODI Data Skills Framework shows the broad range of skills required to get the most from data.

In this course we will be exploring many areas of the framework, including:

* How to organise, structure and clean data
* How to start analysing data to find insight
* How to visualise and tell a story from data
* What the future of data platforms looks like
* The role of data ethics to minimise harmful impacts

In ‘Module 2: Organising and Validating Data’ you will apply data skills to clean data and increase its usability.

## Module overview

In Module 1, you explored what data is, and how it can be used to create value. Value is recognised by stakeholders when we communicate the key insights and findings in the data to ensure we maximise that value, we need to ensure that data is properly gathered, organised, structured and cleaned.

Module 2: 'Organising and Validating Data' introduces the skills to perform these tasks.

#### Learning outcomes

By the end of this section you will understand how to organise and clean data, as well as undertake activities to quality check and enable greater utility and interoperability of data.

**To achieve this, you will:**

* Investigate the gathering of data, including various sources of data and the practices employed in finding the data
* Examine the structure of data and determine approaches to organise the data into the appropriate structure as well as apply cleaning techniques to increase the utility of data
* Understand the purpose of data schemas and how to create and use them

## The rising value of data

#### Data: Gathering, organising, structuring and cleaning

The value of data increases as it turns into a story. This is why data science and journalism skills must be used cooperatively. Those who work with the data must be able to communicate why it is important or useful, otherwise the value is lost.

Creating value from data is a sequence. This sequence repeats in a cycle and follows the order: **Gathering, Organising, Structuring and Cleaning.**

**The following video explains why this is important.**

In the next section we introduce the skills and knowledge involved in gathering data.

#### Gathering data

Data plays a key role in storytelling, however not all data is easy to find. The growing demand for data has led to an increase in human friendly data services, including data portals and simple download buttons. However, downloadable data represents only a small fraction of the available data on the Web. The majority of data available on the Web is hidden from the human eye. But machines can find and read this data.

**In this section we explore the following:**

* The difference between downloadable and hidden data
* Finding downloadable data
* Finding hidden data
* The benefits of using hidden data

#### Organising and Structuring data

Once we have gathered data, we need to ensure it is usable. A common challenge in ensuring data quality is difficulties people face when using spreadsheets. When data is properly managed, it is much easier to answer fundamental questions and conduct analysis. Correctly managed data also ensures you aren't making decisions based on faulty evidence. Knowing how to structure and organise data in a spreadsheet is fundamental to ensuring consistency in your data. In this section we look at how to effectively structure a spreadsheet for raw data collection.

**We will cover:**

* Spreadsheet layouts
* Column titles
* Header rows and the freeze function
* Data types

At the end of this section you can try out what you've learnt on a spreadsheet.

#### Choosing the right structure and format for data

How should data be structured? Who needs to consume the data? Humans, machines or both. So, how should data be structured? Choosing the right format helps ensure the data can be simply managed and reused. To maximise reuse of data, it may be necessary to use a number of structures and formats available across different platforms to suit users' needs.

**In this section we'll explore:**

* The difference between machine and human readable data
* How to create machine readable tabular data
* Other data structures and formats
* How to choose the right structure and format

#### Cleaning data

One of the biggest challenges when working with any data is dealing with errors. Often errors are not even noticed by data publishers because the data can change over many years. In other cases, errors can be the result of human mistakes in data entry, like mistyping or incorrect abbreviations. When working with any data, it is important to know how to find errors and correct them to make the data more useful.

**In this section we'll explore the following:**

* Common data errors
* Useful data cleaning tools
* Reasons for cleaning data

## [Optional] Getting started with Open Refine

#### Open Refine

Whenever you want to carry out analysis, you should always take the time to inspect the data you have been given to see if there are any errors like misspelling, repeated entries, mixed numerical scales and mixed ranges, as well as data that might be missing. Otherwise you run the risk of transferring these mistakes into your analysis.

Open Refine is a powerful desktop app used for re-organising and cleaning data.

#### Watch the video below to get an introduction to Open Refine.

### 

### Installing Open Refine

To get started with this exercise you will require a copy of Open Refine which is available from [openrefine.org](http://openrefine.org/download.html).

For Windows, Refine comes as a pre-built .EXE file that you simply run and does not require installing.

Once running, Refine will create a web server on your local machine that you interact with via the URL <http://127.0.0.1:3333>.

### OpenRefine Community Support

If you require support in using OpenRefine, there is an active community available through their [website forum](https://forum.openrefine.org/).

OpenRefine also maintain a comprehensive [user manual](https://openrefine.org/docs).

#### Applying your skills and knowledge

Throughout this module you have been building your skills in the gathering, organising, structuring and cleaning of data. These skills are transferable across all different types of data, including the data you use at the Wildlife Trust.

The following assignment has been designed to allow you to apply your skills in organising, structuring and cleaning a dataset.

**The assignment requires you to:**

* Reorganise and restructure a messy dataset into a single dataset that can be analysed simply by both human and machine.
* Create a single CSV file (you can export this from tools like Excel) that contains just the dataset.
* The dataset should contain enough data to enable analysis of the data by country, type, species and ownership.

**Where is the data?**

You can download the dataset [here](@@PLUGINFILE@@/WLT%20exercise%20dataset.xlsx).

**What tools can I use?**

You can make use of any tools you are comfortable with, however the submission must be a CSV file.

**How will I receive feedback?**

Once you have completed your assignment you should upload the dataset via the submission portal below. Providing that you submit a valid CSV file, our AI assistant tutor will attempt to assess your submission and give customised feedback.

The AI assistant tutor uses AI that we have specifically trained to deal with this task. It has been given an example a model answer and some other marking criteria. The assistant tutor will look to mark against four criteria:

1. How the data is structured
2. The consistency of the data in each row/column
3. The cleanliness of the data
4. The ability for a machine to reuse the data

**IMPORTANT**: We do not expect the assistant tutor to be 100% reliable and accurate. We have applied it here in order to examine the ability for AI to give instant feedback rather than wait for a human tutor to be available. We believe there is some value in doing this, but also recognise the need for a human to be 'in-the-loop'. For this reason we are building in checks and processes that allow a human tutor to override the AI assistant tutor. We are also building in ways for you to report your feedback to us so it can be flagged and we can use the learning to improve (or remove) the assistant tutor. Think of them as being on probation.

#### Submit your answer via the link below.

#### Choosing and designing schemas

Data schemas are structured definitions or models that specify the structure, format, and constraints of data within a particular system or database. They define the organisation and expected characteristics of the data, including the data types, relationships, and rules that govern the data. Data schemas provide a blueprint for how data should be structured and stored, ensuring consistency and integrity.

