Your 2023 Data Science Roadmap

evidence based >

DATA SCIENCE INFINITY

About THE AUTHOR

The founder of Data Science Infinity, Andrew Jones is one of the best-known Data Science instructors in the world – helping countless students move ahead of the competition and into amazing roles in the field. His 15+ year career includes time at global tech giants Amazon & Sony PlayStation, he has 5 patents to his name, has interviewed hundreds & hundreds of candidates & authored a book on Data Science recruitment.

This e-book is based upon Andrew's experience in the field, as well as his conversations with hundreds of hiring managers & recruiters, making it the most reliable source of information about the skills & tools that are in demand for a career in Data Science & Analytics.

Subscribe on Youtube

Follow Andrew on LinkedIn

WHAT'S INSIDE

2023 A Big Year For Data Science

Data Science is exciting, lucrative, and future-proof – and 2023 will see the field accelerate even further.

With ever-increasing volumes of data being generated and collected, advancements in computing technology, and the recent explosion in Artificial Intelligence, companies from all industries are hiring Data Scientists so they can:

- Build exciting products to drive new revenue streams
- Optimise & automate processes & cut costs
- Understand their customers & predict what they will do next
- Make the right decisions in real-time to stay ahead of the competition

"I've seen a huge uptick in Data Science opportunities since early January. We have been very busy focusing on this area for our clients"

Joshua Smith
Associate Director, Data Science Analytics Engineering Search at MBN Solutions

When demand is high, those who have the right skills & tools can expect to be valued highly.

Salary insights company PayScale calculated that the median salary for a Data Scientist in the US is **$103k** (and remember that is the median, so half of Data Scientists are indeed earning more than that!)

Many fail on this journey. You will succeed.

There is an almost endless number of skills & tools to potentially focus on learning – it can be overwhelming.

To be successful, you need to put your time & energy into learning the skills that hiring managers genuinely need & want.

This document is a high-level distillation of my conversations with hundreds of hiring managers & recruiters in the field. Use this as a framework, and you'll be set for success!

*"I landed my new role at Amazon totally thanks to Andrew, DSI, and this wonderful community. DSI has been the **best academic choice of my life** - it has given me **better results than two degrees**"*

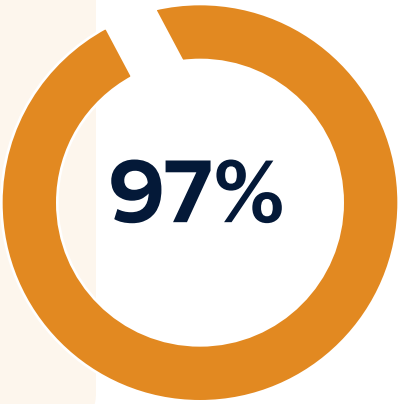
Andrea

I asked “Does your team utilize SQL for Data Science?"

I asked hundreds of Data Science hiring managers that question.

97% said yes. This is a must-have skill.

Thankfully, learning SQL, or at least learning the 90% of SQL that you need to know for Data Science is not that difficult.



As an idea of what you need, make sure you build up confidence with:

- General query structure, so SELECT, FROM, WHERE, ORDER BY
- Aggregation functions and the GROUP BY clause
- Conditional logic using CASE WHEN
- Combining data using INNER JOIN, LEFT JOIN, and CROSS JOIN
- The use of Window Functions
- How to stack data using UNION and UNION ALL

As a bonus, learn how to link together multiple queries into one using either **TEMP TABLES or CTE**. These can come in really handy in interviews as they can help you cleanly break the question down into multiple parts!

Long story short, **SQL** is non-negotiable if you want to become a Data Scientist or Data Analyst, but what about more pure “programming languages”?

PYTHON VS. R

Both are great programming languages – but which should you learn?

According to my research...55% of hiring managers in the field said their team was using R for Data Science.

55%

For Python - this was 87%

87%

While R is an amazing language, if your goal is to become a Data Scientist it just makes sense to put the odds in your favor and learn Python.

I would advise strongly against trying to learn both languages at the start of your career. Some courses look to teach both simultaneously and this can make the whole learning process much harder. Pick one, and focus on that!

"I had over 40 interviews without an offer... After DSI I quickly got 7 offers including one at KPMG and my amazing new role at Deloitte!"

Ritesh

Stay focused when learning Python...

