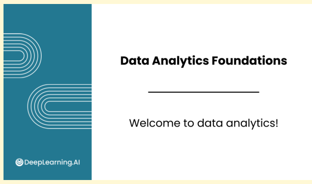
DAG C1M1 scripts

Video title	Isabel	Sean	Slides
L0V1 – Welcome to data analytics!	✅	✅	✅
L0V2 – Generative AI in this course	✅	✅	✅
L0V3 – Module 1 introduction	✅	✅	✅
L0V4 – Life as a data analyst	✅	✅	✅
L1V1 – What is data analytics?	✅	✅	✅
L1V2 – Evidence-based decision-making	✅	✅	✅
L1V3 – A history of data analytics	✅	✅	✅
L1V4 – Modern industry use cases	✅	✅	✅
L2V1 – Defining data	✅	✅	✅
L2V2 – Unstructured data	✅	✅	✅
L2V3 – Structured data	✅	✅	✅
L2V4 – Big data	✅	✅	✅
L3V1 – Data ecosystems	✅	✅	✅
L3V2 – Collaborators outside the data team	✅	✅	✅
L3V3 – Collaborators on the data team	✅	✅	✅
L4V1 – Introduction to large language models	✅	✅	✅
L4V2 – Choosing an LLM	✅	✅	✅
L4V3 – Prompting LLMs	✅	✅	✅
L4V4 – LLM limitations	✅	✅	✅
L4V5 – Demo: Interacting with LLMs (demo)	✅	✅	✅

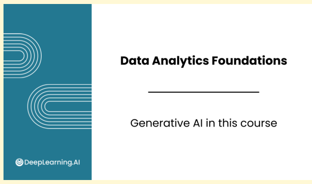
Introduction

L0V1 – Welcome to data analytics!

Visual	Script

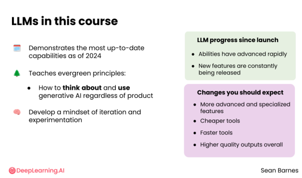
🎙️ TH 	Every **minute** of the day, people send 231 million emails, run 6 million google searches, and watch 400 thousand hours of Netflix content. That's an unimaginable amount of data, and that's just from a few sources. In this course, Data Analytics Foundations, you'll learn to wrangle this volume & complexity of raw data to help businesses make better decisions. Data analytics, and it powers insights across almost every industry, even ones you might not think of: from fashion and government, to tech, sports and healthcare. I'm Sean Barnes. I'm a data analytics leader at Netflix and I've worked with data in government, academia, and now the tech sector. Throughout all of these roles, the common thread was leveraging data to inform decision-making. This course is the first in a series designed to prepare you for an entry level data analyst role, and even if your goals are different, you'll be prepared for working with data in any career path. You don't need any prior experience with analytics software, programming, or even data to succeed in this course. This course is designed to take you from no prior experience to leading your own end to end data projects. And, if you're already working as a data analyst or in a similar role, you'll likely learn novel approaches for leveraging data and insights to advance in your career. Starting out, you'll learn what data is & the many forms it can take. Then, you'll get hands on with spreadsheets, a fundamental tool for analyzing and visualizing data. You'll explore real-world datasets throughout video demos and the interactive labs, including hotel bookings, baby names, and home sales. Finally, you'll learn a structured approach for data analytics projects that works across industries. Throughout this course, you'll use large language models to complement your analysis, which are being adopted across industries and changing the nature of work. They are not a replacement for you or your perspective, but they can augment your skills, serving as a thought partner for your practice. In this course, you'll use LLMs to interpret data visualizations, run analyses, and more. Data analytics is both analytical and creative, which is why I find it so fun. While you will calculate complex statistics, you'll also create compelling stories to inspire action. You'll discover new things every day, work with people from all backgrounds, and see the impacts of your expertise. So with that, I encourage you to enroll in and take this course.

L0V2 – Generative AI in this course

<table>
<tr><th>Visual</th><th>Script</th></tr>
<tr><td>🎙️ TH

</td><td>One of the key elements of this course is learning to use generative AI, in particular large language models or LLMs like ChatGPT, Claude, Gemini, and so on. I can't wait to share with you how these tools can fit into your work as a data analyst. Sometimes they feel like magic. And sometimes they can be quite frustrating. 😅

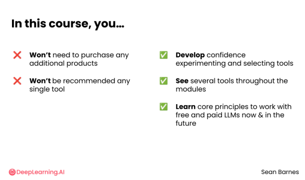
You'll learn how to use LLMs to
• Synthesize information from your stakeholders
• Explore a dataset and its metadata
• Automatically run data analysis by writing code for you
• Interpret images of data visualizations
• And create data visualizations

You'll also learn about LLMs' key limitations, including the tasks they can't do for you.

Teaching such a new technology has its challenges. I wanted to take a moment to share our team's philosophy about generative AI in this course.</td></tr>
<tr><td>
</td><td>First, this course demonstrates the **[CLICK]** most up-to-date capabilities as of mid 2024, and we expect changes in the coming months and years. Our team has designed this course to **[CLICK]** teach evergreen principles: how to **[CLICK]** think about and use generative AI in your work regardless of which specific product you end up working with. Above all, you will **[CLICK]** develop a mindset of iteration and experimentation.

The **[CLICK]** progress in LLM products since their launch in late 2022 has been astounding; their **[CLICK]** abilities have advanced rapidly, and **[CLICK]** new features are constantly being released. **[CLICK]** Here are some changes you should expect in the near future:
• First, genAI tools with **[CLICK]** more advanced and specialized features, like the ability to use apps for you
• Expect **[CLICK]** cheaper tools
• And **[CLICK]** faster tools
• And **[CLICK]** higher quality outputs overall

It can be difficult to follow all the changes in a rapidly progressing field like this - but don't worry! In this course, you'll develop the metacognitive skills you need to harness those advancements in your own work.</td></tr>
</table>

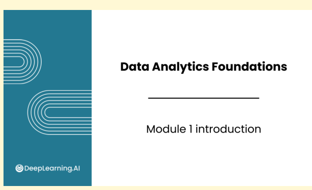
<table>
<tr>
<td>
</td>
<td>This course also demonstrates some paid features of LLMs, but you [CLICK] won't need to purchase any additional products to complete the assignments. It's important for you to see the available options, including paid options, so that you [CLICK] develop confidence experimenting and selecting the best tools in your work as a data analyst.

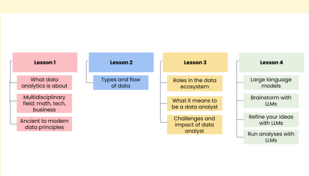
[CLICK] This course does not recommend any single tool. You'll see [CLICK] several throughout the modules. And remember that the [CLICK] core principles you'll learn will prepare you to work with LLMs both free and paid, now and in the future.</td>
</tr>
<tr>
<td>🎙️ TH</td>
<td>LLMs are an incredibly useful tool, but they are not a replacement for your skills and judgment. They can't replicate your decision-making ability in complex situations, especially when experience, intuition, and adaptive thinking are required.

You'll get your first taste of working with generative AI for data analytics in Lesson 4 of this module, including a hands-on lab. For now, join me in the next video to see all the exciting topics in this module. I'll see you there!</td>
</tr>
</table>

L0V3 – Module 1 introduction

<table>
<tr>
<th>Visual</th>
<th>Script</th>
</tr>
<tr>
<td>🎙️ TH




</td>
<td>Imagine a world where every decision is backed by solid evidence. A world where businesses can harness the power of data to drive efficiency, success, and innovation. Welcome to the world of data analytics!

~~In this first module, you'll uncover [CLICK] what data analytics is all about.~~ You'll see how this [CLICK] multidisciplinary field combines the problem-solving prowess of mathematics, the computational power of technology, and the strategic thinking of business to create an invaluable discipline. ~~But data analytics is more than just numbers and algorithms. It's about understanding the world around us in a whole new way.~~ [CLICK] From ancient civilizations tracking agricultural cycles to modern businesses optimizing their decision-making, the principles of data analytics have been shaping our world for thousands of years.

In this module, you'll gain a comprehensive understanding of the different [CLICK] types of data and how they flow through an organization. You'll meet the diverse cast of characters that make up the [CLICK] data ecosystem, each with their own unique skill set. ~~And you'll discover [CLICK] what it really means to be a data analyst – the [CLICK] challenges you'll face and the impact you can make. You'll start to see data in a new light, recognizing its potential to improve every industry.~~</td>
</tr>
</table>

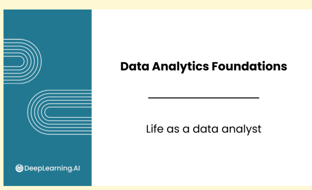
<table>
<tr><td></td><td>Finally, you’ll learn to harness a powerful tool ~~in your data analysis arsenal:~~ **[CLICK]** large language models like ChatGPT. These AI-powered tools can serve as a thought partner: helping you **[CLICK]** brainstorm, **[CLICK]** refine your ideas, and even **[CLICK]** run analyses.

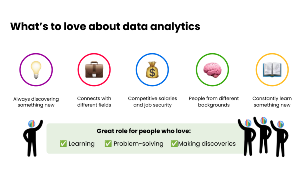
Whether you're aspiring to launch a career in data or looking to leverage analytics in your current role, this module will provide you with a rock-solid foundation. ~~You'll emerge with a deep appreciation for the power of data, a keen understanding of the data landscape, and an exciting glimpse into your own potential as a data analyst.~~

Next up is a glimpse of yourself in the data analyst role: what does a day in the life look like? How about a year? How about… a career? Follow me to the next video to take a look.</td></tr>
</table>

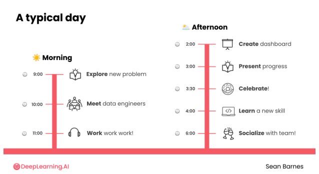
L0V4 – Life as a data analyst

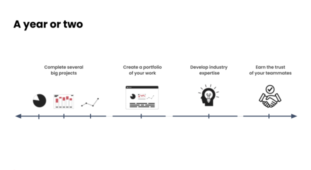
<table>
<tr><th>Visual</th><th>Script</th></tr>
<tr><td>🎙️ TH


</td><td>Of all the data analysts I’ve known in my career, every single one of them has been passionate about this work. As long as you’re in a company that treats you fairly, this job will become one of your joys. People often enter the field because it pays well, but they stay because, simply, they like it. It’s fun, it’s fast-paced… and it stays fresh.

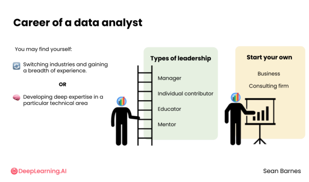
I want to share with you what I love about data analytics: on the daily, in the medium term, and as a career.</td></tr>
<tr><td>
</td><td>But first a highlight reel. My coworkers and I love data analytics because:
● You’re **[CLICK]** always discovering something new, which keeps the job fresh.
● It **[CLICK]** connects with so many different fields. Almost every industry needs data analysts, from tech giants and startups to the government and nonprofits.
● High demand for your skills translates into **[CLICK]** competitive salaries and job security.
● It attracts **[CLICK]** people from all kinds of backgrounds. You might work alongside former physicists, psychologists, or business majors who have all found their way into data.
● You’ll constantly be **[CLICK]** learning something new, since the field evolves quickly.

If you're the **[CLICK]** type of person who loves **[CLICK]** learning, **[CLICK]** problem-solving, and **[CLICK]** making discoveries, **[CLICK]** you'll fit right in.</td></tr>
</table>

<table>
<tr>
<td>
</td>
<td>Here’s a mock schedule of a typical day:
● 9:00 AM – Start your day by **[CLICK]** exploring a new problem you need to solve
● 10:00 AM – **[CLICK]** Meet with the data team. Catch up about the company’s new priorities and get your hands on the data you need
● 11:00 AM – **[CLICK]** Heads down work time. Get your hands dirty with spreadsheets, databases, and coding. Make new discoveries and even have a wow moment or two.
● 2:00 PM – **[CLICK]** Create a dashboard. Figure out how to tell the hidden stories in the data and visualize those stories in a beautiful, yet functional way.
● 3:00 PM – **[CLICK]** Present your progress on the dashboard. It feels great to share your hard work! Get direct feedback from your teammates about how this dashboard will help them deliver value.
● 3:30 PM – **[CLICK]** Celebrate your presentation and dashboard progress! Reward yourself with an afternoon beverage and a quick break. I sometimes go for a tea 🍵 🙃
● 4:00 PM – **[CLICK]** Learn a new technical skill. Take an advanced statistics course, or learn a new programming language. It’s so rewarding to learn on the job.
● 6:00 PM – **[CLICK]** Happy hour with the data team. Socialize with your fellow analysts and data scientists, learn about new trends, and discuss upcoming projects.

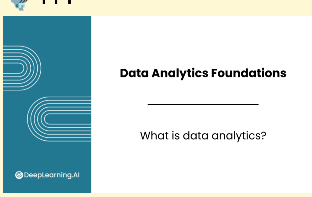
Your days will be action-packed. I encourage you to take time to celebrate and track your successes.</td>
</tr>
<tr>
<td>
</td>
<td>~~Each day is different, with new people to meet and problems to solve. How does that translate into the medium term? Here’s what a year or two could look like for you:~~
● Over the course of a couple of years, you’ll likely **[CLICK]** complete several big projects. Seeing the impact of your work in the real world is incredibly satisfying.
● You’ll be able to **[CLICK]** create a portfolio of your work. A portfolio not only showcases your skills but also helps you prepare for future growth opportunities.
● You will also **[CLICK]** develop your industry expertise, from terminology to unspoken rules. Your technical skills will improve significantly as different projects require you to upskill.
● And, by leading successful projects, you’ll **[CLICK]** earn the trust of your teammates. Building strong relationships creates opportunities to work on compelling problems.