While they provide similar features to data validation in Excel. Data validation rules in Excel are primarily applied to individual cells or ranges and focus on immediate data validation within the spreadsheet. However, they do not provide a comprehensive and standardized structure for an entire dataset or database.  
  
**In this section we explore:**

* What is a schema?
* Why do schemas matter?
* Schemas in practice
* Choosing or designing schemas

#### Applying your skills and knowledge

So far in this module we have explored the creation of well structured, machine readable data that conforms to a number of schemas. In the last part of the module, we have put together a tool that allows you to explore how to create schemas for yourself to validate any data. You can either pick the dataset from the previous exercise, or a dataset of your own which you can create a schema for and find out if you can validate the data.

In order to give you something to take away, we will also provide you with the JSON schema you can use with other tools.

**The assignment requires you to:**

* Identify a dataset you want to create a schema for
* Create a schema that can be used to verify your data
* Download the schema for sharing with others

**Where is the data?**

For this exercise you can use your own data, however be aware that the tool is limited to accepting data in a single CSV file.

**What tools can I use?**

You can make use of any tools you are comfortable with, however the submission must be a CSV file.

**How will I receive feedback?**

Once you have completed your assignment you should upload the dataset via the submission portal below. Providing that you submit a valid CSV file and have created a good schema, it should validate and pass the activity.

There is an easy way to pass this exercise. Create a schema that accepts any input. For this reason our AI assistant tutor will attempt to assess your submission and give customised feedback on how the schema could be improved.

The AI assistant tutor uses the data you upload as well as the schema to provide feedback and potentially a new schema.

#### Submit your answer via the link below.

Launch submission portal

If you are viewing this page using the Moodle app, click here to launch the e-learning

CLOSE AND RETURN TO COURSE

Loading...

## Reflecting on your practice

### Reflecting on Module 2: Organising and structuring data

To support your learning we recommend taking the opportunity to reflect on the module and examine how the lessons covered can be applied to your own professional practice.

**To aid the process, we've provided a SDS Module 2 Reflective Template (.docx).**

The purpose of this document is to give you an opportunity to explore how your learning can apply to your professional practice. It is not an assessed piece of work. Instead, it is intended to provide a discussion framework for you and your line manager, mentor or colleagues.

We recommend that the reflections you make during this course are used as part of a portfolio of learning and development, used in conjunction with your own professional development frameworks.

## Module summary

#### Summary: organising and validating data

This module has introduced the skills required to ensure we maximise the value of data through the proper gathering, organising, structuring and cleaning. This is step one of the four step process of storytelling with data, and is likely to be a step where you spend much of your time.

By spending the appropriate amount of time organising and validating our data, we can ensure that the outputs that our stakeholders receive is as effective and valuable as possible.

In this module you have undertaken activities that quality check and enable greater utility and interoperability of data. This has included you:

* Examining the structure and organising the data into the appropriate structure as well as apply cleaning techniques to increase the utility of data
* Building an understanding of  the purpose of data schemas and how to create and use them, further supporting your approaches to organising data effectively
* Applying these skills to different datasets and contexts to explore the transferable nature of these practices across the different datasets you use in your professional practice

In module 3 we introduce data analysis and explore the skills applied in establishing the shape of data, predicting trends and appropriately filtering data in order to start identifying key insights and building your data story.

## Module overview

**"There are three kinds of lies: lies, damned lies, and statistics."**

Benjamin Disraeli

Having spent a lot of time gathering and preparing data, the temptation is to immediately begin to analyse the data using any technique that seems relevant. First, however it is important to explore the data to discover its shape and any limitations it might have. We also need to understand the benefits and risks of applying each statistical technique to data.

So before we perform any analysis, this module looks at three key things:

1. How numbers and summary statistics can lead us to incorrect conclusions in "Seeing through a world of numbers"
2. Why establishing "the shape of data" is critical to discover its limitations and biases.
3. Why plotting "Trends in data" is more challenging than taking a ruler to a plot.

Following this we'll then look at how some key statistical techniques and tools can be applied to data.

In this module we'll be analysing the performance of the London Fire Brigade and look at the impact on the service of closing 10 fire stations in London in 2014 in order to cut back on spending. Was it a good idea? What has happened since? Is the performance good or bad? Is it fair? What else should happen now?

Module 3 introduces the skills to perform these tasks in order to see what can be discovered from the data.

#### Learning Outcomes

By the end of this section you will have the skills to evaluate data from different sources in order to establish its quality, and carry out exploratory data analysis to produce insights.

**To achieve this, you will:**

* Evaluate the benefits and risks of applying different statistical techniques to data
* Apply several exploratory data analysis techniques to create summary statistics and visual representations
* Practise the skills to establish the shape of data and examine what it means, and how it can be used to determine the state of data
* Use tools to filter and explore trends in data
* Evaluate the difference between qualitative and quantitative data analysis

## The rising value to the public

#### Refining, observing, analysing and calculating

Once we have collected data that is relevant to our projects, we need to then employ several skills to filter that data and make it usable for our purpose and objective.

The steps to determine the state and quality of the data are:

1. Refine,
2. Observe,
3. Analyse, and
4. Calculate.

The following video explains why this is important.

In the next section we introduce the power of numbers, and the skills required to see through the issues and establish accurate and appropriate insights.

#### Seeing through the numbers

Public spending, health risks, environmental disasters, who is rich, who is poor, pensions, the best and worst schools and hospitals, immigration – life comes in numbers. The trick is seeing through them.

To some, numbers can be intimidating. They can also create illusions and lead us to see what is not actually there. If numbers are not understood then how can we possibly understand what conclusions we can draw from them and the impact this has on people!

#### The shape of data

Data, it turns out, has shape. That shape has meaning.

The shape of data tells you everything you need to know about your data from its obvious features to its deepest secrets.

Establishing the shape and meaning of data prior to using it in analysis is essential to avoid basic mistakes that can lead to harmful impacts.

#### Trends in data

One of the key benefits of data is that of finding insight, spotting patterns and being able to make predictions. However, there are risks with mixing descriptive statistics, used for summarising existing data, and inferential statistics which focuses on drawing conclusions, making predictions, or generalising from a sample to a larger population.

This module explores:

* The difference between descriptive and inferential statistics
* Inferential statistics applied to data
* How to avoid overfitting
* Working with confidence

#### Filtering and pivot tables

At this point, we have explored the organising and cleaning of data, and we've looked at exploratory data analysis - what it is and how it helps us to start to analyse data. Through this initial analysis we are also identifying which parts of the data are most useful, and which are not useful at all. We can now start to filter our data and set aside anything that is out of scope for our current objectives. Through filtering the data we can start to identify trends in the data.