There are an enormous number of Python libraries & packages. When you're getting started, don't let this confuse you, or bog you down. Start with these:

>> Base Python

Gain an understanding of different data types (float, integer, boolean, string) as well as the different data structures that are available (lists, sets, tuples, dictionaries). Be able to explain why each exists.

>> Pandas

Pandas is a Python library used for data manipulation & helping Data Scientists understand & explore the data they're working with.

Get to grips with; importing data, creating Data Frames, accessing specific rows/columns/cells, sorting data, joining & merging data, aggregating data, and dealing with missing values.

>> Numpy

Numpy is a Python library used for fast mathematical processes on data stored in array format. It is also often utilized for storing & manipulating image data!

Get familiar with creating & manipulating arrays of differing dimensions as well as applying mathematical operations on the data stored within!

Python...

>> Matplotlib

Matplotlib is the most used Data Visualization library in Python. Data Scientists rely on this to showcase insights & trends in the data.

Get familiar with creating different chart types, utilizing subplots, formatting plot features, colors & styles, and adding text to aid the interpretability of the plot for the viewer!

>> Scikit-Learn

Scikit-Learn is the most popular ML library in Python, containing dozens of algorithms for Machine Learning as well as a whole host of data preprocessing techniques.

>> Streamlit

Streamlit is an amazing library that allows you to easily turn your code & projects into interactive web apps. This is perfect for bringing your learning journey to life and for showcasing your projects when applying for roles!

*"The DSI content is truly the best. It genuinely **surpasses what I learned in my Master's in Data Science** program"*

Math & Statistics

Time and time again, you are going to hear about

"all the mathematics you need to know before you're allowed to become a Data Scientist"

My single piece of advice here: **"Don't be scared off by this"**

There is no shortage of people in this field who will sneer at newcomers who don't know X, Y, or Z (I call them "gatekeepers") but the reality is as follows:

Yes, you do need to know some mathematics.

There is no getting around that. It underpins a lot of what we do in the field.

But you DO NOT need to spend a year reading dusty textbooks before you're allowed to progress, or build things, or touch anything, or before you're allowed to land your first role.

Don't let anyone tell you otherwise.

Start with statistical concepts such as:

- Types of Data
- Statistical Distributions
- Hypothesis tests
- The Central Limit Theorem
- Confidence Intervals.

From these foundations, you can start learning more advanced mathematical concepts like Linear Algebra (which covers things like regression algorithms, and gradient descent).

A great way to do this is to learn as you start applying things like Machine Learning algorithms. It's so much more fun (and productive) learning while testing and modifying things, and seeing what changes, and why!

MACHINE LEARNING

Don't try to learn every algorithm.

Focus on these instead!

Many new Data Scientists fall into the trap of learning as many ML algorithms as they can.

This is detrimental to your progress, and your career prospects. The following list of algorithms & concepts are used to solve 90%+ of business problems that require ML - get a deep understanding of these, and you'll position yourself ahead of other candidates.

Start by getting a good understanding of the difference between Supervised Learning vs. Unsupervised Learning. Simply put, these are two areas within Machine Learning with slightly different end goals, but it's important to understand what each is, and which algorithms might be useful for tasks that fall into each area.

Once you've got the grips with that, you want to learn how the following algorithms work, and how to apply them in practice (these are the ones hiring managers said were most commonly used to add value in their teams)

Supervised Learning

>> Linear Regression

This often is the go-to approach for regression tasks (i.e. where we're predicting a number such as sales, or salary, or price)

>> Logistic Regression

While still a "regression" algorithm, this often is the go-to approach for classification tasks (i.e. we're predicting a class or a type, so whether a customer will leave the business next month, yes or no)

>> Decision Tree

These can be used for both regression & classification tasks, and are famously easy to interpret for stakeholders!

>> Random Forest

Made up of many Decision Trees working in unison. These are very powerful, and can also be used for both regression & classification tasks.

>> K-Nearest-Neighbours

Often simply referred to as KNN, this type of algorithm uses the distances between data points to understand what its prediction should be.

Unsupervised Learning

>> K-Means

Often used as a clustering algorithm to group together data points based on distance, to create useful products such as customer segmentations!

>> Principal Component Analysis

A dimensionality reduction technique often just referred to as PCA. It can come in handy to use in conjunction with both k-means for clustering or any supervised learning tasks.