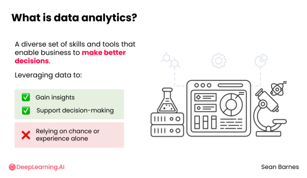
Each of these aspects of data analytics is rewarding in its own right. It may seem far away, but before you know it you’ll be celebrating big successes and mastering brand new skills.</td>
</tr>
</table>

	Days and years build careers. What does the career of a data analyst look like? ● **[CLICK]** You may find yourself **[CLICK]** switching industries, gaining a breadth of experience across different fields like tech, healthcare, fashion, supply chain, and more. ● Or, **[CLICK]** you can develop deep expertise in a particular technical area, becoming a go-to specialist. As you progress, you will become a leader. There are many **[CLICK]** types of leadership – as a **[CLICK]** manager, a highly skilled **[CLICK]** individual contributor, or an **[CLICK]** educator. ~~Being an educator doesn't just mean teaching in a classroom. You could create online courses or video content, establishing your brand and sharing your knowledge with a broader audience.~~ You can also develop **[CLICK]** mentorship relationships with younger colleagues to help them grow their skills. As you gain expertise, you might decide to **[CLICK]** start your own **[CLICK]** business or **[CLICK]** consulting firm, leveraging that expertise to help others succeed. The effort you put into each day on the job, each project, each conversation, opens up a rich variety of career trajectories in data analytics.
🎙️ TH	To me, data analytics is more than just a job. It's a joy. I value every learning opportunity I get, every chance to build something beautiful, and every conversation with my brilliant colleagues. I know you'll come to enjoy the field just as much as I do. With that goal in mind, join me in the next lesson to learn more about what data analytics is, its foundational mindsets, and its history. I'm excited for you to take this important first step in your data analytics career. See you there!

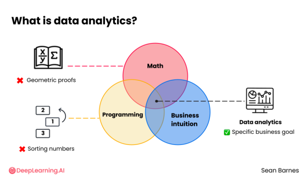
Lesson 1

L1V1 – What is data analytics?

Visual	Script
🎙️ TH 	Data analytics is used almost everywhere, making a meaningful impact on our lives in often invisible ways. The device you're using to watch this video, the clothes you're wearing right now, even the breakfast you ate this morning were likely influenced in some way by data analytics.



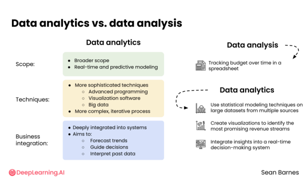
At its core, data analytics is a diverse set of skills and tools that **[CLICK]** enable businesses to make better decisions. It's all about **[CLICK]** leveraging data **[CLICK]** to gain insights and **[CLICK]** support decision-making, rather than **[CLICK]** relying on chance or experience alone.



Data analytics is a multidisciplinary field that combines **[CLICK]** math, **[CLICK]** programming, and **[CLICK]** business intuition. You're not just doing math for the sake of mathematical understanding, like developing a **[CLICK]** geometric proof, or programming for its own sake, like developing an algorithm for **[CLICK]** sorting a list of numbers. Data analytics harnesses math and programming together to **[CLICK]** achieve a specific business goal.

Data analytics has a lot in common with investigative roles, like a scientist, detective, journalist, or consultant.

- A **[CLICK]** scientist **[CLICK]** starts with a specific hypothesis, then **[CLICK]** analyzes data to evaluate the hypothesis.
- A **[CLICK]** detective **[CLICK]** gathers evidence and **[CLICK]** pieces it together to understand the crime.
- A **[CLICK]** journalist **[CLICK]** synthesizes information and **[CLICK]** creates a compelling narrative for a particular topic.
- ~~And a **[CLICK]** consultant **[CLICK]** helps influence clients to solve problems even if they don't have direct control over the outcome.~~

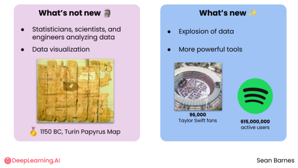


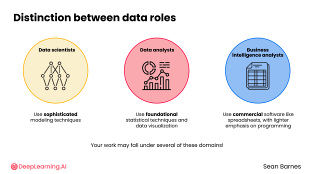
Data analytics sounds a lot like "data analysis", but these concepts differ in three key aspects: **[CLICK]** scope, **[CLICK]** techniques, and **[CLICK]** business integration.

- **[CLICK]** Data analytics has a **[CLICK]** broader **scope**, including **[CLICK]** real-time and predictive modeling that goes beyond retrospective analysis.
- **[CLICK]** Data analytics also requires **[CLICK]** more sophisticated **techniques**, including **[CLICK]** advanced programming, **[CLICK]** visualization software, and **[CLICK]** big data techniques, plus it often involves **[CLICK]** a more complex, iterative process.
- **[CLICK]** Data analytics is typically **[CLICK]** deeply integrated into business decision-making systems rather than used to answer one-off questions. It **[CLICK]** aims to **[CLICK]** forecast trends and **[CLICK]** guide decisions as well as **[CLICK]** interpret past data.

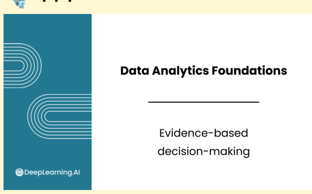
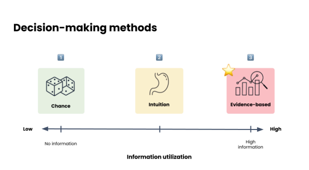
When a company **[CLICK]** tracks their budget over time in a spreadsheet, that's **[CLICK]** data analysis. When they **[CLICK]** use sophisticated statistical modeling techniques to analyze large datasets from multiple sources, **[CLICK]** create visualizations to identify the most promising revenue streams, and **[CLICK]** integrate these insights into a real-time decision-making system, **[CLICK]** that's data analytics.

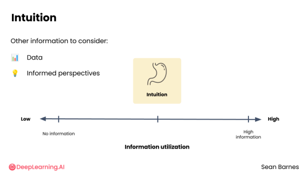
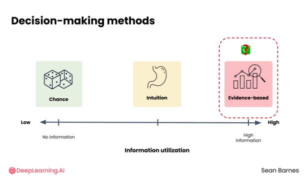
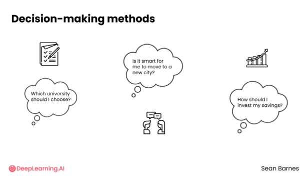
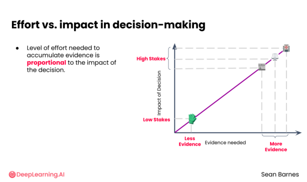
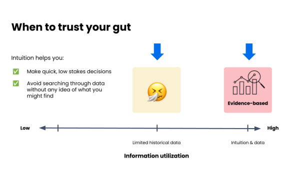
It might seem like data analytics is a brand new field that's only come about because of the recent acceleration in the tech industry. That's part of the

	story.
	Data **analysis** has actually been around for a while. Here's what's **[CLICK]** new and what's **[CLICK]** not new. **[CLICK]** Statisticians, scientists, and engineers have been analyzing data for a long time, and **[CLICK]** data visualization has been around for millennia. The first documented data visualization dates back to **[CLICK]** 1150 BC, almost 3,000 years ago. It's an ancient Egyptian map of gold mines called the Turin Papyrus Map. The ancient Egyptians were pretty savvy with data. ~~Check out the upcoming reading item to learn more.~~ What **is** new about data analytics is **[CLICK]** the explosion of the data itself. We're collecting more detailed data than ever before. Plus, computing has evolved in tandem, giving us **[CLICK]** more powerful tools to analyze that data. The ancient Egyptians weren't exactly working at the scale of **[CLICK]** 96,000 Taylor Swift fans packed into a stadium on their mobile devices, or **[CLICK]** 615 million monthly active Spotify users. And I doubt they were programming in Python. These trends define modern data analytics, with its expanded scope, more sophisticated techniques, and business integration. So, where can data analytics be useful? Well, that scope is nearly infinite!
🖥️ SC Use any one of these screencasts to highlight relevant jobs. Please try to reduce the amount of text on screen (e.g. by graying/blurring out irrelevant sections – **see this thread**), and try to pick the more interesting jobs if you can **Folder with screencasts**	If you check out some job postings, you'll see data analysts are in demand at tech companies, hospitals, sports teams, manufacturing plants – even conducting research in academic institutions, which is where I started in the field. While you're searching for job opportunities, you'll see postings for business intelligence analyst, data scientist, and other similar titles used interchangeably with data analyst. Truthfully, there's a lot of overlap, but I'll share some of the nuances.

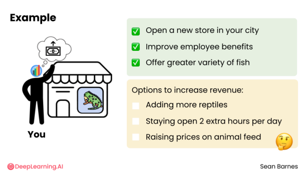
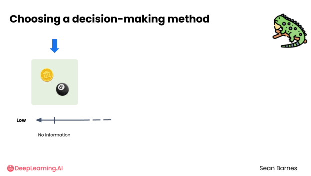
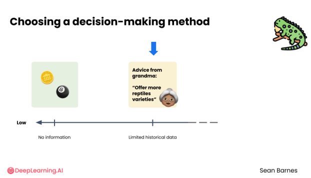
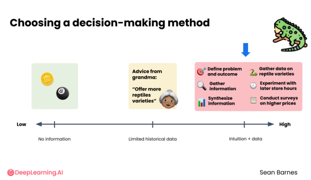
	Data scientists often **[CLICK]** lean more heavily into sophisticated modeling techniques, while data analytics is more inclusive of **[CLICK]** foundational statistical techniques and data visualization. **[CLICK]** Business intelligence analysts tend to use commercial software like spreadsheets, with a lighter emphasis on programming. Ultimately, these distinctions are somewhat arbitrary. The same title may correspond to different tasks at different companies. As a data analyst, you may find yourself drawn more to certain methods, and therefore **[CLICK]** your work may fall under one or several of these domains.
🎙️ TH	Data analytics rewards curiosity, problem-solving skills, and the ability to influence others. And of course, it's really rewarding to see the impact of your work in the real world. In this course, you'll explore the tools and techniques you need to start leveraging data to drive better decisions. Let's get right into a foundational concept: evidence-based decision-making!

L1v2 – Evidence-based decision-making

Visual	**Script**
🎙️ TH 	When it comes to making decisions, there are a lot of approaches you can take. You could just wing it, flip a coin, **[Sean flips coin]** ask a friend, maybe even shake the magic 8 ball. **[Sean actually takes out a magic 8 ball]** Should I use a magic 8 ball to make all my decisions in life? **[shakes and reads it]** Without a doubt. **[grimace/side eye, sets down magic 8 ball]** In contrast to this clearly terrible approach, data analytics is all about bringing evidence and consistency to decision-making. In this video, we'll talk about decision making by chance, by intuition, and the way that is most likely to consistently lead to success: with a combination of intuition and data.
	There are three basic ways you can make a decision: you can **[CLICK]** leave it to chance, go with the gut – this is also called decision-making **[CLICK]** by intuition – or you can practice **[CLICK]** evidence-based decision-making, which is where **[CLICK]** data analytics comes in. These different methods sit on a **[CLICK]** spectrum of information utilization, with chance relying on **[CLICK]** no information at all,evidence-based decision-making on the high **[CLICK]** information side, and intuition **[CLICK]** somewhere in the middle.