In these exercises we make use of core functionality that is available in most spreadsheet applications.

#### Analysing the performance of the London Fire Brigade

In this module you will be focusing on data analysis through the lens of the London Fire Brigade's operations and the impact of the 2014 station closures. You will be guided through real-world data, enhancing your analytical skills and your ability to interpret and present complex information.

###### **Backstory**

In 2013, it was announced that cuts to the London Fire Brigade meant that ten fire stations in the capital had to be closed.   
These closures happened a year later in 2014, but what actually happened as a result of the closures? What more could be done to improve the efficiency and overall performance of the London Fire Brigade? Where is the data?

This time we will not provide you with a dataset to begin with; one of the key components of analysing data is combining domain expertise with statistical knowledge.

You should be able to find several relevant datasets online with a simple search, and you can use the following questions to support you in deciding which data to use:

* What data is available?
* How usable is the data?
* Which fire stations closed?
* When did the fire stations close?
* What was the performance before these closed?
* Are there any data collection techniques or targets that influence the data?

There may be many more questions that you might need to answer to inform your analysis.

In addition, the [Open Data Institute built a tool](http://london-fire.labs.theodi.org/map.html) that predicted the impact of closing the proposed fire stations, which you may find useful to explore.

As you work through this exercise, think about the different contexts in which you might apply these skills. The analysis undertaken in this section is transferable and is applicable across all different types of datasets.

#### Important

**Before you get started, please download this dataset.**

## Analysing charitable donations to the Wildlife Trust

#### Applying your skills and knowledge

Now you have the opportunity to explore a dataset that is more recognisable. In this exercise, we provide you with data relating to one-off donations made to Trusts in South Yorkshire and the East Midlands.

This is an opportunity to apply what you've learnt about exploratory data analysis. During this process you make observations and generate hypotheses. You are seeking to identify patterns and spot anomalies, and ultimately build evidence to support your hypothesis.

In this exercise you will use the following dataset: [Analysing Charitable Donations](@@PLUGINFILE@@/Analysing%20Charitable%20Donations%20%282%29.xlsx).xlsx

**Note:** the dataset is synthetic and does not refer to real people or detail any real donations made.

The dataset contains data on donations made to Wildlife Trusts in South Yorkshire and the East Midlands, all made during August 2022. Work with this data and try to identify the following:

* The trust with the highest number of donations
* The median donations of all trusts
* Which trust receives donations of the highest value
* The modal value for each trust

## Introduction to qualitative data analysis

### Introduction to Qualitative Data Analysis

Not all data comes in the form of structured tables or accurately located geographic data. Often it can be qualitative data that is the hardest to analyse and work with. Qualitative data is information that cannot be measured and is very subjective as a result. Even colour can be subjective.

The aim of qualitative data analysis is to reduce and make sense of vast amounts of information, often from different sources, and to offer an explanation, interpretation or thematic summary of the data. Inputs to qualitative analysis can take many forms including interview transcripts, documents, blogs, surveys, pictures or videos.

Qualitative data analysis is a more natural process for humans who naturally seek to distil inputs into themes and key outcomes, as is especially true of meetings or focus groups. People will often use mind-mapping or post-it based thought maps to help group together and categorise wide-ranging discussion into key themes.

Qualitative data analysis ought to pay attention to the spoken word’s context, consistency and contradictions of views, frequency and intensity of comments, their specificity as well as emerging themes and trends.

There are two main ways of analysing qualitative data, **framework analysis** and **thematic network analysis**.

### Framework Analysis

Framework analysis involves building a predefined set of criteria that clearly reflect your aims, objectives and interests. Using this set of criteria, the relevant pieces of information can be extracted from the data and compared with other inputs in the framework. Using a framework allows many researchers to do the extraction while minimising the chances of qualitative analysis bias. A bias can still be introduced at framework design stage which may result in key information being missed.

An alternative approach is to not construct a framework, but rather to apply thematic network analysis. Thematic network analysis is a more exploratory approach which encourages the analysis of all of the input data which can shape the output in unexpected directions.

In reality, the majority of qualitative data analysis will involve a combination of the two approaches.

Whichever approach is chosen, the first step in any qualitative data analysis involves familiarisation with the data; reading and re-reading responses. At the same time, it is a good idea to start codifying the data by writing down keywords and topics which attempt to reduce and interpret the data. The result of the coding process could be the thematic network analysis or a framework by which all responses need to be coded. Either way, coding can be a long, slow and repetitive process, however there are a number of tools that can help with thematic network analysis.

### Thematic Network Analysis

Entity recognition tools provide one such technique that can help analyse and enrich qualitative data. Essentially, entity recognition seeks to locate and classify named entities in text into predefined categories such as the names of persons, organisations, locations, expressions of times, quantities, monetary values, percentages, etc.

Calais (shown below) tagging of BBC News: NI farms 'can expect same support' post-Brexit - Gove

Refinitiv, provided by London Stock Exchange Group (formerly known as Calais from Thomson Reuters), is one such example of an entity recognition engine. More than just entity recognition, the Calais service links those entities to data records that exist within the Thomson Reuters permID database of entities and financial records.

More widely, such entity recognition is used by services such as [TheyWorkForYou](https://www.theyworkforyou.com/) in order to track the activities of politicians and provide this in an easy way to the public.

Calais is a good example of an online tool that can perform entity recognition and provide links into additional data on each topic. Other techniques can be much simpler but just as effective, such as word cloud generators.

## Reflecting on your practice

### Reflecting on Module 3: Introduction to data analysis

To support your learning we recommend taking the opportunity to reflect on the module and examine how the lessons covered can be applied to your own professional practice.

**To aid the process, we've provided a SDS Module 3 Reflective Template (.docx).**

The purpose of this document is to give you an opportunity to explore how your learning can apply to your professional practice. It is not an assessed piece of work. Instead, it is intended to provide a discussion framework for you and your line manager, mentor or colleagues.

We recommend that the reflections you make during this course are used as part of a portfolio of learning and development, used in conjunction with your own professional development frameworks.

## Module summary

#### Improving performance of data

The exploration of data is an important step in maximising the utility of our data. Once gathered, we filter the data to ensure it is appropriate for our purposes. We may remove parts of the data set that are superfluous to our needs, or we may fix data quality issues. But in order to do this, we must explore the data to find out more.

The exploration of data involves the practice of applying data analysis techniques and the creation of summary statistics and visual representations. Through this practice we can establish the shape of data. The shape of data tells us everything we need to know about data from the obvious to the hidden.