Bonus Algorithms

>> Association Rule Learning

An approach that discovers the strength of relationships between different data points. Think of the "customers who purchased product A are likely to also purchase product B" you see on sites like Amazon!

>> Causal Impact Analysis

Measures the change in a metric after some event has taken place. This could be the uplift in sales after a promotion, the additional clicks, conversions, or signups generated by an online ad campaign, or it could be the change in share price after a market event. A very important concept!

"Since completing DSI I've had two interviews and have received offers for both! This is definitely the best course for those moving into Data Science"

Angelo

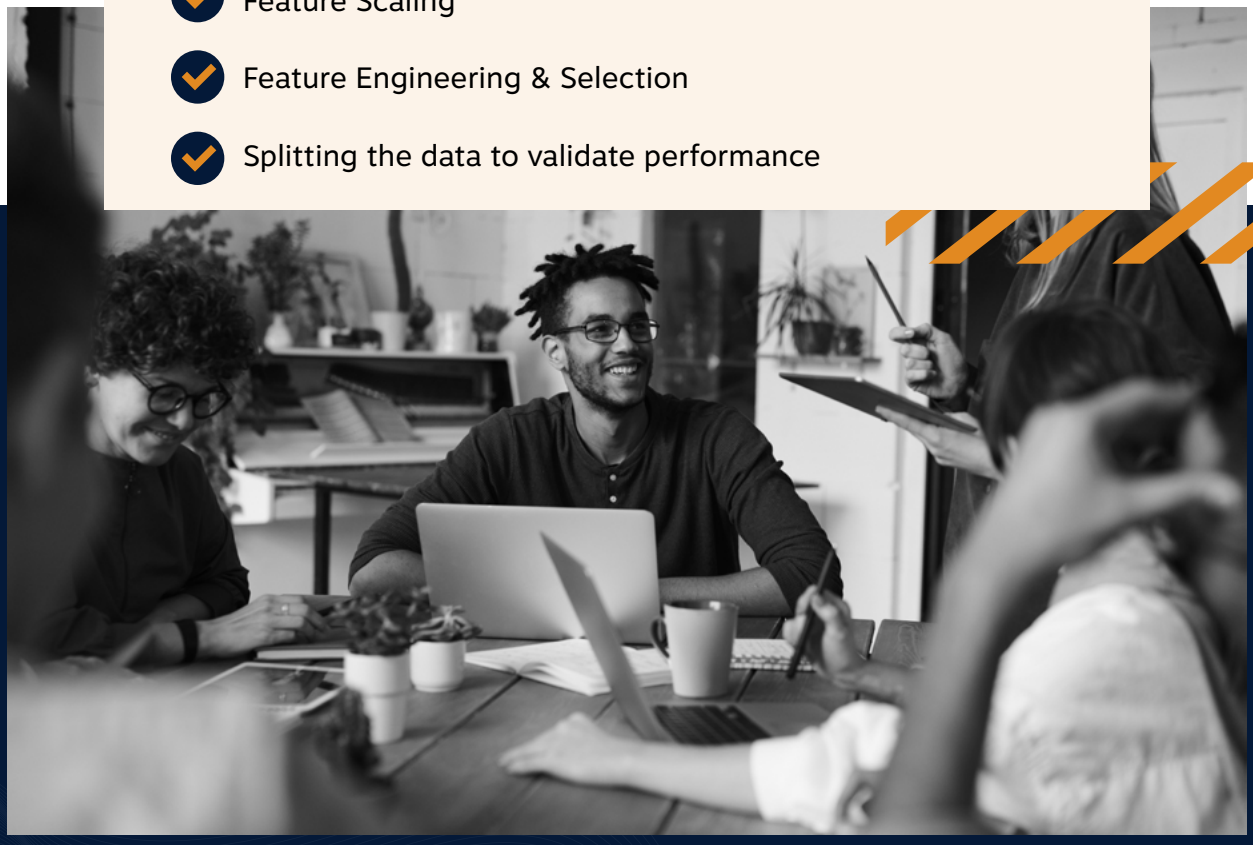
IMPORTANT!

Garbage in, Garbage out...

Even the best ML algorithm can't learn from bad data. Good knowledge of data preparation & cleaning techniques will differentiate you from other candidates, and ensure your ML projects are successful.