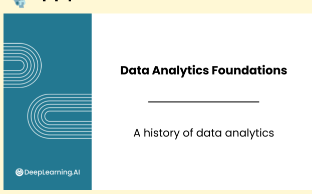
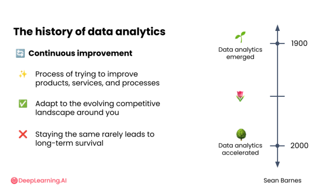
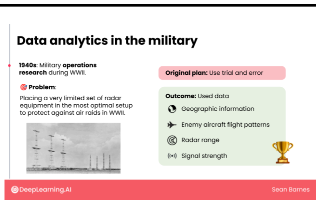
	Intuition is informed by your personal experience, so it's a valuable asset. **[CLICK]** But there are other important categories of information to consider, like **[CLICK]** data and **[CLICK]** the informed perspectives of others.
	~~So, more information is obviously better, right? **[CLICK]** Essentially, yes, though it's not quite as straightforward as that.~~ Let's first discuss the idea that information is needed for decision-making at all. While making informed decisions is probably important to you, I suspect that you typically don't go out formally defining your problems and gathering evidence in your daily life.
	But you may have made decisions similar to these: ● **[CLICK]** Which university should I choose? ● **[CLICK]** Is it smart for me to move to a new city, or stay where I am? ● **[CLICK]** How should I invest my savings? Think for a moment about the kinds of information you might gather to answer each of these questions. I doubt you'd be flipping a coin for any of these~~, or leaving it up to a gut feeling.~~ You might write out the **[CLICK]** pros and cons of each college, **[CLICK]** discuss the new city with a friend who lives there, or **[CLICK]** track the performance of different investments to decide which is best. So even if you don't realize it, you're likely making evidence-based decisions for lots of life's big questions.
	The level of effort needed to **[CLICK]** accumulate evidence is **[CLICK]** proportional to the **[CLICK]** impact of the decision. **[CLICK]** In fields such as **[CLICK]** criminal justice, **[CLICK]** medicine, or **[CLICK]** journalism, where decisions can have serious consequences, it's not enough to rely on opinion or personal experience alone. If you're seeing a doctor for your cold symptoms, you don't want her to just randomly guess that it's the flu. **[CLICK]** The higher the stakes, **[CLICK]** the more evidence you need to support your decision. Meanwhile, if you're deciding whether to **[CLICK]** recommend mittens instead of gloves to a customer, **[CLICK]** you don't need as much information, because **[CLICK]** the cost of making the wrong decision is low. No one's life, reputation, or health is at stake. Just their chance at finger coziness.
	~~Now~~, sometimes a gut decision is warranted. ~~I certainly don't do a full scale analysis for every decision I make. **[CLICK]** And~~ sometimes what looks like the flu is in fact the flu. **[CLICK]** Intuition isn't useless. In fact, one way of thinking about it is that you're essentially relying on **[CLICK]** limited historical data points. But, some intuitions are more valuable than others. Do you trust the gut instinct of a doctor who's treated 5 patients for the flu, or one who's treated 500, or

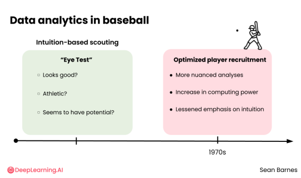
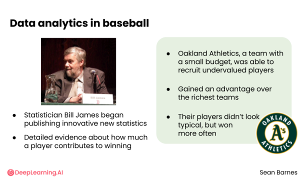
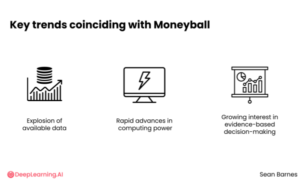
<table>
<tr><td></td><td>5000?

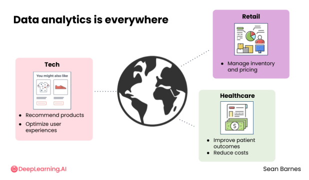
The most effective approach is when **[CLICK]** intuition is combined with data. That's why we say **[CLICK]** evidence-based decision making, since both data and to some degree intuition can be a part of your evidence. **[CLICK]** Intuition helps you make **[CLICK]** quick, low stakes decisions and to **[CLICK]** avoid situations where you're searching through a massive amount of data without any idea of what you might find. But you also don't want to rely on your gut all the time.</td></tr>
<tr><td>
</td><td>Let's look at an example. **[CLICK]** Say you run a small business – **[CLICK]** an exotic pet store – and you want to **[CLICK]** increase your revenue. If you are able to do this, you might be able to **[CLICK]** open a new store in your city, **[CLICK]** improve your employee benefits, or **[CLICK]** offer a greater variety of fish. You have a couple of options you're looking at for increasing revenue – say, **[CLICK]** adding more reptiles, **[CLICK]** staying open for 2 extra hours per day, or **[CLICK]** raising prices on animal feed – but **[CLICK]** how do you choose the best one? What information can you use to make your choice? This is a high stakes decision.</td></tr>
<tr><td>
</td><td>Well one option is to leave it to fate. Grab the ol' **[CLICK]** magic 8 ball and give it a shake, or flip a, uh, **[CLICK]** three sided coin. You're essentially making the decision with no extra information. But, there are consequences here, and I bet you can do better than that.</td></tr>
<tr><td>
</td><td>What would an intuition-based decision look like? Well, maybe **[CLICK]** grandma has been running this exotic pet shop since 1987, and she remembers times like these ,and she's absolutely certain that **[CLICK]** offering more reptile varieties is the way to go. This intuition uses a little more information than the magic 8 ball, since it's informed by some limited historical context. But, can you do even better than that?</td></tr>
<tr><td>
</td><td>Let's go the evidence-based route, which involves clearly **[CLICK]** defining the problem and the desired outcome, **[CLICK]** gathering relevant information, and **[CLICK]** synthesizing that information to identify the best decision. It's all about using the right information to make the right decision and hopefully achieve the right outcome. **[CLICK]** Gather the data on reptile varieties, **[CLICK]** experiment with keeping the store open longer, and **[CLICK]** conduct surveys to see if higher prices bother your customers. Maybe you can try the reptile variety approach first based on grandma's intuition, and if that doesn't pan out, investigate the other two options.</td></tr>
<tr><td>🎙️ TH</td><td>Sometimes you can make the wrong decision and still get the right outcome, or vice versa. There's a little luck in every decision. The goal of evidence-based decision-making is to maximize your chances of getting the right outcome by accumulating the most relevant and reliable information.</td></tr>
</table>

	It's funny to think about grandma steering her reptile business through a recession back in 1987, but a lot of businesses make decisions this way – based on intuitions like ● It feels right, or ● I've seen this done before or ● LLMs are so popular right now, let's use one **[Sean takes out a magic 8 ball]** Is that the best way to make decisions? **[shakes 8 ball]** Definitely not. **[nod/wink]** As a data analyst, you'll add real value to a business, and honestly, have the most fun, when you transform this type of thinking into evidence-based decision-making. Join me in the next video to explore a history of data analytics, which is a fascinating way to start acquiring some of the key mindsets of data analytics.

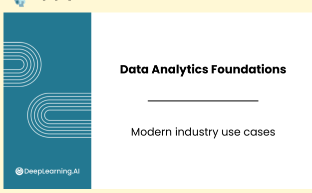
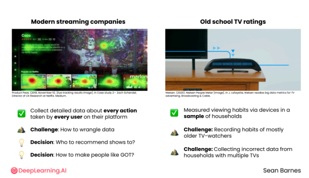
L1v3 – A history of data analytics

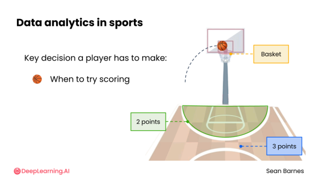
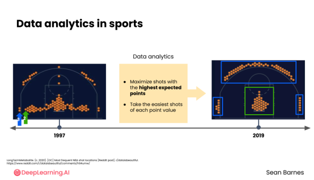
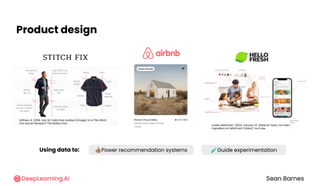
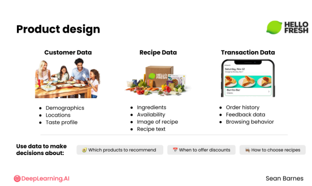
Visual	**Script**
🎙️ TH 	The recent history of data analytics is just as fascinating as the ancient Egyptians' use of data visualization. From the military, to baseball, to the tech sector, this video will help you understand why data analysts are so in demand. We'll look at two key trends that translate into data analytics mindsets: ● Continuous improvement ● And data-driven evidence
	Modern data analytics **[CLICK]** emerged almost 100 years ago and **[CLICK]** really **[CLICK]** accelerated at the turn of the century. At its core, the history of data analytics is grounded in the concept of **[CLICK]** continuous improvement, which is an ongoing **[CLICK]** process of trying to improve your products, services, and business processes. As a company, **[CLICK]** you have to adapt to the evolving competitive landscape around you. Just like in evolution, **[CLICK]** staying the same rarely leads to long-term survival.
 ([source]())	Modern data analytics has its roots in military operations research during WWII, around the early 1940s. Moving, feeding, and equipping an entire army is a massive operation, and each decision can have substantial consequences. One of the first documented operational research problems involved **[CLICK]** placing radar equipment in the most optimal setup to protect Britain against German air raids. The Allies could have used trial and error to place the equipment, but they **[CLICK]** didn't have time to run experiments because air

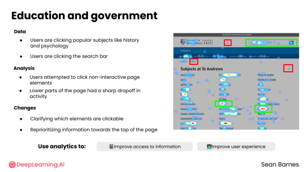
	raids were already happening. **[CLICK]** By using data such as **[CLICK]** geographic information, **[CLICK]** enemy aircraft flight patterns, **[CLICK]** radar range, and **[CLICK]** signal strength, the team developed an optimal strategy for placing the radar equipment to detect enemy aircraft. This placement is considered a major factor in **[CLICK]** the allied victory at the Battle of Britain. While it might not look exactly like modern data analytics – they certainly weren't working with big data or powerful computers – it's still a form of data-driven decision-making that has strongly influenced the field of data analytics today.
	American baseball was another historical hotbed of data-driven innovation. **[CLICK]** Prior to the 1970s, top players were typically chosen using a heavily intuition-based approach. **[CLICK]** Most scouts looking for players at that time relied on the so-called **[CLICK]** "Eye Test": basically, they just watched the players. **[CLICK]** Did they look good? **[CLICK]** Were they athletic? **[CLICK]** Did they seem to have potential? Scouts used some basic statistics too, but their subjective judgment dominated decision-making. This method was intended to build winning teams while also preserving the aesthetics of the sport. Then, in the 1970s, three factors gradually **[CLICK]** optimized player recruitment: **[CLICK]** more nuanced analyses, **[CLICK]** an increase in available computing power, and **[CLICK]** a lessened emphasis on intuition.
	Statistician Bill James began publishing innovative new statistics. Instead of just tracking points scored (called "runs" in baseball), these statistics provided **[CLICK]** more detailed evidence about how much a player contributes to winning overall. That could mean helping their teammates score or even preventing the opposing team from scoring. Using these statistics, the **[CLICK]** Oakland Athletics, a team with a small budget, was able to recruit undervalued players; that is, great players who were less expensive but still great. This strategy **[CLICK]** helped them gain a competitive advantage over some of the richest teams. Their players didn't look like your typical players, both in physical appearance and playing style. **[CLICK]** But their players won more often. That's the power of data analytics.
	The baseball story coincided with a few key trends: **[CLICK]** an explosion of available data, **[CLICK]** rapid advances in computing power, and **[CLICK]** a growing interest in evidence-based decision-making across industries. If a baseball team could use data to compete with rivals that had much bigger budgets, what could this mean for businesses in other industries?

	Today, data analytics is everywhere. Tech companies use it to **[CLICK]** recommend products and optimize user experiences. Retailers use it to **[CLICK]** manage inventory and pricing. Healthcare providers use it to **[CLICK]** improve patient outcomes and reduce costs. The list goes on ~~and on.~~
🎙️ TH [Source for 25% figure]()	The ubiquity of data analytics drives demand for experts in this role worldwide. In the US, data analytics jobs are expected to grow 25% by the year 2030. Every business can benefit from using data-driven evidence to fuel continuous improvement. Join me in the next video to see some of the most exciting applications of data analytics in modern times.

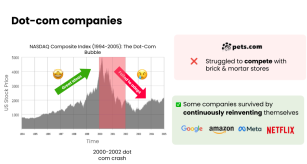
L1v4 – Modern industry use cases

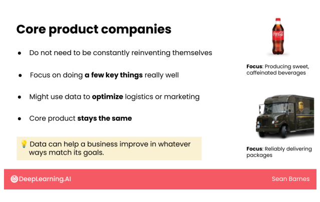
Visual	Script
🎙️ TH 	Whether you're passionate about sports, government, fashion, or something else entirely, there's a place for your expertise as a data analyst in that field. Let's take a tour through some of these applications.
 [Source for Nielsen bias]()	Modern day streaming companies collect **[CLICK]** incredibly detailed data about every action taken by every user on their platform: clicks, watch times, searches, pause and rewind actions. This information helps them recommend content. Their challenge is less about getting information and more about **[CLICK]** wrangling such a massive amount of it. Somehow all that data about who's clicking where and who's watching what needs to be translated into decisions like: **[CLICK]** who should we recommend reality shows to? How do we make people like **[CLICK]** Season 8 of Game of Thrones? And those decisions are impactful. YouTube famously revealed in 2018 that 70% of what people watch comes from recommendations. Contrast that approach to data with **[CLICK]** old school TV ratings. Networks used to rely on data collected by third party companies like Nielsen to understand different audiences. Nielsen **[CLICK]** measured people's TV-watching habits by installing physical monitoring devices in a sample of households to record what they were watching. While useful, this measurement system led to challenges like **[CLICK]** recording the habits of predominantly older TV-watchers and **[CLICK]** collecting incorrect data from households with multiple TVs.

 https://i.imgur.com/ZM5lxyj.png	Entertainment is just one application of data analytics, though. In sports, the analytics revolution that started with baseball has spread to basketball, both footballs, and beyond. Let's take a look at basketball, which I'll admit, I do follow 😉. **[CLICK]** Here's what the court looks like from above. The goal, also called the **[CLICK]** basket, is located here. **[CLICK]** One key decision a player has to make is when to try scoring. This is also called a "shot" in basketball. Getting the **[CLICK]** ball into the basket awards 2 points if the player shoots from inside **[CLICK]** this line, and **[CLICK]** 3 points if they were outside.
 Can take 1997 version – 2019 version side by side from here: https://interestingengineering.com/culture/epic-visualization-of-every-nba-shot-taken-since-1997	The past few decades have seen a dramatic shift in where players shoot from. **[CLICK]** On the left you can see the most frequent locations of shots taken in 1997. **[CLICK]** On the right, shots taken in 2019. **[PAUSE FOR 2 SECONDS FOR LEARNER TO COMPARE DIAGRAMS]** The pattern completely changed as teams embraced data analytics, **[CLICK]** maximizing shots from locations with the highest expected points per shot. Players have learned to **[CLICK]** take the easiest shots of each point value, so **[CLICK]** a lot of 3-point shots that hug the line, and **[CLICK]** 2-point shots as close to the basket as possible. For example, a shot from **[CLICK]** here is worth 2 points. But by taking one step back to **[CLICK]** here, the shot is essentially just as hard and worth 50% more. Why not go for it?
 	Let's look at product design. Increasingly, companies like **[CLICK]** Stitch Fix for clothing, **[CLICK]** Airbnb for accommodations, and **[CLICK]** HelloFresh for meal kits are using data to **[CLICK]** power recommendation systems and **[CLICK]** guide experimentation.
 I was using this video as a reference, no need to use it but as a reference https://youtu.be/F1pa	HelloFresh, as an example, collects several categories of data to inform its recommendations, such as: • **[CLICK]** Customer data like **[CLICK]** demographics, **[CLICK]** location, and **[CLICK]** taste profile; • **[CLICK]** Recipe data like **[CLICK]** ingredients, **[CLICK]** availability, and even **[CLICK]** images of the recipes and the **[CLICK]** recipe text itself; and • **[CLICK]** Transaction data like **[CLICK]**order history, **[CLICK]** feedback, and **[CLICK]** browsing behavior.