Through this module you should now be confident in performing exploratory data analysis activities, such as:

* Creation of summary statistics and visual representations
* Establishing the shape of data
* Using tools to filter data and explore trends

In module 4 we continue building on the skills that improve the performance of data and use our findings from exploratory data analysis to start creating effective visualisations.

## Module overview

#### Module 4 overview

The course so far has focused on working with data directly to clean, structure and analyse the data. In this module, we change our focus to data visualisation.

We present data to our stakeholders by telling stories. To accompany the story, we produce visualisations that support the explanation of the scale and impact of our data.

#### Learning outcomes

By the end of this section you will understand how to produce and present appropriate data visualisations that deliver a connected and relevant narrative.

To achieve this, you will:

* Explore data visualisation best practice and understand the principles behind effective data visualisations
* Examine where visualisations can be used to misrepresent or deceive and the impact this has on our outcomes
* Evaluate complex stories and their impact
* Create a narrative that tells the story of the data

## Rising value to the public

### Data Visualisation

When effectively analysed and presented in a clear and compelling way, data has the potential to create impact. Whether that’s changing perceptions, offering counterintuitive insights or prompting action, impact happens when data acts as the catalyst for change. As we work through the stages of rising value, we gather our data and clean it, following that with filtering and analysis.

As per the Pareto Principle (the 80:20 rule), this is where we will have spent much of our time.   
  
  
  
But, at the heart of driving change is the skill of finding and telling stories using, where relevant, compelling visualisations. A modern data scientist will work with many different types of visualisation, from common graphs to advanced statistical plots to geographic representations. It is the elements of visualisation and storytelling that our stakeholders care about, because it’s the elements they will see.

#### Communicating outputs

There’s something breathtaking about data that is communicated well — it’s a lot like marvelling at an architectural masterpiece. Think about how you felt last time you looked at an engineering or architectural masterpiece, how did it make you feel? Did you feel elated? Motivated? Educated?

Data communicated well can do exactly the same. Alternatively, when not done well you end up suffering a long presentation of poorly presented, cluttered and misleading charts that don’t seem to be relevant to your problem.

Find out more in this eLearning module about how to start creating an inspiring visual story from data.

#### Data visualisation formats

Visualisations have the power to distil complex data sets into clear and concise stories. Used properly, they have the ability to condense a huge amount of information into a compact and easily digestible form.

The human brain is adept at recognizing and interpreting visual patterns. Stimulating our visual senses and leveraging their natural inclination towards visual stimuli, data visualisations are far more likely to capture attention and be remembered compared to textual or numerical data. By carefully crafting the design and layout, visualisations can guide viewers through a narrative, presenting powerful stories that are engaging and memorable.

In this module we are going to cover:

* The data visualisation toolkit
* Best practices for creating impact with different types of data visulisation
* Using infographics
* Adding interactivity
* Selecting the right visualisation format

#### Fundamentals of visual perception

Data Visualisations come in all shapes and sizes, they can be simple and they can be incredibly complex, but they should always communicate the message in an understandable way. In this section we introduce the underlying science of how the visual cortex works and how you can use that knowledge to create effective visualisations.

This forms the basis of data visualisation best practices. These can then be applied to maximise the communication of your key messages and insights.

#### Visual deception

The use of visual deception can create intrigue, but it can also be used to mislead. Here, we examine examples of visual deception and the complexities they create when communicating a message. As you learn, consider the data practices we explored in modules 1, 2 and 3, and how they might contribute to misleading information.

A modern data scientist should be able to choose deception techniques that engage and inform, and avoid those which intentionally mislead.

#### Turning a visualisation into a story

Data visualisations are a powerful way of bringing data to life. They help convey messages quickly and succinctly. However on their own they can lack context and the risks of misinterpretation are high.

To add context to the visualisation, try to build a strong story with clear headline and narrative.

This module goes through the steps you should take in putting your story into words for the best possible impact

#### Telling the story of the London Fire Brigade

In the last module we asked you to analyse some data published by the London Fire Brigade, covering a short period before and after several stations were closed in 2014. For the next exercise we are revisiting that topic in the context of data storytelling; how would you build an engaging narrative using this data? Think about the concepts we have covered in this module, and see if you can apply those principles to select the most appropriate options from the choices below

## Reflecting on your practice

### Reflecting on Module 4: Telling stories with data

To support your learning we recommend taking the opportunity to reflect on the module and examine how the lessons covered can be applied to your own professional practice.

**To aid the process, we've provided a SDS Module 4 Reflective Template (.docx).**

The purpose of this document is to give you an opportunity to explore how your learning can apply to your professional practice. It is not an assessed piece of work. Instead, it is intended to provide a discussion framework for you and your line manager, mentor or colleagues.

We recommend that the reflections you make during this course are used as part of a portfolio of learning and development, used in conjunction with your own professional development frameworks.

## Module summary

#### Telling stories with data

Through this module you should now be confident in working with your data to create effective and appropriate data visualisations by:

* Understanding the principles behind data visualisation best practices
* Examine data visualisations and determine whether or not they are deceiving or misleading
* Evaluate complex stories and their impact
* Create a narrative that tells the story of data

Throughout these modules we have been looking at how to use data to help inform decision making. But who benefits from the decisions being made? How do we ensure that we are not just compliant but also ethical? The next module introduces the critical area of data ethics and how to start operationalising it within projects.

## Module overview

#### Module 5 overview

All decisions made have an impact on someone. Increased responsibility is being placed on data and algorithms that process it to automate, speed up, or make decisions at scale. In everything from provision of healthcare, allocation of benefits and credit to predicting crime and even deciding how to optimise the performance of emergency services. Data informed decision making is everywhere. But are the benefits fair? Are people protected from the harmful impacts? Who decides?

At the Open Data Institute (ODI), we advocate for and support practices that increase trust and trustworthiness: building ethical considerations into how data is collected, managed and used; ensuring equity around who accesses, uses and benefits from data; engaging widely with affected people and organisations. In using data, the full extent of the consequences have not always been explored.

Find out more in the video below.

In this section we will be exploring data ethics and the tools available to help everyone ensure that the monitoring of ethics is more akin to compliance monitoring as opposed to a review that gets done at the start of the project and forgotten. Ensuring ethics is at the centre of a project helps deliver successful outcomes people trust (even if they don't agree with them).

#### Learning Outcomes

By the end of this section, you will gain the skills to evaluate data's impact, identify data ethics concerns, and generate recommendations to mitigate negative effects.