Key Data Preparation Concepts To Know

- Processing missing values
- Dealing with duplicate & low-variation data
- Dealing with incorrect & irrelevant data
- Processing Categorical Data using encoding techniques
- Assessing & processing outliers
- Feature Scaling
- Feature Engineering & Selection
- Splitting the data to validate performance



Should I take on Deep Learning?

Yes, absolutely. But only when the time is right!

Don't skip the foundational skills first - knowledge around concepts like Linear & Logistic Regression make up a large part of Deep Learning, so you want to know these intimately first.

Deep Learning is being used to solve some very cool problems in the field these days, but 99% of Data Science tasks that even need something along the lines of Machine Learning can be solved using the list we discussed earlier. Get the core skills first and then take on Deep Learning!

Other key Data Science tools to know

There is no shortage of tools listed on Data Science job descriptions - but when starting out, it's important to be able to focus your time on the ones that are truly in demand. Here are two tools that are widely used, and that many hiring managers will want to see:

>> Github

GitHub is a version-control platform used by virtually all Data Science teams, as it provides a suite of tools for managing code, collaborating with others, and sharing work with the world!

>> Tableau

Tableau is a powerful data visualization tool that allows users to easily connect, visualize, and share data insights. It enables users to create interactive and dynamic dashboards, reports, and charts without requiring programming skills.

Tableau supports various data sources including spreadsheets, databases, cloud services, and big data. With Tableau, users can explore, analyze, and communicate insights from data, helping organizations to make data-driven decisions.

Do I also need to know about Data Engineering?

Yes, and no. Creating & training a Machine Learning model locally is great. Getting it to a place where it can be used by end-users or customers is what will drive a lot of business value. Often, this work will be done by Data Engineers but as a Data Scientist, it is good to know the basics of:

>> ML Pipelines

These are simply what tie together all the steps from data ingestion, to preprocessing & cleaning, to prediction, into one standalone object.

>> Parameter Tuning

Techniques to help us find the optimal parameters for our algorithms, ensuring they provide the best possible accuracy or performance

>> Deployment

The process of placing the trained model pipeline where it can be used for prediction, i.e. in a web app sitting behind a website. To start with, get familiar with GitHub, and build something simple with the Python library *Streamlit*

Understanding where the data comes from, and knowing how to best work with Data Engineers is certainly useful, but you don't need to cover both skill sets!

"100% worth it, it is amazing. I have never seen such a good course and I have done plenty of them!"

Khatuna

Think Data Science is just about technical skills? *Far from it...*

It is so important to emphasize that Data Science is not all about technical skills. If you're transitioning from another role or field, this is where your initial advantage lies.

In fact, the best Data Scientists I've worked with in my career are not the “smartest" ones, by definition. Absolutely, they know their stuff in terms of coding, statistics, and other key data concepts but what differentiates them is this:

They understand what the business problem is, or what the business is trying to achieve. They use data, and their unique skillset in clever and often simple ways to solve these problems or to add tangible value to the team, business, or end-user.

“The best Data Scientists start with the business problem, and then work back to a Data Science solution from there - not the other way around!”

Communication is vital!

A good Data Scientist knows a lot of technical concepts. A great Data Scientist can simplify these down in a way that gets everyone in the business onboard.

As Data Scientists we're here to solve problems, not introduce new ones...we're here to enhance, and accelerate business decision-making - not get in the way of it!

If you're transitioning from another field, these softer skills can genuinely put you above the competition - so you're in a strong position to become a great Data Scientist!

Something I say all the time to the aspiring Data Scientists in Data Science Infinity is “No one is going to pay you just to be good at coding, or just to be good at maths, or just to know a lot of machine learning algorithms...but they will pay you, and they'll pay you extremely well, to add tangible value to their business or to the end-user”

“DSI is the best investment I have ever made. It gave me the confidence to code & I’ve now landed an amazing role at an incredible company. A special thanks to Andrew for being such a great teacher and for his constant support through this journey! I would 100% recommend this course to every aspiring Data enthusiast”

Manasi

*inside-knowledge!

From Learning to $Earning

Now you know a bit more about the skills you need for this exciting, future-proof & lucrative field - the next step is to move into a great role.

This is often easier said than done as competition for roles is high.

Let me tell give you some inside knowledge from my time as an interviewer - to help you move ahead of the pack, and into that role you want!

"DSI gave me the skills and confidence to step out of my comfort zone and start applying for jobs. As a result, I've now landed a job as a Data Scientist in an absolute dream company!"

Luka

Build a Portfolio that will get you HIRED!