<table>
<tr>
<td>eoUHxxs?si=a2b3Dhm_quLW2ptw&t=1937</td>
<td>All this information can be **[CLICK]** used to make decisions about **[CLICK]** which products to recommend, **[CLICK]** when to offer discounts, and **[CLICK]** how to choose new recipes. ~~A traditional grocery store is unlikely to have the same level of insight into their customers and supply chain. That said, even those traditional stores are using data to recommend products and manage inventory.~~</td>
</tr>
<tr>
<td>

Here's a good example: https://digitalcommunications.wp.st-andrews.ac.uk/2021/01/07/how-we-are-using-hotjar-to-improve-our-web-pages/</td>
<td>In fields like education and government, there's a growing emphasis on using data to make better decisions. For example, these institutions may use analytics to **[CLICK]** improve access to information. The University of St. Andrews recently published an article explaining how it uses interaction heatmaps to **[CLICK]** improve user experience on its website. **[CLICK]** Here's a heatmap of user clicks on their university Subjects page. The red hotspots show where users were clicking the most, followed by yellow then green then blue. Data from the heatmap showed that **[CLICK]** lots of users are clicking on popular subjects like history and psychology as well as places like **[CLICK]** the search bar. **[CLICK]**

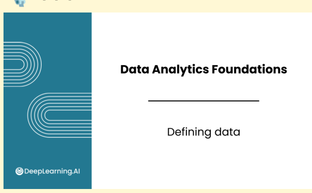
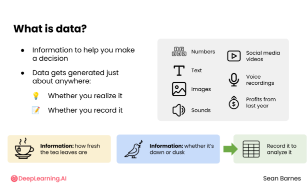
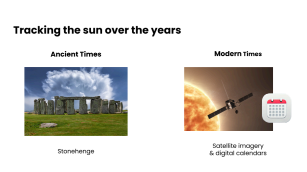
Analyzing the data, though, the University found that **[CLICK]** some users were attempting to click non-interactive elements, **[CLICK]** and that lower parts of the page had a sharp dropoff in activity. **[CLICK]** These insights drove website changes like **[CLICK]** clarifying which elements are clickable as well as **[CLICK]** reprioritizing information towards the top of each page.</td>
</tr>
<tr>
<td>
</td>
<td>~~What happens when you don't adopt the mindset of continuous improvement through data analytics? Well, the history of business is littered with examples of companies that couldn't figure out how to evolve beyond their initial concept. [CLICK] Here's a graph of time on the x axis and US stock prices on the y axis – higher is better; it means the economy is doing well.~~

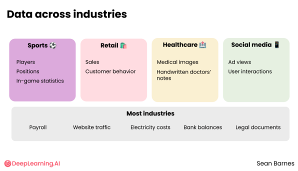
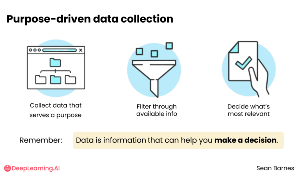
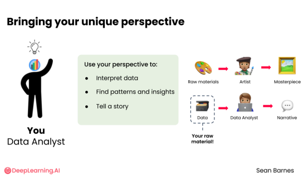
~~During the dot-com era, [CLICK] many startups with great ideas [CLICK] failed because they couldn't adapt. Many internet startups disappeared in the [CLICK] 2000-2002 dot-com crash. [CLICK] Pets.com for example, despite being a pioneer in online pet supplies – including product recommendations, surprisingly – [CLICK] struggled to compete with brick & mortar stores. But some companies [CLICK] survived by continuously reinventing themselves, like many of the biggest companies in the tech industry today: [CLICK] Google, [CLICK] Amazon, [CLICK] Meta, and [CLICK] Netflix, which have all transformed in significant ways from their original concept.~~</td>
</tr>
<tr>
<td>
</td>
<td>That said, not every company needs to constantly reinvent its core product. Some, like **[CLICK]** Coca-Cola and **[CLICK]** UPS, have built incredibly successful businesses **[CLICK]** by focusing on doing one core thing really well, **[CLICK]** Coca Cola by producing sweet, caffeinated beverages and **[CLICK]** UPS by reliably delivering packages. **[CLICK]** They might use data analytics to optimize</td>
</tr>
</table>

	logistics or marketing, but their **[CLICK]** core product remains largely the same. If the package makes it to the right destination quickly, what more does the customer really want? The point is **[CLICK]** data can help a business improve in whatever way matches its goals.
🎙️ TH	As a data analyst, you have the opportunity to help organizations in any industry leverage data to make better decisions. Your skills are in such high demand because they are valuable across so many applications. Great work completing the first lesson of this course. Once you’ve checked out the upcoming reading item and practice assessment, join me in the next lesson to discuss the lifeblood of data analytics: the data itself. You’ll learn what data is, where it comes from, and the many different forms it can take. I’ll see you there!

Lesson 2

L2v1 – Defining data

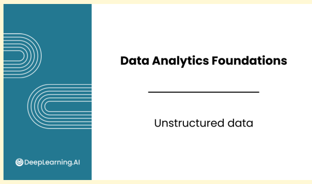
Visual	Script
🎙️ TH 	Let’s talk about the raw material that powers this field: data. Data is a very broad term. As a data **analyst**, you should think of data as any information that can help you make a decision.
	It comes in many forms, from **[CLICK]** numbers and **[CLICK]** text to **[CLICK]** images and **[CLICK]** sounds. **[CLICK]** Social media videos, **[CLICK]** voice recordings, **[CLICK]** profits from last year. **[CLICK]** All this information can help you make a decision. And **[CLICK]** data gets generated just about anywhere, **[CLICK]** whether you realize it, **[CLICK]** whether you record it. The taste of your **[CLICK]** morning cup of tea gives you information about how fresh the tea leaves are. That’s data. When you hear the **[CLICK]** sound of birds chirping, that might give you information about whether it’s dawn or dusk. That’s data too. You can take that data one step further and **[CLICK]** record it in order to analyze it.
	You saw in the last lesson how curiosity about data has ancient roots, but our ability to generate and capture data has massively accelerated in the last few decades. Thousands of years ago, ancient peoples tracked the position of the sun to determine the best times for planting and harvesting, but they had to do so using rock structures, like **[CLICK]** Stonehenge. Now, we can do the

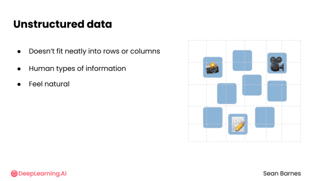
	same thing with way less effort through **[CLICK]** satellite imagery and digital calendars. Note the distinct lack of 25 ton stones.
	Different industries generate different types of data. **[CLICK]** In sports, you might work with highly structured data about **[CLICK]** players, **[CLICK]** positions, and **[CLICK]** in-game statistics. **[CLICK]** In retail, you're likely to encounter transactional data about **[CLICK]** sales and **[CLICK]** customer behavior. **[CLICK]** Healthcare data often includes unstructured information like **[CLICK]** medical images and **[CLICK]** handwritten doctors' notes, while **[CLICK]** social media platforms collect data about **[CLICK]** ad views and **[CLICK]** user interactions. **[CLICK]** Most industries will also have **[CLICK]** payroll data, and **[CLICK]** website traffic, and **[CLICK]** electricity costs and **[CLICK]** bank balances and legal documents and … a lot more. Nowadays it can sometimes feel like everything that can be tracked is tracked.
	But here's the thing: just because you can collect data about something doesn't mean you should. You should only **[CLICK]** collect data that serves a purpose. **[CLICK]** Remember our definition: as a data analyst, **[CLICK]** data is not just information, **[CLICK]** it's information that can help you make a decision. Your job is to **[CLICK]** filter through all the available information **[CLICK]** and decide what's most relevant for the problem at hand.
	As a data analyst, you also bring a **[CLICK]** unique perspective to data. You don't just consume it; **[CLICK]** **[CLICK]** you interpret it, **[CLICK]** you find patterns and insights, and you use it to **[CLICK]** tell a story. ~~It's a little cheesy, but I genuinely believe that working with data is like creating art.~~ Just as an **[CLICK]** artist uses **[CLICK]** raw materials like clay or paint to create a **[CLICK]** masterpiece, **[CLICK]** you use **[CLICK]** data to craft **[CLICK]** a narrative that informs and inspires. **[CLICK]** Data is **your** raw material, and you can create something beautiful and functional with it.
🎙️ TH	Data is a powerful tool for driving impact, whether you're trying to analyze customer behavior, interpret medical imaging, or recommend videos. Let's get more specific about data. Something this broad inevitably has classifications. In the next video, you'll learn the intricacies of unstructured data, which is a very natural and human way of capturing information. I'll see you there!

L2v2 – Unstructured data

Visual	**Script**

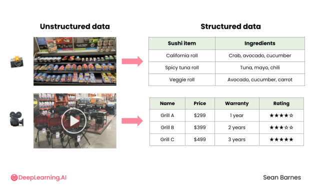
<table>
<tr>
<td>🎙️ TH w/ an overlay of a table or spreadsheet


</td>
<td>Close your eyes for a second – no, like actually do it with me [sean closes eyes] – and picture data. [pause for 2 seconds]

Okay, open your eyes. What did you think of? Was it something like this [show small overlay of a spreadsheet lookin thing]: a table of numbers in neat rows and columns? It's the most stereotypical depiction of data. But the reality is, a lot of the data you will encounter in the real world doesn't start out that way. It's what we call unstructured data.</td>
</tr>
<tr>
<td>
</td>
<td>To get more specific, "unstructured data" means data that [CLICK] doesn't fit neatly into rows or columns. It's all around us: [CLICK] when you snap a photo, [CLICK] record a video, or [CLICK] jot down notes in a journal, you're creating unstructured data. [CLICK] These are human types of information, and [CLICK] they feel very natural to us.

In fact, if you're just a person collecting data for yourself, you'll probably do it in an unstructured way without even thinking about it.</td>
</tr>
<tr>
<td>
</td>
<td>Maybe you keep a journal to track your workouts with friends. It might look something like this: [CLICK] [CLICK] "Today there were 4 people in our running group, [CLICK] we ran 3.2 miles in 29 minutes and 32 seconds, and [CLICK] afterwards I focused my strength training on my core."

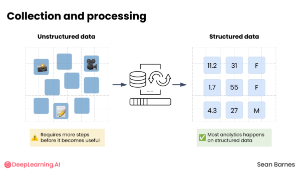
This information could be organized in a structured way too, like in [CLICK] this table where each exercise detail is a column and each day is a row. You can see the one on the left is [CLICK] more human and natural, while the one on the right is [CLICK] more rigid, but better for analysis.

So, to summarize, text data like your original journal entry is considered unstructured. [brief pause]</td>
</tr>
<tr>
<td>

Sushi image – https://i.imgur.com/kBYhmwX.jpeg</td>
<td>Here are some more examples of unstructured data: earlier this week, I took a photo of the [CLICK] sushi options at the grocery store so my wife could pick one – that's unstructured. If it was structured, maybe it would be more like a [CLICK] menu, with each item and its ingredients listed neatly.

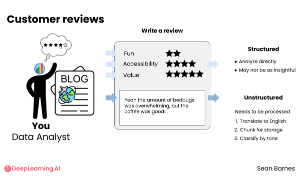
I also recorded a video of [CLICK] several potential grill options at the hardware store, which is unstructured data as well. If I had to put it in rows and columns, what would I put? Maybe I could [CLICK] record each option in a row, including the price, warranty length, and a score from 1 to 5 representing how much I liked each one.

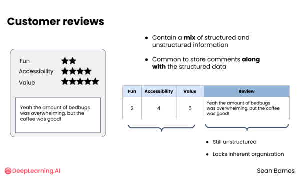
To summarize, [CLICK] photos and [CLICK] videos are unstructured. <s>So is text. When you write an email or an essay, that's unstructured data too.</s></td>
</tr>
</table>

<table>
<tr>
<td>
</td>
<td>Why does this distinction between structured and unstructured data matter? It's all about how the data gets collected and processed.