**To achieve this, you will:**

* Explore the concept of data ethics
* Examine examples of unethical data practices
* Apply the consequence scanning toolkit to evaluate the impact of real-world data ethics case studies

## What is data ethics?

### What is Data Ethics?

“Data ethics is a branch of ethics that evaluates data practices with the potential to adversely impact on people and society – in data collection, sharing and use” The Open Data Institute

When we consider the broader philosophy of ethics, we consider what is right and what is wrong. The concept of right and wrong are personal to the individual and are informed by one's life experiences, cultural values and the systems in which they have grown up.

That being said, the practice of data ethics is not to be confused with ethics more broadly. Data ethics is a branch of ethics and considers data practices. In short, in our professional practice we undertake actions, build tools and provide services that have an impact.

By considering data ethics as a professional practice we explore the impact of these outputs and consider the intended consequences and the unintended consequences.

### Understanding the impact of products and services

Research undertaken by Doteveryone (Miller & Coldicutt, 2020) provided evidence that tech workers believe in the power of their products to drive positive change. However, the level of positive change is dependent on their ability to raise concerns, work with broad ranges of stakeholders with different expertise and ultimately understand the possible outcomes of their work - both positive and negative.

A key finding from the research published by Doteveryone found that:

"More than a quarter (28%) of tech workers in the UK have seen decisions made about a technology that they felt could have negative consequences for people or society. Nearly one in five (18%) of those went on to leave their companies as a result."

Within this research, **DotEveryone found that 78% of workers want practical tools** to support them in exploring the impact of their products.

### The data lifecycle

Depending on your objective, you will be using data at any point, and possibly multiple points across the data lifecycle. This may include any of the following:

* Creation, curation or acquisition of data
* Storage of data (including retention)
* Organising and structuring of data
* Algorithmic model building
* Algorithmic model deployment
* Sharing or selling of data, and finally
* The disposal of data.

Each stage in the data lifecycle has different ethical considerations that affect people and need to be considered. The process of examining our impact is not always clear, or straightforward.

### Data Ethics in Practice (... or not)

Unfortunately, there is no shortage of examples of unethical data practices. Let’s explore some:

#### Auto Cropping images on Twitter

Back in January of 2018, Twitter introduced an auto-cropping AI that detects the most interesting part of your image and crops the ‘preview’ photo to match.

Watch the video to learn more about the effects of Twitter’s auto cropping algorithm.

#### Diagnosing Skin Cancer

With skin cancer. Like most diseases, early diagnosis is a key contributor to positive health outcomes. In the case of a Melanoma diagnosis, an individual is 100% likely to survive over 1 year if it is diagnosed at the earliest stage. This statistic decreases to 53% when diagnosis occurs at the latest stage. Individuals with non-white skin, whilst less likely to develop skin cancer, have been found to have lower 5-year survival rates (70%) than white individuals (92%).

Watch the video to learn more about the effects of bias in training data sets.

### Why is data ethics important?

Data has the potential to highlight the bias and assumptions that exist across society. Within products and services, these ‘data-driven’ biases can sometimes create outcomes that are incorrect, exclusionary or outright harmful.

Understanding the potential for both positive and negative consequences is integral to the delivery of effective and valuable products and services. However, we must also recognise that important skills of the future are not only technical, but rely on the skills of a broad range of stakeholders that can translate data into information and ultimately into knowledge.

The value of data-driven decision making may not always be the panacea it’s intended to be. Understanding the implications of bias at any point during the data lifecycle, whether that is at the point of collection, analysis, model development and deployment, as well as at the point of disposal is imperative to ensuring a human-centred outcome.

Exploring the data ethics in and around the vast technological developments is vital to ensuring that contributions to the future of AI and tech-enabled solutions are both effective and equitable.

## The Consequence Scanning Toolkit

#### There is never a lack of data, or technology

There is never a lack of technology or data. Both are often seen as the “golden bullet” to solving a whole host of vaguely defined problems. In fact having too much technology and data can cause more problems, both legal and ethical. Too much data and technology becomes more challenging to manage, more disparate, siloed and not to mention aged and part of the legacy/debt of the organisation. Having the right processes and people in place to support the data and technologies and being focussed on your purpose will help ensure that technologies and data are suitable for the purpose and not there because it is easy to get both.

### The Consequence Scanning Toolkit

Consequence Scanning is an agile practice for responsible innovators. It's a framework of practice to enable teams who work on data enabled projects to explore the impact of their practice by identifying the intended and unintended consequences, and defining the appropriate actions to mitigate negative effects and promote positive impact.

In this section, we will explore the consequence scanning toolkit and how it can be used in the practice of data ethics.

### How to use the Consequence Scanning Toolkit

You start by asking yourself three questions:

These questions help your team to share knowledge and expertise in order to map the potential impact of your work. Most importantly, these questions are designed to be asked to a varied group of stakeholders ensuring that everyone has the opportunity to input into the project.

* **Intended consequences describe your purpose:** What are you trying to achieve? What is the positive goal of the project? Who will be positively impacted?
* **Unintended consequences describe the potential side-effects** of your project and focus on who could be negatively impacted or excluded from your project, impact on wider society, economy or even environment.

#### When should you do Consequence Scanning?

You should carry out consequence scanning during the early stages of product planning and return to your thinking and outputs throughout development and maintenance.

### Striking a balance between positive and negative effects

To assess if a potential unintended consequence is a real risk it is important to evaluate the likelihood and the severity of the impact. A common method of doing this is to complete a risk matrix and assign a risk score.

It is important to remember that while the impact of different risks could be the same, e.g. “Damage to organisations reputation”, the likelihood and severity scores could be significantly different depending on the risk and wider context of the risk within the organisation. It is also likely to vary widely depending on the organisation's attitude towards the risk. For example, is sharing poor quality data, which is likely most of the time, actually high in severity? Especially if you tell people the data is likely to contain errors due to the way it was collected.

### The recommended structure

If you're working in a group and looking to run your own Consequence Scanning workshop, the following structure will support you in ensuring each individual has the opportunity to construct their own ideas as well as collaborate with others.

Phase one:

1. Introduction: explain the focus of the event and post up intended consequences.
2. Quiet time: ask participants the first question. Everyone takes a moment to answer based on their experiences and expertise.
3. Affinity sorting: participants write down ideas, similar ideas are grouped.
4. **Quiet time:** participants have another opportunity to add new consequences based on other posts.

Phase two:

1. **Action sorting:** sort the consequences into action categories (act, influence and monitor).
2. **Dot voting:** Once sorted, ask questions 2 (positive consequences to focus on) and 3 (what are the consequences we want to mitigate). Everyone can vote all at once on the consequences.
3. **Discussion:** What do the consequences you can act on mean for your product? Assign responsibilities.
4. **End:** Start working on implementing the actions.

#### An example: the unintended consequences of publishing crime data

Let’s explore a simple example. What do you think the likelihood and severity is related to the following situation?