A portfolio of projects can be an excellent way to showcase your skills when you're early in your Data Science or Analytics career.

But, I want to quickly bust a myth about portfolio projects (based on my experience interviewing & screening hundreds & hundreds of candidates at companies such as Amazon & Sony)

Myth Data Science portfolio projects need to use huge volumes of data and/or use extremely complex Machine Learning or Deep Learning solutions!

Truth A portfolio containing varied, impactful, and clearly communicated projects is much more likely to get you hired.

Hiring Managers & Recruiters have very little time to get into the depths of your projects so you must make it quick and easy for them to see your value, and the types of tasks you have the ability to solve.

There are no right or wrong projects for a Data Science portfolio - it really all comes down to how well they are written up!

Break your project write-ups into these sections:

1. Project Overview:

Make it easy for the reader! Right at the top, showcase the highlights from the full write-up

2. Concept Overview:

Showcase your understanding of the concepts you'll be applying

3. Data Overview & Preparation:

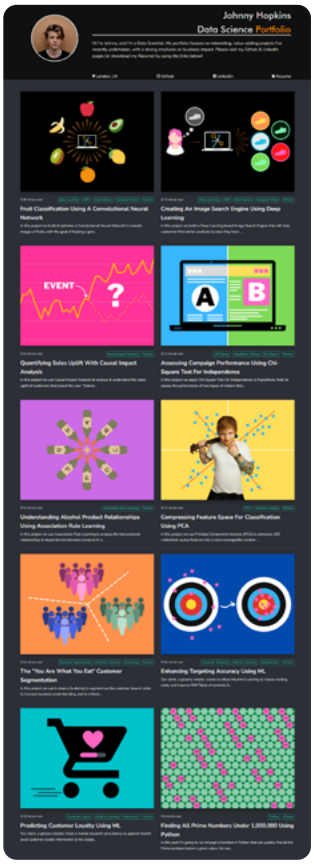
So important in the real world, but almost always missed in portfolio projects!

4. Application:

The application, the code, the nuts & bolts!

5. Analysing The Results:

The results, and importantly, what they mean!



6. Growth & Next Steps:

Explain what would you do to improve this, or what you would do if you had more time!

***"I got it!** Thank you so much for all your advice and help with preparation - **it truly gave me the confidence to go in and land the job!**"*

Marta

10 Project Ideas To Get You Noticed!

A varied portfolio of 5-10 clear & impactful projects will beat out a single long & convoluted capstone project every time. This approach can really help a hiring manager get a feel for the types of things you can do, and the ways you can tackle problems. In Data Science Infinity, we build the following projects based on ideas hiring managers were interested in seeing.

Please use them for inspiration for your own portfolio!