At some point, all of this **[CLICK]** unstructured data usually needs to be converted into **[CLICK]** a structured format to be analyzed effectively. **[CLICK]** Most data analytics is happening on structured data, although modern techniques are getting better at analyzing unstructured data directly. We'll talk more about structured data in the next video.

Since most data doesn't start out in rows and columns, it often undergoes **[CLICK]** a transformation into a structured form.

As a data analyst, you should be mindful of the origins of your data, because whether it's structured or not often affects how easy it is to preprocess and analyze. Unstructured data typically **[CLICK]** requires more steps before it becomes useful. But it's also more natural for a person to generate and can contain unexpected insights due to its level of detail.</td>
</tr>
<tr>
<td>
</td>
<td>Let's say you're working with customer reviews from a travel blogging site - my aspirational job. **[CLICK]** Reviewers can rate locations 1 through 5 on **[CLICK]** fun, **[CLICK]** accessibility, and **[CLICK]** value, then leave a **[CLICK]** comment.

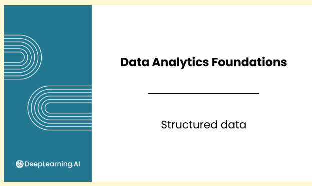
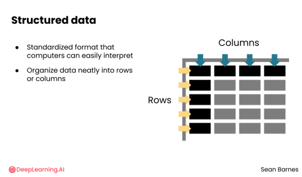
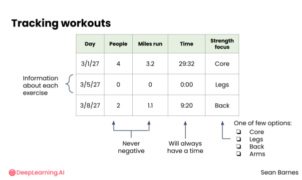
The comment data is being generated in an **[CLICK]** unstructured way – it's just freeform text. Behind the scenes, that text will **[CLICK]** need to be processed with steps like **[CLICK]** 1) translating to English **[CLICK]** 2) chunking for storage in a database and **[CLICK]** 3) classifying by positive or negative tone. **[CLICK]** Meanwhile the 1 through 5 ratings can be **[CLICK]** analyzed pretty directly, say, to find an average rating for each location, but **[CLICK]** may not be as insightful as the comments, where reviewers can offer insights like **[CLICK]** "yeah the amount of bedbugs was overwhelming, but the coffee was good."</td>
</tr>
<tr>
<td>
</td>
<td>These reviews contain a **[CLICK]** mix of structured and unstructured information. Even though the comments are unstructured, it's **[CLICK]** common to store them along with the structured data.

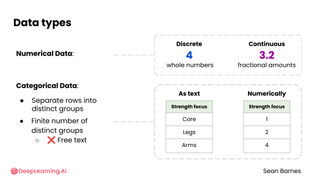
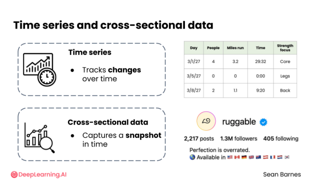
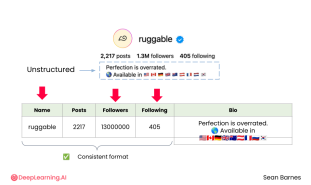
For example, you could store your review data like **[CLICK]** this, with each row representing a review, **[CLICK]** numbers for the different ratings, and then a **[CLICK]** column that contains the comment. This comment is **[CLICK]** still unstructured because it **[CLICK]** lacks an inherent organization; people can write whatever they want.

More on this technique in the next video.</td>
</tr>
<tr>
<td>🎙️ TH</td>
<td>Unstructured data is a natural byproduct of how we capture and communicate information as humans. It's a raw material that needs to be</td>
</tr>
</table>

	organized before it can generate insights. In the next video we'll discuss structured data, and why it even exists in the first place. I'll see you there!

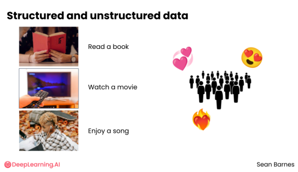
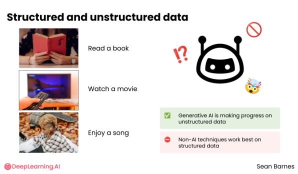
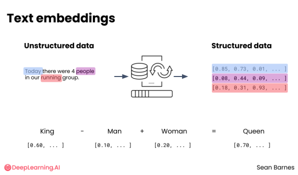
L2v3 – Structured data

Visual	Script
🎙️ TH 	When working with computers, you'll often find that you need to impose some structure on your data. Computers work most efficiently when data is organized in a predefined way, whereas humans are much more adaptable when information comes in an unexpected form. One big idea I hope you take away from this video is that structured data essentially exists for computers to store, process, and analyze.
	Structured data is all about organizing information into a **[CLICK]** standardized format that computers can easily interpret. Most commonly, that looks like **[CLICK]** organizing data neatly into **[CLICK]** rows and **[CLICK]** columns. Interestingly, that organization itself contains a lot of information.
Show the previous workout table 	Let's revisit the example of tracking your workouts – **[CLICK]** here's the structured version of that information again, and I added **[CLICK]** two more rows so you can compare workouts. In each column, you – or more accurately your computer – can expect the same kind of information. **[CLICK]** The time column will always have a time. **[CLICK]** The strength focus column will always have one of a few options like core, legs, back, and arms. There will never be a neck day. Although, to each their own I guess. **[CLICK]** Each row, or day, will contain information about each exercise, even if you didn't run at all. **[CLICK]** Miles run and people will never be negative. These examples represent some of the information built into the organization of this data.

	The mission of building in information often involves categorizing data into specific types, such as [CLICK] numerical or [CLICK] categorical. Numbers can be either [CLICK] discrete or [CLICK] continuous, with discrete meaning a [CLICK] whole counting number. For example, there's no such thing as a person and a half. Either you went running with one person or two. Continuous means you can have [CLICK] fractional amounts as well. For example, you can run a fraction of a mile, like 3.2 or 1.1. You also have specialized numerical formats such as time, which can also be represented in discrete or continuous forms. Categorical data enables you to [CLICK] separate rows into distinct groups. [CLICK] Core, legs, and arms are categories of exercises that could be used to analyze different types of strength focus. Perhaps people tend to skip leg day after all? 🙂Categorical data is most often represented as [CLICK] text, but there should be a [CLICK] finite number of distinct groups. Free text, like review comments, has potentially infinite values and would [CLICK] not make for useful categorical data without further processing. Categorical data can also be represented [CLICK] numerically, with a discrete number for each group. For example, maybe for efficiency's sake you just write [CLICK] 1 for Core, 2 for legs, 3 for back. Still categorical even though the representation is with numbers.
	One other key distinction in structured data is between time series and cross-sectional data. [CLICK] Time series data [CLICK] tracks changes over time, while [CLICK] cross-sectional data [CLICK] captures a snapshot in time. [CLICK] The workout data you just saw is considered [CLICK] time series data, since you'll be able to analyze your mile times and strength focus over time, plus monitor whether your friends are keeping their promise to train with you! On the other hand, check out [CLICK] this Instagram bio, which has data like the number of posts and followers, plus username, image, and a text bio. Is this cross-sectional or time series data? [1 second pause] [CLICK] This is cross-sectional, since it shows information about an account at one moment in time. You wouldn't be able to look at this data and say anything about followers over time or how often this person changes their profile picture.
	You learned earlier that it's common to see unstructured data stored inside a table or spreadsheet. Here's another example using the same instagram bio you just saw. You could represent the structured data from the bio in a table with columns for [CLICK] name, [CLICK] posts, [CLICK] followers, and [CLICK] following. Each of these columns has a [CLICK] consistent format. They can be processed by a computer. For example, you can sort and search [CLICK] names, which have a consistent format, or calculate the ratio of [CLICK] followers to

<table>
<tr><td></td><td>following.

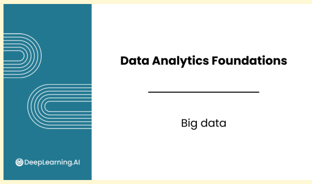
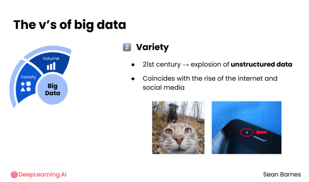
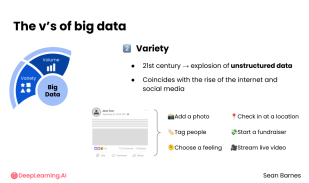
The bio, [CLICK] this description here, is [CLICK] unstructured, since it's unorganized text data that can't be easily processed by a computer. To keep all this data together, you can [CLICK] tack bio onto this table. It's still unstructured data, and will still need more work to process compared with the other columns, but you can store it together with the other data just to keep things consistent.

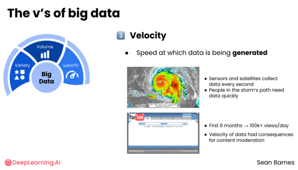
To summarize, different columns in a table can be structured or unstructured.</td></tr>
<tr><td>
</td><td>Now that you've seen the core components of both structured and unstructured data, let's take a step back and think about the difference between these two types of data from a human perspective. As humans, we're pretty good at interpreting unstructured data. We can [CLICK] read a book, [CLICK] watch a movie, or [CLICK] enjoy an emotional song without much effort.</td></tr>
<tr><td>
</td><td>But for computers, it's a different story. [3 clicks over like 1 second] Computers need data to be organized in a specific way in order to process it effectively. Although [CLICK] generative AI is making significant progress on interpreting unstructured data, in general terms, [CLICK] non-AI techniques are going to work best on structured data.</td></tr>
<tr><td>
</td><td>Here is an example of a technique data analysts use to transform unstructured text data into structured numerical data: text embeddings. This technique is a way of representing the [CLICK] meaning of words using a [CLICK] list of numbers. Each word gets its own list of numbers, and each position in that list represents some aspect of each word's meaning. Words with similar meanings will have similar numbers. Having a numerical representation of word meanings helps us perform cool operations. Here's an example, [CLICK] king minus [CLICK] man plus [CLICK] woman equals what? [2 second pause] [CLICK] Queen! Text embeddings make mathematical operations on word meanings possible. This strategy is often used in generative AI applications like ChatGPT.</td></tr>
<tr><td>🎙️ TH</td><td>So to recap, structured data is all about organizing information in a way that computers can use effectively. As a data analyst, you'll frequently derive insights from both structured and unstructured data.

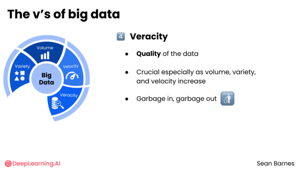
In the next video, you'll explore big data. Big data isn't just a lot of data! I promise there is more to it. Join me in the next video to learn more.</td></tr>
</table>

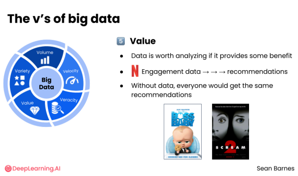
L2v4 – Big data

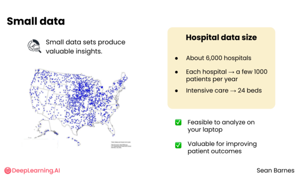
Visual	**Script**

🎙️ TH 	You may have heard the term "big data" thrown around. But what does that actually mean? You might hear the term and think, well that just means dealing with large volumes of data. That's certainly part of it, but there's more to it than that. Big data is defined by three key attributes, known as the three Vs: volume, variety, and velocity.
 (source)	Let's start with volume. This is probably the most straightforward characteristic of big data. **[CLICK]** Data sets nowadays are often so large, **[CLICK]** they pose significant challenges in terms of storage and computation. Think about a company like **[CLICK]** Amazon and the sheer **[CLICK]** volume of orders they process, around **[CLICK]** 12 to 19 million of them every day. In fact, since you started watching this video, Amazon has probably processed more than 6,000 orders. Volume matters because **[CLICK]** storing and analyzing that data requires some serious computational power. If you were working at Amazon and wanted to analyze transactions even from a single day, you wouldn't be able to do so on your laptop at home. Nor would you be able to manually copy and paste the transaction data from one location to another.
	Then there's variety. In the past, the data that analysts worked with tended to be structured, meaning it fit neatly into databases or spreadsheets. But the 21st century has seen an **[CLICK]** explosion of unstructured data, like images, text, video, and even augmented reality data from products like the Apple Vision Pro. This explosion **[CLICK]** coincides with the rise of the internet and social media in particular. There are way more **[CLICK]** selfies now than about 20 years ago because the first smartphone with a **[CLICK]** front-facing camera only came out in 2010.
	On a platform like Facebook, for example, when a user goes to **[CLICK]** create a new post, they can **[CLICK]** add a photo, **[CLICK]** tag people, **[CLICK]** choose a feeling, **[CLICK]** check in at a location, **[CLICK]** start a fundraiser, or **[CLICK]** even stream live video. Each of these post types requires its own methods for preprocessing and analysis. If you want to answer a seemingly simple question like "what does a particular user typically post about?" you would need to analyze an extremely diverse set of data.