* **Action:** Releasing crime data with detail at street level
* **Intended positive impact:** Provide a rich dataset for use in consumer applications, such as travel planning apps that can now find out where high crime areas are and recommend people avoid them, making society safer.
* **Potential unintended consequence:** A significant reduction in house prices in high crime areas
* **Perceived negative impact:** People in high crime areas being unable to sell or rent out their houses

What do you think the likelihood of the unintended consequence is?

View answer

Many countries have published highly detailed crime data. Although many had concerns, there has been no significant impact as a result of the data being published despite the belief that the release of crime data would affect house prices.

In reality, people simply didn’t know what high crime was and it is more about experiences than statistics. There are different things that impact house pricing, and what this perception failed to do was consider the other reasons that drive people to buy houses.

The perceived impact was high, but in actual fact the impact was low.

### Deciding your actions

Once your team has had the opportunity to explore the intended and unintended consequences and rated those highest in risk. You can then group these consequences into the following categories:

* **Act**: The things the core team can take forward and action immediately. Within these, prioritisation and time scales should be applied.
* **Influence**: Consequences that need to be handled by other stakeholders, either within the organisation or externally, but within reachable networks.
* **Monitor**: These are consequences that you cannot control, such as political climate or global trends.

From this point, you can articulate the required roles and responsibilities, allocate the actions and start implementing them against your projects.

### Applying the Consequence Scanning Toolkit

The Consequence Scanning toolkit is designed to be adaptable and fit into how you already work with your team. It has been designed to be:

1. **Lightweight**: minimal materials, preparation or technical expertise required.
2. **Simple to start:** easy to understand what you should do, how and why.
3. **Difficult to master:** critical and creative thinking is required. Participants must think openly and speak candidly about the potential harms.

Although it may be difficult at first, consequence scanning becomes easier the more you embed it into your practices. In the next section we'll introduce some real-world data ethics examples and provide an opportunity for you to consider the consequences, both intended and unintended as well as positive and negative.

## Applying data ethics to real-world scenarios

#### Identifying consequences

In this section, we present data ethics case studies from the real world. For each case study, you should focus on the first phase of the Consequence Scanning Toolkit and try to:

1. identify the intended and unintended consequences,
2. decide on which positive consequences to focus on, and
3. decide which consequences need to be mitigated.

To support your learning each case study is accompanied with sample outcomes.

### Should we close fire stations?

Throughout this course we have been looking at real policy decisions made to close 10 fire stations in London. Given the sensitivity of such a topic it is an ideal candidate for evaluating the ethical challenges and risks of taking such action.

### Closing fire stations: Consequence Scanning

Ask yourself the following:

* What are the intended consequences and benefits?
* What are the potential obstacles that might be faced?
* What do we know about the current situation from the data?
* Can any inferential statistics be used to model the potential consequences?
* How can we present our findings clearly to audiences to paint a clear picture of the benefits and how we will address the risks?

View Possible Answers

Closing the fire stations will provoke an emotive and negative response. However, looking at the data we have seen that performance is not fair across London, in fact it is worse where people live. Presenting the data in clear and concise ways will help people understand the situation. More than this however, being open and transparent with the data, analysis, expected positive and negative outcomes helps even more.

### How expensive is it to play Pokemon Go?

When Pokemon Go was released, users noted that pokestops and gyms occurred much less frequently in poor neighbourhoods, and black and other minority neighbourhoods. This was due to bias in the dataset used to create the locations.

##### What is Pokemon Go?

Pokemon (short for the original Japanese title of ‘Pocket Monsters’) is a popular franchise owned by Nintendo with two other companies (Game Freak and Creatures). Pokemon are imaginary creatures and players (known as trainers) can catch these animals and use them to battle other trainers.

Pokemon Go is a free-to-play, location-based augmented reality game developed by Niantic in collaboration with Nintendo. In the game, players use their mobile device GPS to locate, capture, battle and train virtual Pokemon. Players collect ‘pokeballs’ – needed to catch pokemon, and therefore to play the game – from pokestops (check points) and battle other pokemon at ‘gyms’, which are based on their GPS location. Without access to pokestops, players must pay for pokeballs and other items. Players with easy access to pokestops and gyms, on the other hand, can earn more experience points and collect more items for free.

##### Where did the data come from?

The locations of pokestops and gyms in Pokemon Go were derived by Niantic from the locations of ‘portals’ in a previous augmented-reality GPS-based game, Ingress.

##### Who collected the data?

Locations for Ingress were predominantly crowd-sourced, both via an existing database – Historical Marker Database – and by Ingress players. The Historical Marker Database reflected contributions from approximately 3,000 volunteers across the US who were mostly male. Ingress players also appeared to be mainly male, young and English-speaking.

### Pokemon Go: Consequence Scanning

Ask yourself the following:

* What are the intended consequences of Pokemon Go?
* What are the unintended consequences of Pokemon Go?
* Which intended consequences should Pokemon Go look to maximise?
* Which unintended consequences should Pokemon Go mitigate?
* Which of the consequences are within Pokemon Go's capability to manage?

View possible answers

Studying the locations of Pokestops, derived from Ingress data, exposes the limitations of crowd-sourced data. In Pokemon Go – which has had significantly broader reach than Ingress – gaps in the underlying pokestop location data stand to potentially drive up the cost of playing Pokemon Go for users in predominantly black, hispanic or low socio-economic neighbourhoods.

The underlying data sources for the locations of pokestops in Pokemon Go are non-personal, however these geolocation gaps stand to negatively impact on certain Pokemon Go users (by increasing the overall cost of participation). A data ethics framework that helps people to identify how data use affects different demographics could have helped identify this issue before the launch of the game and allowed the developer to take steps to mitigate the issue and broaden the pool of players.

Of course, these are only some of the possible consequences. You may have identified others.

### The demand for Uber

Uber uses data about their users to provide them with personalised pricing and create a more efficient transport market where supply matches demand. Although no personal data was exposed, Uber’s model nevertheless put people at risk in an emergency situation. In 2014, a hostage situation in Sydney led to people feeling fearful for their lives and wanting to escape the area by any means possible. Many didn’t feel safe to do so via public transport, thus tried to book an Uber, only to find that prices had surged to four times higher than usual, a level never seen by many.

Uber issued a statement on Twitter saying: ‘We are all concerned with events in CBD. Fares have increased to encourage more drivers to come online & pick up passengers in the area’. However, the fare-calculating algorithm raised those fare prices by up to four times the normal rate. This meant the minimum fare per trip was $100.

This led to an outcry on social media causing Uber to respond by offering free rides and refunds for people charged with higher rates.