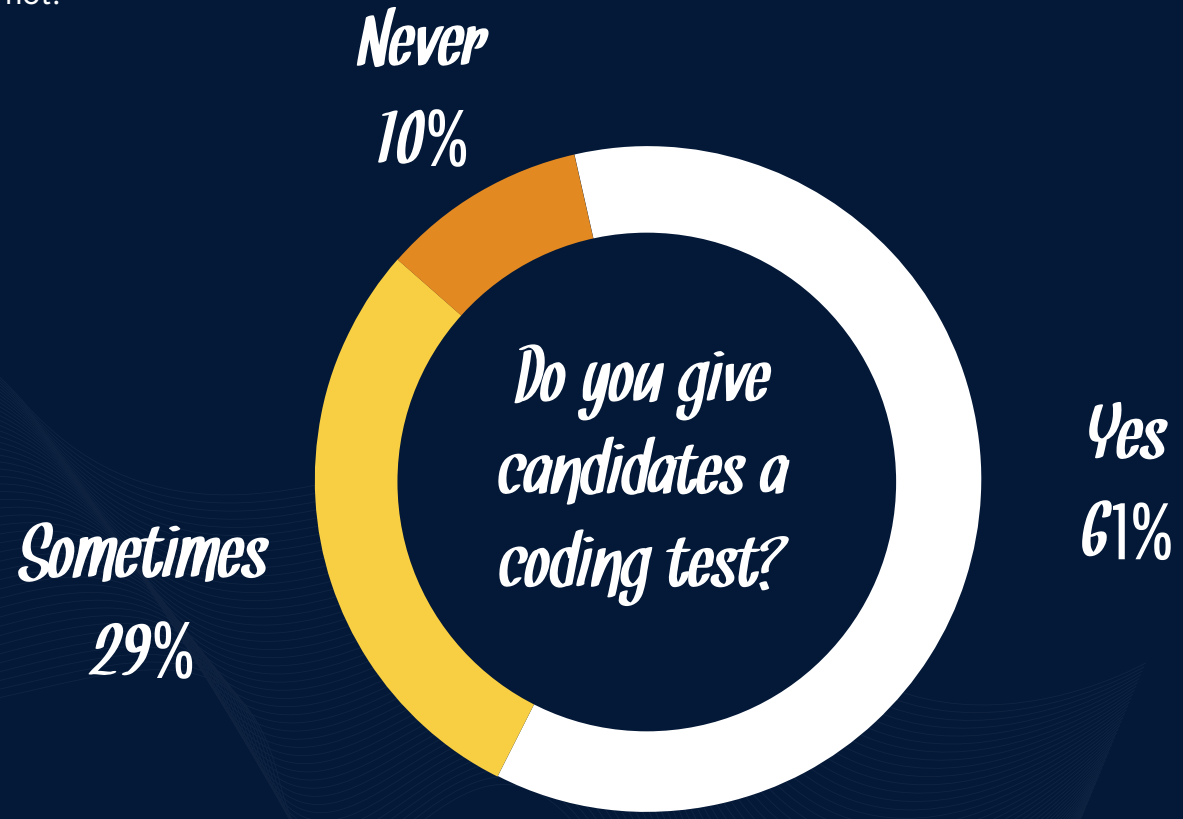
- Assessing Campaign Performance using Chi-Square (A/B Testing)
- Predicting Customer Loyalty with Machine Learning (Regression)
- Enhancing Targeting Accuracy with Machine Learning (Classification)
- The "You Are What You Eat" Customer Segmentation (Clustering)
- Compressing Feature Space for Classification (PCA)
- Assessing Alcohol Product Relationships (Association Rule Learning)
- Quantifying Sales Uplift using Causal Impact Analysis
- Fruit Image Classification using CNN (Deep Learning)
- Creating An Image Search Engine (Deep Learning)
- Finding All Prime Numbers Under 1,000,000 Using Python

***"DSI is incredible** - everything is taught in such a clear and simple way, even the more complex concepts"*

Arianna

Ace The Coding Test!

From my research with Data Science leaders, hiring managers, and recruiters from various companies around the world while creating Data Science Infinity - it came out that 61% did get candidates to undertake a coding or technical test, 29% did sometimes, and the remaining 10% did not.



Those numbers would suggest that is it very important to be prepared for something like this when interviewing!

For most entry-level Data Science or Data Analyst positions the test will simply be based in SQL, or potentially Python.

The level of difficulty will vary - and often it will start with basic questions and move up to more complex ones.

If the role is focused more on insights & analytics, then the test will probably be less complex than for a role focused more on data engineering or software engineering.

Get as much information about what it will entail - and then practice. There is nothing better for this type of test than being in the right mindset, and this comes from consistent practice.

There are lots of online resources for this, but make sure you're focusing on the right thing. Don't get super stressed trying to answer all sorts of insanely complex software engineering-type problems when you aren't going to need that.

When completing the test, if you don't know the answer, or the exact syntax - don't worry.

Put what you think should happen in words. You can still get a lot of kudos for knowing what steps to follow, or what considerations are important.

Don't leave any question blank!

*"I've taken a number of Data Science courses, **and without doubt, DSI is the best**"*

william

Ace The Behavioral Interview!

In non-technical interviews, you'll mostly be asked to discuss projects that you've worked on in the past.

The simplest way to nail questions in these interviews is to prepare in the same way you'd want to answer. Here is a high-level framework from Data Science Infinity called the **CRAIG System**

For each project, make sure you are well versed on;

C = Context

Give context around the business problem and why it needed to be solved – this pulls the interviewer into your story

R = Roles

What was your role in the project – quick and easy!

A = Action

The specific actions you took from inception to conclusion. Refine this to the most succinct & compelling narrative but ensure you keep auxiliary context up your sleeve, for example, "why you chose solution C over solution A and B". A good interviewer *will* ask this!

I = Impact

What was the result of your work? Super important – but often missed or underemphasized. Use tangible figures, for example "drove $x sales" or "saved y hours"

G = Growth

Ask yourself "If I could have started the project again, what would I do differently?"