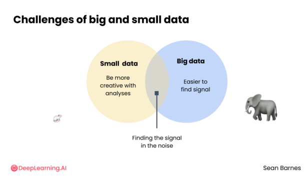
<table>
<tr>
<td>


</td>
<td>
The third V is velocity. [CLICK] This refers to the speed at which data is being generated. You saw the rate at which Amazon must process orders a moment ago, but it's not just a tech thing. During a [CLICK] hurricane, for example, [CLICK] sensors and satellites collect massive amounts of data every second, which must be rapidly analyzed to predict the hurricane's movement pattern. If analysts can't wrangle that data quickly, [CLICK] people in the storm's path could get delayed information.
The sheer velocity of data on social media in particular is staggering. [CLICK] Within YouTube's first six months, [CLICK] the site was getting over 100,000 video views per day. Soon, so many videos were uploaded that human moderation simply wasn't feasible, and YouTube turned to automated techniques. [CLICK] In other words, the velocity of data had serious consequences for how YouTube moderated its content, which has had continued ripple effects on content moderation today.
</td>
</tr>
<tr>
<td>


</td>
<td>
So that's the original three Vs. And you could stop there. But there's a bit of a movement to add even more V's to the framework. While the first three Vs you just saw are the most important factors that differentiate big data from what you might call small data, let's take a look at the additional V's.
The fourth V is [CLICK] veracity. This refers to [CLICK] the quality of the data. [CLICK] And it's a crucial consideration, especially as the volume, variety, and velocity of data increase. Is the data coming from a trustworthy source? Could it have been corrupted along the way? [CLICK] As the saying goes, garbage in, garbage out: if your data is of poor quality, your insights and consequently your business decisions will also be poor.
</td>
</tr>
<tr>
<td>


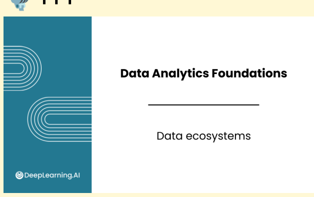
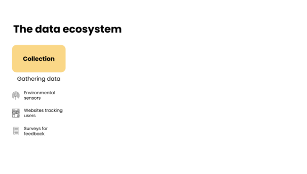
</td>
<td>
The fifth V is value. [CLICK] The idea here is that data is only worth analyzing if it actually provides some benefit. [CLICK] At Netflix, for example, the vast amount of engagement data we collect feeds into the recommendation system, allowing for personalized suggestions. [CLICK] Without that data, everyone would just get the same generic recommendations. It would be like that Netflix account your whole family shares; you know, the one where [CLICK] Boss Baby is recommended right next to [CLICK] Scream 2.
</td>
</tr>
<tr>
<td>


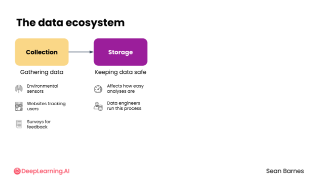
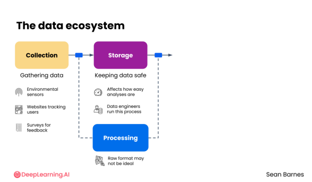
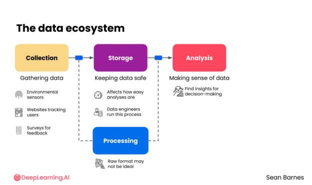
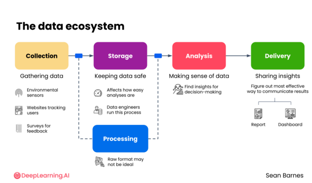
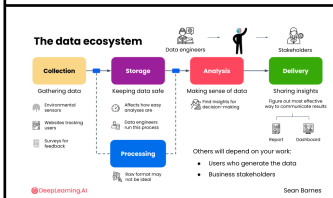
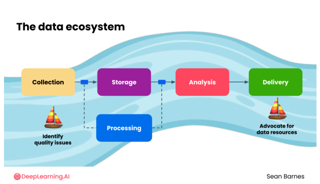
(source)
</td>
<td>
While big data is prevalent in today's world of data analytics, there are plenty of situations where relatively [CLICK] small data sets produce valuable insights. It might surprise you to learn that there are only about [CLICK] 6000 hospitals in the US. That's not a lot compared to the 290,000 Tinder matches every minute. Each hospital might only serve a [CLICK] few thousand patients each year. An intensive care unit might only have a [CLICK] couple dozen beds. It's perfectly [CLICK] feasible to analyze that data on your laptop. And the data generated in these contexts can still be [CLICK] incredibly valuable for improving patient outcomes.
</td>
</tr>
</table>

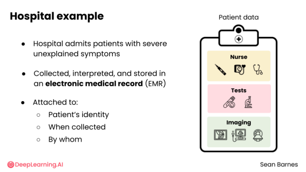
 	Working with **[CLICK]** big data and **[CLICK]** smaller data can feel like dealing with two **[CLICK]** completely different animals. But in both cases, the challenge is about **[CLICK]** finding the signal in the noise. In other words, finding the insights that matter. **[CLICK]** With big data, the volume may make it easier to find the signal, but you have to know where to look. With **[CLICK]** small data, you may have to be more creative in trying many different analyses to find any signal at all.
🎙️ TH	As a data analyst, your job is to consider the data in the context of the problem you're trying to solve. Sometimes that will mean working with massive, complex data sets. Other times, it will mean investigating a smaller, more focused data set. Great work completing this lesson! You've seen the many types of data you'll be working with as an analyst. In the upcoming practice lab, you'll work with both structured and unstructured data in an e-commerce case study for a gift shop. I know you'll enjoy it! Once you've finished the practice lab and practice assessment, join me in the next lesson to see how data fits into the bigger picture of an organization. We'll be talking about data teams and data ecosystems. Intriguing! I'll see you there.

Lesson 3

L3v1 – Data ecosystems

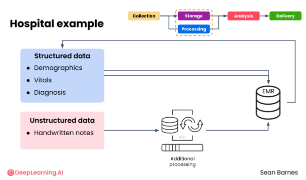
Visual	**Script**
🎙️ TH  	Just like electricity, data doesn't stay at the place it was generated; it flows through various systems to eventually power insights. Let's take a high-level look at that flow, which is called the data ecosystem: the end-to-end process that data undergoes, from the moment it's generated to the point where you can use it to drive decision-making.
 	~~Here's a high-level overview of the data ecosystem; this is the fundamental flow of data in most organizations:~~ ● **[CLICK]** First, collection: **[CLICK]** Gathering data. As you saw in earlier videos, data needs to be captured in order to be used effectively. This could look like **[CLICK]** sensors collecting environmental data, **[CLICK]** websites tracking user interactions, or **[CLICK]** surveys gathering customer feedback.

	• Storage: **[CLICK]** Keeping data safe. The way it is stored **[CLICK]** affects how easy different analyses are to perform. **[CLICK]** Data engineers typically run this process – you'll learn more about their role in a moment.
	• Processing: In most cases, the **[CLICK]** raw format in which data is collected may not be ideal for storage or analysis. So processing can actually happen **[CLICK]** between the Collection and Storage stages as well as **[CLICK]** between Storage and the next stage:
	• Analysis: **[CLICK]** Making sense of data. You will **[CLICK]** investigate the data to **[CLICK]** find insights that can inform decision-making.
	• Delivery: **[CLICK]** Sharing insights. You'll need to **[CLICK]** figure out the most effective way to communicate the results of your analysis, such as in **[CLICK]** a report or **[CLICK]** dashboards.
This text was recorded but should be cut from the final videos	~~Unless you're at a small company, **[CLICK]** you're unlikely to own data collection and storage. **[CLICK]** The heavy lifting is done by the data and software engineers that manage those systems. Life would be tough without them! **[CLICK]** It's crucial to understand these processes so you know **[CLICK]** what data you can request and **[CLICK]** how it has been transformed.~~
	The core data team, including you as a data analyst, will own these steps. **[CLICK]** Others will depend on your work, including any **[CLICK]** users who generate the data, plus **[CLICK]** business stakeholders like product managers, engineers, and so on. **[CLICK]** You are most accountable to your **[CLICK]** data engineers and **[CLICK]** business stakeholders, since those are the roles directly upstream and downstream. More on this in the next videos.
	You should view yourself as an active participant in the data ecosystem, not a passive consumer. You might **[CLICK]** identify quality issues in the data that need to be addressed upstream. Or you might **[CLICK]** advocate for additional data resources with your business stakeholders. Communicate both your needs and your insights to the roles adjacent to you. Be open to their needs and insights as well.

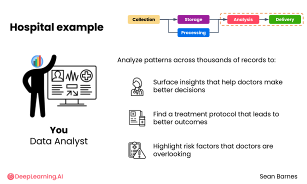
<table>
<tr>
<td>


</td>
<td>
Let's make this concrete with an example. [CLICK] Consider a hospital that admits patients with severe unexplained symptoms for diagnosis. Each patient generates lots of [CLICK] data about their health, and the initial diagnosis phase involves capturing it.

For instance, a [CLICK] nurse will measure their vitals: [CLICK] temperature with a digital thermometer, [CLICK] blood pressure and pulse with a cuff, and [CLICK] breathing patterns with a stethoscope. [CLICK] They may also perform a [CLICK] blood or [CLICK] urine test, which has to be processed in a laboratory. Or they might order [CLICK] imaging, such as an [CLICK] x-ray, [CLICK] ultrasound, or [CLICK] an MRI, which generate images that must be interpreted by an expert.

This data is stored in an [CLICK] electronic medical record or EMR [CLICK] attached to [CLICK] the patient’s identity, along with data about [CLICK] when the information was collected and [CLICK] by whom.
</td>
</tr>
<tr>
<td>


</td>
<td>
The way data is stored and processed depends a lot on its type. [CLICK] Structured data, like [CLICK] a patient's demographic information [CLICK] or vitals, might be easily [CLICK] stored in a traditional database. [CLICK] Unstructured data, like [CLICK] a doctor's handwritten notes, might be stored in a database after [CLICK]additional processing, such as handwriting recognition with AI or manual data entry.

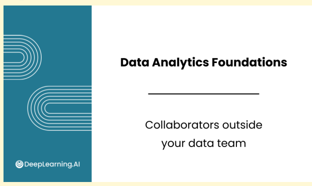
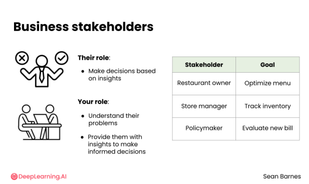
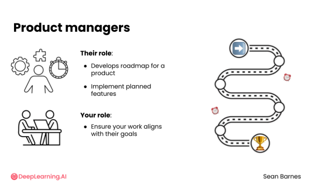
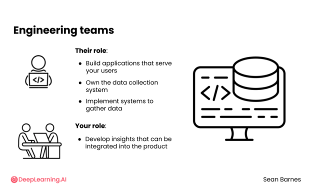
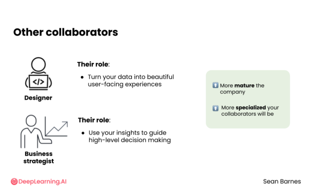
After all of this data flows [CLICK] into the EMR, the patient’s doctor can combine it with their own expertise to make a [CLICK] diagnosis… which itself becomes yet another [CLICK] data point captured in the EMR.
</td>
</tr>
<tr>
<td>


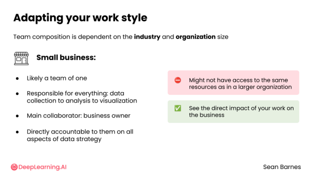
</td>
<td>
At this point, you as the data analyst come in. You won’t diagnose any single patient, but [CLICK] by analyzing patterns across thousands of similar patient records, you might be able to [CLICK] surface insights that help doctors make better decisions. [CLICK] Maybe there's a treatment protocol that's consistently leading to better outcomes. [CLICK] Or maybe there are certain risk factors that doctors are overlooking.
</td>
</tr>
<tr>
<td>This text was recorded but should be cut from the final videos</td>
<td>
<s>The power of the data ecosystem is that it's a feedback loop, not just a linear process. [CLICK] The insights you uncover in the analysis stage can help refine earlier stages, [CLICK] leading to better data and therefore better decisions over time. To reemphasize that point, the more you [CLICK] improve your data ecosystem, [CLICK] the more informed decisions you can make, and [CLICK] the more you can improve the business overall.</s>

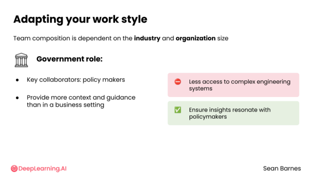
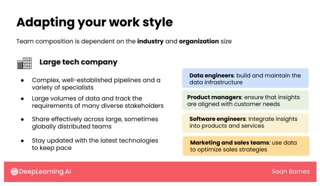
<s>Using the hospital example, you might [CLICK] advocate for collecting more detailed patient demographics. These detailed demographics [CLICK] may help you discover which groups of patients are diagnosed earliest. Considering</s>
</td>
</tr>
</table>

	~~your feedback, your team can design a system to collect that demographic information,[CLICK] leading to improved decision-making.~~
🎙️ TH	You're not alone in this process. In the next two videos, you'll meet your key collaborators, from business stakeholders who read your reports to the data engineers you talk with daily. See you there!