Uber faced a huge backlash. Even if the surge model is designed to encourage more drivers onto the road, is it ethically acceptable to (1) Encourage drivers to move into a potentially dangerous area and (2) charge people such high fares to escape the area.

#### Uber Surge Pricing: Consequence Scanning

Ask yourself the following:

* What are the intended consequences of surge pricing?
* What are the unintended consequences of surge pricing?
* Which intended consequences should Uber look to maximise?
* Which unintended consequences should Uber mitigate?
* Which of the consequences are within Uber’s capability to manage?

View possible answers

Following the 2014 crisis, Uber issued an apology: “We didn't stop surge pricing immediately. This was the wrong decision.” Since then, Uber has defended its surge pricing strategy in other cities, and reached an agreement with regulators in the US to restrict the policy during national emergencies (capped at 2.1x).

In November 2017, Uber responded to requests of drivers to adjust its surge pricing policy in the US and Canada, and implemented a public program (‘180 Days of Change’) to make improvements to the driver experience. This included the following benefits to drivers and customers:

* Making surge periods last longer to provide more time for drivers to surge areas
* New surge pricing will be made based on the location of the driver, not the passenger ie when drivers reach the surge area, an extra dollar amount will be added to their next trip regardless of where the passenger is located
* Surges will be simpler to understand eg by telling drivers upfront exactly how much additional they’ll be paid in surge fees for the trip

More recently, regulators in NYC have introduced measures to regulate the rideshare industry and create more ‘fairness’ for drivers- such as introducing minimum pay rate for rideshare drivers, and regulating the number of licences/vehicles to prevent saturating the market. However, this could lead to more surge-pricing periods with a capped supply of drivers.

### Weighing up the wrong risk factors

Online car insurance firms use predetermined algorithms to assess the risk of a user filing a claim against their policy. At the beginning of 2018, this practice faced public backlash when the Sun newspaper found that insurance quotes for drivers with the traditional English name ‘John’ were far lower than quotes of the same for drivers named ‘Mohammed’.

##### The impact

Under particular scrutiny was the insurance company Admiral where their deal on Go Compare for fully comprehensive insurance on a 2007 Ford Focus in Leicester was priced at £1,333 for ‘John Smith’ and £2,252 for ‘Mohammed Ali’ - a price difference of £919. Sixty quotes run across ten different cities on GoCompare and other comparison sites revealed the companies in question consistently charged more if the driver was called Mohammed.

##### Sourcing the data

These algorithms are based on key indicators about users which are measured against a data set to weigh risk and calculate premium rates. Common rating factors vary but include age, demographic, marital status and driver history i.e. if you have had accidents in the past, an algorithm will assume that you are likely to have more accidents in the future.

In this case, the training set for the model contained all of these factors but also included the drivers’ names which should not be relevant. As a result the model probably ended up selecting and assigning significant weight to names that occurred frequently in the data, whether as a simple one-to-one mapping (very common name = more incidents) or as a proxy for a combination of other, more relevant attributes like age, gender and location (young males in urban areas = higher risk).

### Car insurance: Consequence Scanning

Ask yourself the following:

* What are the intended consequences of insurance risk models?
* What are the unintended consequences of insurance risk models?
* Which intended consequences should insurance providers look to maximise?
* Which unintended consequences should insurance providers mitigate?
* Which of the consequences are within insurance providers' capability to manage?

View Possible Answers

The algorithms are based on key indicators about users which are measured against a data set to weigh risk and calculate premium rates. Common rating factors vary but include age, demographic, marital status and driver history i.e. if you have had accidents in the past, an algorithm will assume that you are likely to have more accidents in the future.

In this case, the training set for the model contained all of these factors but also included the drivers’ names which should not be relevant. As a result the model probably ended up selecting and assigning significant weight to names that occurred frequently in the data, whether as a simple one-to-one mapping (very common name = more incidents) or as a proxy for a combination of other, more relevant attributes like age, gender and location (young males in urban areas = higher risk).

Following public outcry, the Financial Conduct Authority (FCA) announced it would extend its [consumer data research](https://ethicsandinsurance.info/2018/04/10/dual-pricing-in-insurance/) to examine whether names can impact the cost of car insurance cover. This is part of a widespread investigation into dual pricing and ‘optimal pricing’ tactics within the insurance industry during 2018/19.

In future, insurers will need to better understand these issues to prepare a response that the regulator can judge to be informed and reasonable. Insurers need to be clear on the principles that underpin their pricing strategy, and have a clear template for how those principles are being applied within all those pricing algorithms in widespread use e.g. what signals are pricing algorithms looking for in the data.

### Considerations for data ethics

**A data project can involve only non-personal data and still be unethical**

Issues of bias, inaccuracies and inconsistencies in data can arise regardless of the nature of its source, whether personal or non-personal. Researchers analysing trends in aggregated search-engine data, for example, must account for gaps in who has contributed to that data (a predominantly digitally literate population).

Identifying and mitigating issues within data sources that could negatively affect certain population demographics if left untreated is part of treating data ethically.

## Reflecting on your practice

### Reflecting on your practice

To support your learning we recommend taking the opportunity to reflect on the module and examine how the lessons covered can be applied to your own professional practice.

**To aid the process, we've provided a SDS Module 5 Reflective Template (.docx).**

The purpose of this document is to give you an opportunity to explore how your learning can apply to your professional practice. It is not an assessed piece of work. Instead, it is intended to provide a discussion framework for you and your line manager, mentor or colleagues.

We recommend that the reflections you make during this course are used as part of a portfolio of learning and development, used in conjunction with your own professional development frameworks.

#### There is never a lack of data, or technology

There is never a lack of technology or data. Both are often seen as the “golden bullet” to solving a whole host of vaguely defined problems. In fact having too much technology and data can cause more problems, both legal and ethical. Too much data and technology becomes more challenging to manage, more disparate, siloed and not to mention aged and part of the legacy/debt of the organisation. Having the right processes and people in place to support the data and technologies and being focussed on your purpose will help ensure that technologies and data are suitable for the purpose and not there because it is easy to get both.

##### Consequence Scanning

Consequence Scanning is an agile practice for responsible innovators. It's a framework of practice to enable teams who work on data enabled projects to explore the impact of their practice by identifying the intended and unintended consequences, and defining the appropriate actions to mitigate negative effects and promote positive impact.

In this section, we will explore the consequence scanning toolkit and how it can be used in the practice of data ethics.

Launch e-learning

If you are viewing this page using the Moodle app, click here to launch the e-learning

CLOSE AND RETURN TO COURSE

Loading...