This sort of thinking around your career and the projects within it, can be so much more impactful than you might think.

It shows you have an awareness of business impact, it shows you have an understanding of the nuance of what you do from a technical point of view as you're thinking about why different solutions would work more effectively, and in general, it just shows a growth mindset, that you're always looking to build, and to improve. And, trust me, that is a lethal combination!

"I did a lot of research before choosing DSI, I asked other students and their experience had been really good. It was definitely worth it - I feel so confident in Data Science now!"

Lovepreet

I hope this has provided some direction for your exciting journey into this fascinating, lucrative, and future-proof field.

If you have any questions about Data Science or indeed Data Science Infinity, please do reach out using the links below - I'm here to help!

Andrew

Subscribe on Youtube

Follow Andrew on LinkedIn

Contact Andrew Directly

Your Time is Now

DATA SCIENCE INFINITY

The program focused entirely on your results

Data Science Infinity focuses on the journey and obsesses over the results of students. Created by former Amazon & PlayStation Data Scientist, Andrew Jones,

Learn the skills & tools that will get you hired

The curriculum of 300+ tutorials, downloadable resources, and quizzes are based upon input from hundreds of hiring managers & recruiters in the field. The curriculum grows & evolves over time - and therefore you as a Data Scientist do as well.

SQL | Python | Statistics | Tableau | A/B Testing | Github | Data Preparation & Cleaning | Machine Leaning | Deep Leaning | Project Best Practices | Interviewing & Application Tips & Inside Knowledge

Learn at your own pace, in a way that works for you

You get *unlimited* access to all current & future content. Content is pre-recorded so you can take things at your own pace, and revisit topics whenever you want in the future.

Get ahead of the competition with expert support & mentorship

You can opt for unlimited and dedicated guidance and support & mentorship for your entire learning journey including personalized support through the application & hiring process,

Be part of a community

Join a community of equally invested peers also chasing success

Build a portfolio that will get you hired

Get a professionally designed portfolio site, and then you add 10 pre-built projects designed specifically to impress hiring managers. This is moving students ahead of the competition when applying for, and interviewing for roles.

The DSI Data Science Professional Certification

Prove your skills, and showcase the value you bring to hiring managers

Check DSI Out For Yourself!

Start Your Data Science Journey With A 25% Discount

More student feedback

*"The **best program** I've been a part of, hands down"*

Christian

*"DSI gave me the confidence to apply for **& land my amazing new role!** The support provided for technical questions & tackling the hiring process is phenomenal"*

Qasem

*"I'd completed my Master's in Business Analytics, but DSI was the first time I felt I had a solid foundation in Data Science **to go forward with**"*

Scott K

*"DSI is the best program. I learned the right skills and built a mindset for success. I felt confident in interviews and **I've now landed an amazing DS role!**"*

Umar

---

*"I started a bootcamp last summer through a well respected University, **but I didn't learn half as much from them**"*

GA

More student feedback

*"I've learned **more than on any other course**"*

Eric

*"Andrew's guidance with my Resume & throughout the interview process helped me land my amazing new role **(and at a much higher salary than I expected!)**"*

Barun

*"I can't emphasize how good this program is...**well worth the investment!**"*

Dejan

*"One of the **best purchases towards learning** I've ever made"*

Scott F

*"DSI is a **fantastic community** & Andrew is one of the **best instructors!**"*

Keith

*"This is a **world-class Data Science experience**. I would recommend this course to every aspiring or professional Data Scientist!"*

David

More student feedback

*"I'm now at University, and **my Data Science related subjects are a piece of cake** after completing this course! **I'm so glad I enrolled!**"*

Jose

*"The fact this program is facilitated by someone with **Andrew's experience & expertise** was a huge factor when deciding to enroll - and it was **definitely the right choice**"*

Sergio

*"The course has such high-quality content. You get your **ROI even from the first module**"*

Donabel

*"DSI is the **best program** I've been part of. It's given me so much confidence to **move forward and land a great job** in the field"*

Ankit

***"DSI truly surpasses the competition, it's amazing.** You get life-long access, all the key knowledge is there, complicated topics are made accessible, and there is so much hands-on learning to put into practice in the real world"*

Fabrizio

2023

Data Science
Roadmap

DATA SCIENCE
INFINITY

Subscribe on Youtube

Follow Andrew on LinkedIn