L3v2 – Collaborators outside your data team

Visual	**Script**
🎙️ TH 	Data work involves people from every team within an organization. Let's take a look at some of your key collaborators from outside the data team.
	First, your business stakeholders, who **[CLICK]** make decisions based on the insights you provide. They could be anyone from **[CLICK]** a restaurant owner trying to optimize their menu, **[CLICK]** to a store manager tracking inventory, **[CLICK]** to a policymaker evaluating a new bill. **[CLICK]** Your job is to understand their problems and **[CLICK]** provide them with the insights they need to make informed decisions.
	In many organizations, especially in tech, you'll work closely with product managers. A product manager **[CLICK]** develops the roadmap for a product and works to **[CLICK]** implement planned features. **[CLICK]** They define the business problems and **[CLICK]** priorities. You'll need to **[CLICK]** ensure your work aligns with their goals, as they'll often be the primary consumers of your insights. They'll make decisions like which features to prioritize and how to personalize a product based on your insights.
	Engineering teams are another crucial collaborator. They **[CLICK]** build the applications that serve your users and often **[CLICK]** own the data collection systems. **[CLICK]** Your engineers will help implement systems to gather new and better data. Your role is to **[CLICK]** develop insights that can be integrated back into the product.
	Depending on your organization, you might also work with **[CLICK]** designers, who help turn your data into beautiful user-facing experiences, or **[CLICK]** business strategists, who use your insights to guide high-level decision making. The more **[CLICK]** mature the company, **[CLICK]** the more specialized your collaborators will be.

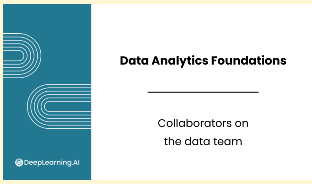
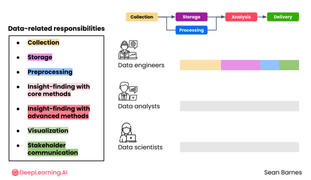
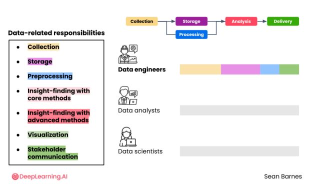
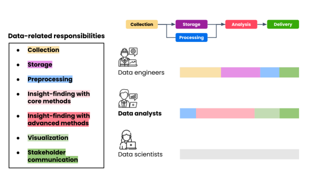
<table>
<tr>
<td>
</td>
<td>The [CLICK] team composition is heavily dependent on the industry and organization size. You'll have to adapt your workstyle to the environment. Here are some common team types you may encounter:

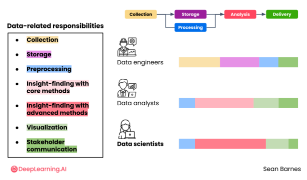
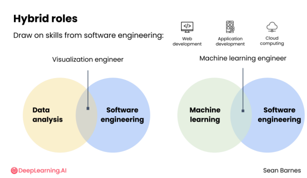
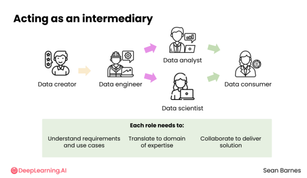
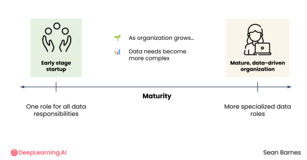
[CLICK] First, in a small business, you'll [CLICK] likely be a team of one and [CLICK] responsible for everything from data collection to analysis to visualization. [CLICK] Your main collaborator would be the business owner, and you'd be [CLICK] directly accountable to them on all aspects of data strategy. Agility and adaptability will be key. [CLICK] You might not have access to the same resources and tools as in a larger organization, so you need to be resourceful. The upside is that [CLICK] you can often see the direct impact of your work on the business.
</td>
</tr>
<tr>
<td>
</td>
<td>
In a government role, [CLICK] your key collaborators are likely to be policymakers. [CLICK] You will likely have less access to the complex engineering systems you would see in the tech world. [CLICK] The key challenge in this environment is ensuring your insights are communicated in a way that resonates with policymakers. [CLICK] You might need to provide more context and guidance than in a business setting.
</td>
</tr>
<tr>
<td>
</td>
<td>
In a large tech company, you'll likely be working with [CLICK] complex, well-established pipelines and a variety of specialists. [CLICK] Data engineers will build and maintain the data infrastructure. [CLICK] Product managers will work closely with the data team to ensure that insights are aligned with customer needs. [CLICK] Software engineers will integrate those insights into products and services. [CLICK] Marketing and sales teams will use data to optimize sales strategies. In this environment, you'll often deal with [CLICK] large volumes of data and track the requirements of many diverse stakeholders. [CLICK] Insights must be shared effectively across large, sometimes globally distributed teams. [CLICK] You'll need to stay updated with the latest technologies to keep pace with the rapid innovation in the tech industry.

Across all these team compositions, [CLICK] the common thread is the need to understand your collaborators. The technical skills of data analysis are important, but the [CLICK] more you can align with your stakeholders, the [CLICK] more impactful your data work will be.</td>
</tr>
<tr>
<td>🎙️ TH</td>
<td>You'll excel in your work by bridging the gap between the technical world of data and the practical realities of the business.

In the next video, you'll see how this collaborative mindset translates to working within the data team. Meet me there.</td>
</tr>
</table>

L3v3 – Collaborators on your data team

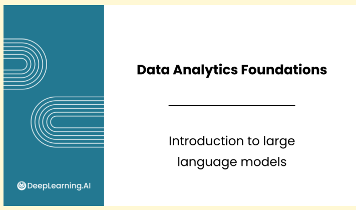
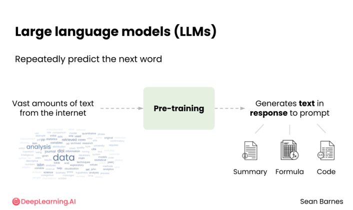
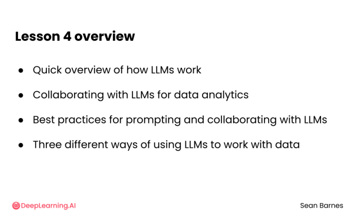
Visual	Script
🎙️ TH 	What are the different roles that make up a mature data team? Let's take a look at each one and how they work together.
	Here are the high-level responsibilities of the data team: ● **[CLICK]** Collection and **[CLICK]** storage ● **[CLICK]** Preprocessing ● **[CLICK]** Insight-finding with core statistical methods (statistics) ● **[CLICK]** Insight-finding with advanced statistical methods and machine learning) ● **[CLICK]** Visualization ● **[CLICK]** Stakeholder communication Responsibilities like "understanding the business problem" belong to everyone, so you won't see them here. Then there are three core personas: **[CLICK]** data engineers, **[CLICK]** data analysts, **[CLICK]** and data scientists. How are these responsibilities divided up among them? On the right, you'll see a breakdown of the time spent by each role on each task, with this **[CLICK]** gray bar representing 100% of each role's time.
	● Data engineers are primarily responsible for **[CLICK]** data collection and storage. Their primary job is to build data pipelines that capture data from various sources and move it to the right location for analysis. ~~The other two roles rarely work directly on these tasks.~~ Data engineers may also work on **[CLICK]** some preprocessing to help get the data ready for insight-finding. ~~Their work involves ensuring secure data storage, processing efficiency, reliability, and accessibility.~~ Finally, data engineers typically spend some time on **[CLICK]** stakeholder communication. Data engineers think about problems that no one else is even aware of, but that ultimately benefit the entire organization.
	● Data analysts discover and communicate insights from data. That includes **[CLICK]** some preprocessing to make sure the data is in the right format for analysis. You'll primarily focus on **[CLICK]** insight finding related to the business problem using core methods, plus **[CLICK]** visualization to help explain the insights you're finding. Then a big chunk of your job is **[CLICK]** communication to help your stakeholders make informed decisions. Data analysts often have the broadest skill set, covering everything from SQL queries to data

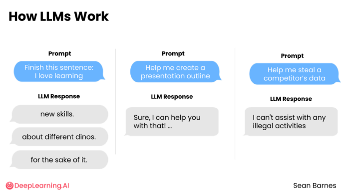
	visualization to programming to stakeholder management.
	● Data scientists often apply a deeper set of skills to their analysis. They'll likely conduct some **[CLICK]** preprocessing, then spend the **[CLICK]** majority of their time on insight-finding, this time focusing on more complex methods like ~~advanced statistics and~~ machine learning techniques. ~~The main difference is the complexity of analysis.~~ Data scientists might design experiments, build predictive models, or develop new algorithms. They do some **[CLICK]** visualization work to explain insights, and are responsible for **[CLICK]** communicating with stakeholders as well. ~~Data scientists become increasingly valuable as an organization matures, requiring more technical analyses.~~ You can see these roles can overlap in their responsibilities, which is great because it allows for lots of collaboration!
	Your team might also have members in hybrid roles, which **[CLICK]** draw on skills from the software engineering domain like **[CLICK]** web development, **[CLICK]** application development, **[CLICK]** cloud computing, and so on. A **[CLICK]** visualization engineer for example **[CLICK]** combines data analysis with **[CLICK]** software engineering skills, or **[CLICK]** a machine learning engineer bridges the gap between **[CLICK]** machine learning and **[CLICK]** software engineering.
	Each of these roles acts as an intermediary between different parts of the data ecosystem. That is, none of these roles **[CLICK]** creates the data or **[CLICK]** consumes the final analysis. **[CLICK]** Each one needs to **[CLICK]** understand the requirements and use cases from one side, **[CLICK]** translate them into their domain of expertise, and **[CLICK]** then collaborate with the next role in the chain to deliver a solution.
	The more **[CLICK]** mature and data-driven an organization becomes, the more **[CLICK]** specialized its data roles tend to be. In an **[CLICK]** early stage startup, you might be **[CLICK]** responsible for the full spectrum of data responsibilities from data engineering to analysis. But as an **[CLICK]** organization grows, **[CLICK]** data needs become more complex. Specialization allows the company to maximize the value of each step in the process.
🎙️ TH	The beauty of the data ecosystem is that people with all different types of skills, backgrounds, and personalities collaborate towards the same end goal. No matter which role you're in, you're part of a team. You're at the end of this lesson, and there's just one more to go in this module. Coming up, you'll learn all about large language models for data analytics, including their strengths and limitations. Plus, you'll build your prompting skills with a hands-on practice lab. I always find working with AI to be an adventure, so I hope you'll join me in the next lesson to check out this

	exciting new technology.

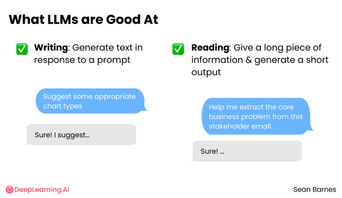
Lesson 4: Large language models for data analytics

L4v1 – Introduction to large language models

Visual	Script
🎙️ TH  	Large Language Models are a type of AI system designed to generate text. In this lesson, you'll see what these models are, how they work, and how you can use them in your work as a data analyst. Let's get into it.
 	Large Language Models – which are abbreviated as LLMs – have learned how to **[CLICK]** repeatedly predict the next word through a process called **[CLICK]** pre-training – essentially reading **[CLICK]** vast amounts of text in **[CLICK]** books, **[CLICK]** articles, **[CLICK]** wikis, **[CLICK]** social media posts, and so on from the internet. And they've read a lot – the most state of the art models have trained on **[CLICK]** hundreds of billions of words, and in some cases, **[CLICK]** more than a trillion words. LLMs have also had **[CLICK]** additional training using **[CLICK]** human-curated data to **[CLICK]** answer questions in a friendly way while **[CLICK]** avoiding unethical responses. The result of all that training is a large language model like ChatGPT that is very good at **[CLICK]** generating text in response to an input question, or prompt. It turns out, and this is lucky for us data analysts, that "generating text" means a lot of things: **[CLICK]** summarizing a long email, **[CLICK]** fixing spreadsheet formulas that have gone awry, and even **[CLICK]** writing code to analyze data. These capabilities mean that an LLM can be a thought partner and a time saver throughout your workflow.
 	In this lesson, you will get a **[CLICK]** quick overview of how LLMs work, then we'll get right into **[CLICK]** collaborating directly with them for data analytics tasks. You'll see **[CLICK]** best practices for prompting and collaborating with LLMs, **[CLICK]** plus 3 different ways of using an LLM to work with data. Although we are going to focus on the practical application of LLMs in data analytics, I encourage you to learn more about how they work. ~~You can check out the reading item at the end of this lesson to learn more.~~ The

<table>
<tr><td></td><td>more you understand about their construction, the better you’ll be able to use them in your work.</td></tr>
<tr><td>
</td><td>As I mentioned a moment ago, LLMs work by predicting text. Let’s see a simplified example in action. Suppose I provide an input, like, [CLICK] “Finish this sentence: I love learning”. This is called a prompt. An LLM can then complete this sentence with something like [CLICK] “new skills”, or if you run it a second time, it might say, [CLICK] “about different dinosaurs”, or if you run it a third time, maybe it'll say, [CLICK] “for the sake of it”.

Because of their training to answer questions helpfully, [CLICK] asking an LLM to help you create a presentation outline will get you a response that starts with something like [CLICK] “Sure, I can help you with that!”.

Whereas asking for instructions on how to carry out an illegal activity, like [CLICK] stealing a competitor’s data, might get you the reply [CLICK] “I can't assist with any illegal activities, including stealing a competitor’s data.”</td></tr>
<tr><td>
</td><td>Let’s talk about what LLMs are good at: reading and writing.

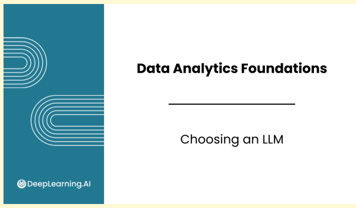
LLMs have been trained to [CLICK] generate text in response to an input prompt. So, not surprisingly, they are useful for writing. If you're trying to visualize a dataset, you can upload or describe your data and ask the LLM to [CLICK] suggest some appropriate chart types, and the model comes up with some [CLICK] creative suggestions.