#### Data ethics scenarios

Something about it here

Launch e-learning

If you are viewing this page using the Moodle app, click here to launch the e-learning

CLOSE AND RETURN TO COURSE

Loading...

## Next steps in data ethics

### Enabling positive impact: next steps in data ethics

For some time, the Open Data Institute has been working on measures to help organisations build trust in how they collect, use and share data, and to foster better use of data overall. Trust is an essential component of society. When trust breaks down, the public lose faith in the institutions that provide them with services, and organisations lose the ability to share data and collaborate in ways that could improve all our lives.

By exploring data ethics in depth, the ODI uncovered common assumptions about data ethics that we wanted to challenge.

* ‘Data ethics is only an issue where activities involve personal data.’ Ethics issues must also be considered in the collection and use of non-personal data. For example, not publishing the location of bus stops in poorer neighbourhoods can mean that the advantages of smartphone-mapping and route-finding services are not available to people who live in those areas, increasing existing inequalities.
* ‘Data ethics is concerned with data protection compliance.’ Complying with legal obligations – under the EU General Data Protection Regulation, for example – is just one part of treating data ethically. Data-related activity can be unethical but still lawful.
* ‘Data ethics is (only) about how organisations use data.’ Data ethics is about the impact that all data activities have on people and society. Collecting and sharing data only about certain groups of people may disadvantage them relative to others. All activities should be subject to ethical examination.

### Data ethics tools

The ODI has developed and is responsible for a series of data ethics tools that focus on supporting individuals and organisations in embedding data ethics practices into their projects and overarching organisational strategies.

Project Level

### [Consequence Scanning Toolkit](https://doteveryone.org.uk/project/consequence-scanning/)

For a quick assessment of a project, with particular emphasis on the intended and unintended consequences on people, communities and the planet. It provides an opportunity to enhance positive consequences, and to mitigate or address potential harms from negative consequences before they happen. It considers what can be acted upon, what can be influenced, and how to monitor.

### [Data Ethics Canvas](https://www.theodi.org/article/the-data-ethics-canvas-2021/)

For deep consideration of a project. It is divided into four areas: ‘Data’, ‘Impact’, ‘Engagement’ and ‘Process’. The Data Ethics Canvas is a tool for anyone who collects, shares or uses data. It helps identify and manage ethical issues – at the start of a project that uses data, and throughout.

It encourages you to ask important questions about projects that use data, and reflect on the responses. These might be:

* What is your primary purpose for using data in this project?
* Who could be negatively affected by this project?

The Data Ethics Canvas provides a framework to develop ethical guidance that suits any context, whatever the project’s size or scope.

Organisational Level

### [Data Ethics Maturity Model](https://www.theodi.org/article/data-ethics-maturity-model-benchmarking-your-approach-to-data-ethics/)

Unlike the other tools, the Data Ethics Maturity Model is not intended for specific data sets. Considering data ethics at project level is good, but deep and lasting developments must happen at organisation level. This is the focus of the Data Ethics Maturity Model. The tool helps assess and benchmark how widely embedded data ethics culture and practices are across your organisation.

It provides a framework to consider current performance and future ambitions, to embed a culture of ethical data practices. This tool builds on the Open Data Institute’s [Data Ethics Canvas](https://www.google.com/url?q=https://www.theodi.org/article/the-data-ethics-canvas-2021/&amp;sa=D&amp;source=editors&amp;ust=1686323699302615&amp;usg=AOvVaw1XyxRe2FklFsMHRqvZOV1j), which helps identify and manage ethical issues within a specific project, and may be particularly helpful to [Data Ethics Professionals](https://www.google.com/url?q=https://www.theodi.org/event_series/data-ethics-professional/&amp;sa=D&amp;source=editors&amp;ust=1686323699303019&amp;usg=AOvVaw2bXydGcgIFNrholHr1ezNL) who are working to embed a data ethics culture within their organisation.

#### Learn more with the ODI

We offer a mix of short introductory courses, and longer courses that offer opportunities to really explore the area of data ethics and even become certified by the ODI as a data ethics professional.

Data Ethics Essentials

The aim of the ‘Data Ethics Essentials in AI’ course is to emphasise the significance of data ethics and enable learners to develop their understanding of data ethics as a domain, examine how data ethics manifests itself in data-enabled products and services, such as machine learning and AI, and empower learners to examine their own direct impact and incorporate ethical considerations into their own professional practice.

The course is for individuals who use data in their professional roles, but are new to data ethics. The sequential modules introduce the concept of data ethics and its importance in today’s society with increasing influences from new technologies, such as generative AI. The course then explores the available data ethics tools and frameworks, providing a basis for learners to apply these tools in their own professional practice.

##### Educational aims

**The course aims to develop greater skills and knowledge in:**

* Understanding the importance of data ethics as a domain
* Explore data ethics in the context of AI and other data related projects
* Examining data projects to identify the intended and unintended consequences
* Evaluate the ethical challenges associated with data, and develop awareness of potential biases that can arise through the use of data-enabled decision making
* Identifying common barriers to adopting approaches to data ethics
* Applying practical tools to investigate data projects through the lens of data ethics
* Developing an actionable approach to managing data ethics consequences

[Learn more about our data ethics learning offering on our website](https://learning.theodi.org/courses/introduction-to-data-ethics)

.

Data Ethics Professionals

Unlock the potential of responsible data creation, sharing and use, and become a sought-after data ethics professional with our comprehensive course designed for forward-thinking individuals passionate about evaluating data practices with the potential to adversely impact on people and society. Learn to build trust, be transparent, and accountable in the data-driven world.

This course is best suited to professionals who manage or use data to inform business strategy as well as day to day decision making. This includes, but is not limited to, performing data organisation activities with raw data, performing analysis either in spreadsheet software or BI Tools as well as interpreting data outputs to share with stakeholders.

Successful participants will be:

* Recognised as competent in the area of data ethics and the practical steps required to help organisations minimise potential harms that come from data collection, use and sharing.
* Able to think critically about data ethics.
* Ready to apply tools and techniques to analyse the current situation within organisations.
* Equipped to suggest potential actions to minimise risk – be it through process adoption as guided by tools such as the data ethics canvas, or changes to the way the data itself is collected, used, or shared.

##### Educational aims

**The course aims to develop greater skills and knowledge in:**

* A deep understanding of data ethics as a domain
* Practical expertise in data and techniques to evaluate ethical and legal risks
* Ability to perform a professional piece of research, drawing on historical and peer-reviewed literature
* Ability to identify and use a number of tools to help evaluate ethical impacts and suggest actions to minimise harmful impacts from data
* [Learn more about the course on our website](https://learning.theodi.org/courses/data-ethics-professional).