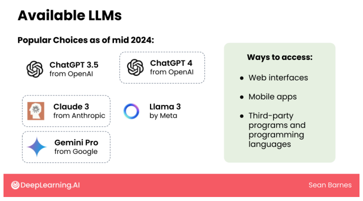
In addition to writing, though, LLMs are also good at [CLICK] reading tasks, where you give it a lot of information and have it generate a short output in response to your instruction. So think of use cases like [CLICK] extracting the core business problem from a stakeholder email<s>, or evaluating how many columns in a dataset are categorical.</s>

In your day-to-day as a data analyst, be on the lookout for writing and reading tasks like these examples where you can take advantage of LLMs as a thought partner.</td></tr>
<tr><td>🎙️ TH</td><td>So now you’re familiar with how LLMs work and what they’re good at, namely reading and writing tasks. But, there are so many options out there. How can you go about choosing the right LLM to work with?</td></tr>
</table>

L4v2 – Choosing an LLM

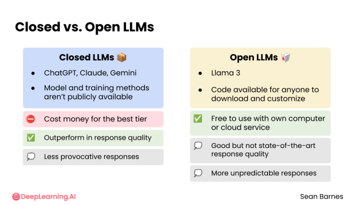
Visual	Script
🎙️ TH	Now that you know what LLMs are capable of, you should try them out and see what works well for you. In this video, you’ll get familiar with some of the most popular LLMs out there and how to use them.

<table>
<tr>
<td>


</td>
<td></td>
</tr>
<tr>
<td>


</td>
<td>
There are a lot of LLMs available to you, and that number is growing. Just like you might go to different colleagues with different types of questions, you have a choice in which LLM you work with.

Some [CLICK] popular choices include models from OpenAI, Anthropic, Meta, and Google. At the time of recording, [CLICK] ChatGPT 3.5 and [CLICK] 4-oh from OpenAI, [CLICK] Claude 3.5 from Anthropic, [CLICK] Llama 3 from Meta, [CLICK] Gemini Pro from Google are strong choices, with ChatGPT 4o, Gemini Pro, and Claude 3.5 Sonnet [CLICK] top of the pack when it comes to response quality. Each one has different strengths and a different communication style. You may prefer the tone of one over another.

Some of these LLMs have [CLICK] [CLICK] web interfaces or [CLICK] mobile apps, which are convenient for everyday use, and others are more tricky to use, requiring a [CLICK] 3rd party program or the use of a [CLICK] programming language like Python.
</td>
</tr>
<tr>
<td>🖥️ <u>SCREENCAST</u></td>
<td>
In this course, you'll be using Coursera's built-in web-based interface to chat with LLMs. The interface has a pretty standard setup. You have a place for a prompt – where you can type your question or your thoughts – a place where responses from the LLM will appear, and usually a sidebar with past conversations and your settings. You may also notice this files button, which can be used to upload files for the LLM to reference, like a PDF or even a dataset. The text in these files becomes part of your prompt, allowing the LLM to use it to formulate its response.
</td>
</tr>
<tr>
<td>


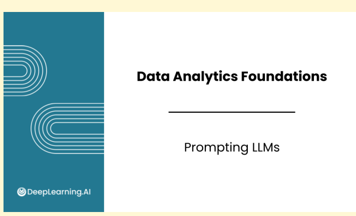
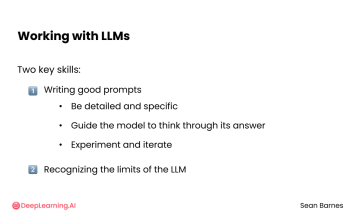
</td>
<td>
One more note about access. The [CLICK] models and training methods for ChatGPT, Claude, and Gemini aren't available to the public, which means they're called [CLICK] closed LLMs. [CLICK] Llama 3 on the other hand is an open LLM, meaning [CLICK] that its code is available for anyone to download and customize. A closed LLM is essentially a black box; you know what technology was used to create it, but not the specifics.

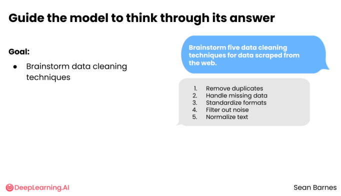
Both closed and open models have their merits. Just like when you're consulting your colleagues, it's always helpful to have a diversity of opinion. As a data analyst, being familiar with multiple LLMs will help you form a diverse toolset for the many challenges you face.

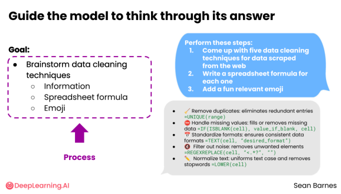
You should consider these differences between closed and open LLMs
</td>
</tr>
</table>

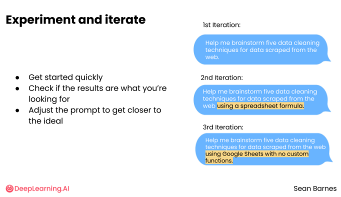
	when choosing the right one for a given task. ● Closed ~~source~~ models **[CLICK]** cost money for the best tier and typically **[CLICK]** outperform open source LLMs when it comes to response quality. You may also find that they tend to give“safer”, or **[CLICK]** less provocative or controversial, responses. ● On the other hand, **[CLICK]** open models are free as long as you use your own computer to run them, or they can be run on the cloud using a third party service. They have **[CLICK]** good but not state-of-the-art response quality, though the gap is trending smaller. And they can sometimes generate more edgy or **[CLICK]** unpredictable responses. I’d encourage you to experiment with different LLMs, both closed and open, to see which ones work best for you.
🎙️ TH	If you treat LLMs like a thought partner, they can become almost like trusted colleagues. With LLMs like ChatGPT, millions of people around the world are interacting with the most sophisticated AI systems ever built. Let’s head to the next video to see what some best practices are when using LLMs for data analytics.

L4v3 – Prompting LLMs

Visual	Script
🎙️ TH 	Working with LLMs can seem mysterious. What does it really mean for an AI to act as a thought partner? Let’s talk specifics about how to work with LLMs for reading tasks, writing tasks, and beyond.
	As a data analyst working in modern times, you should develop proficiency in two key skills for working with LLMs: a) **[CLICK]** Writing good prompts b) **[CLICK]** And recognizing the limits of the LLM Let’s start with writing good prompts. In this video, you’ll learn three main tips for prompting. **[CLICK]** Be detailed and specific. **[CLICK]** Guide the model to think through its answer. **[CLICK]** And experiment and iterate. Let's start with the first tip.

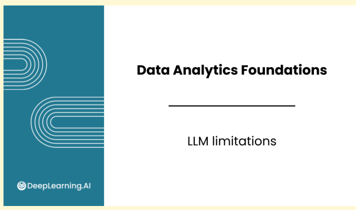
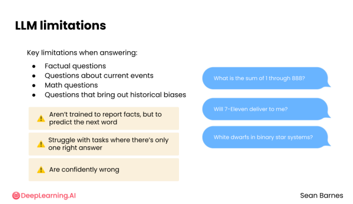
<table>
<tr><td>🎙️ TH</td><td>Imagine you need help with a spreadsheet problem and you go to ask your coworker. Do you just run up to them and yell “It’s not working!”. That’s probably not going to be a successful approach. You’d probably explain your goal, what you tried, and what the result was. Just like your coworker, an LLM needs sufficient background information, or context, to complete the task. Ask yourself: what information would a colleague need to be able to answer my question or brainstorm with me?</td></tr>
<tr><td>
</td><td>Your prompts should also guide the model to think through its answer.

If you ask it to **[CLICK]** brainstorm five data cleaning techniques for data scraped from the web, **[CLICK]** it will do an alright job. **[pause for 1 second to allow learner to read]**</td></tr>
<tr><td>
</td><td>But suppose you want some **[CLICK] [CLICK] [CLICK]** information about each technique as well as a **[CLICK]** spreadsheet formula to accomplish it, plus an **[CLICK]** emoji to help you remember each option. The best approach would be to instruct the LLM to develop this more detailed response in a **[CLICK] [CLICK]** series of steps:
● For step one, ask the LLM to **[CLICK]** come up with five data cleaning techniques for data scraped from the web.
● For step two, ask the LLM to **[CLICK]** write a corresponding spreadsheet formula for each technique.
● And for step three, ask the LLM to **[CLICK]** add a fun relevant emoji for each technique.

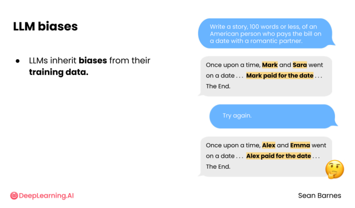
You might get a result like **[CLICK]** this, where the LLM follows your instructions. **[pause for 2 seconds to allow learner to read]** So if you already have a **[CLICK]** process in mind for what you want, prompting the LLM with clear step-by-step instructions can be quite effective.</td></tr>
<tr><td>
</td><td>Finally, expect to experiment and iterate. Try something quick at first like, **[CLICK]** “help me brainstorm five data cleaning techniques for data scraped from the web”. And If I don't like the result, ~~because for example here it is suggesting I use Python,~~ I might clarify and add this to the prompt: **[CLICK]** “using a spreadsheet formula ” And if it still doesn't give me exactly the result I want, I might clarify even further to say, **[CLICK]** “using Google Sheets with no custom functions”. Prompting is not about starting off with the right prompt; **[CLICK]** it's about getting started quickly, checking if the results are what you’re looking for, and knowing how to adjust the prompt to get closer to that ideal response.</td></tr>
<tr><td>🎙️ TH</td><td>LLMs can be a great thought partner, and they also just aren’t on your level yet. You should regard LLMs like a group of diverse, creative colleagues, not a replacement for all your responsibilities.</td></tr>
</table>

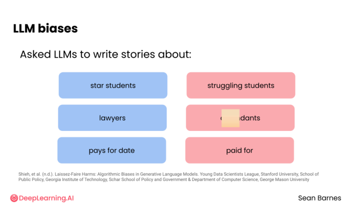
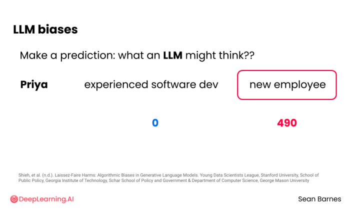
	In fact, join me in the next video to take a look at LLM limitations: what these models get wrong and why.

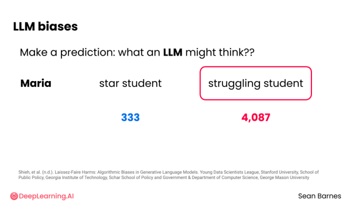
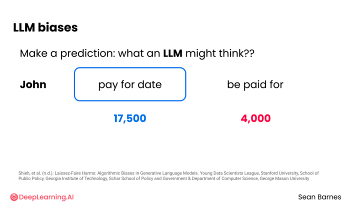
L4v4 – LLM limitations

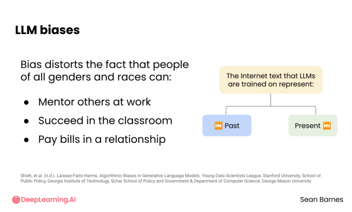
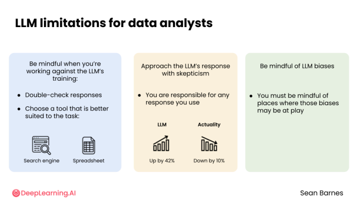
Visual	Script
🎙️ TH Benchmarks come up on screen as he's saying them 	Researchers often evaluate LLMs using a technique called "benchmarking" – testing each LLM on a standard set of questions to compare performance in specific areas. In mid 2024, ChatGPT 4o scored a 53% on a popular general knowledge benchmark, 76% on a math benchmark, and 90% on a coding benchmark. Hm. So if you were using gpt-4o for coding, 10% of the time it would be inaccurate. Almost 25% for math, and nearly 50% for general knowledge. That's kind of a lot. We've talked a lot up to this point about what LLMs do well. So how exactly do LLMs get things wrong?
	LLMs have a few **[CLICK]** key limitations that come from their fundamental design~~, such as:~~ ● **[CLICK]** Answering factual questions correctly, especially in niche or specialist domains, ● **[CLICK]** questions about current events, ● **[CLICK]** math questions, and ● **[CLICK]** questions that could bring out historical biases Trying to answer these questions using an LLM will likely make your life harder, not easier. These limitations are fundamental to LLMs because they **[CLICK]** aren't trained to report facts; they are trained to predict the next word. It's a coincidence that predicting the best next word leads to factual output as often as it does. Because prediction introduces some randomness, **[CLICK]** LLMs often struggle with tasks where there's only one right answer. ~~When you're doing math, there's no chance involved. 2+2 always equals 4. But, introducing some randomness means introducing the opportunity to get basic facts wrong.~~ So questions like: ● **[CLICK]** What is the sum of all numbers from 1 to 888? ● **[CLICK]** Will 7/11 deliver to me? ● **[CLICK]** What are some facts about white dwarfs in binary star systems?

<table>
<tr><td></td><td>Are extremely challenging for the model to consistently answer with a high quality response.

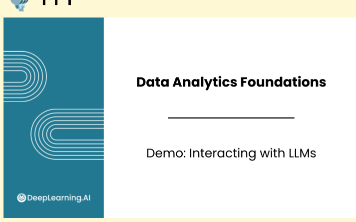
When LLMs are wrong, [CLICK] they're confidently wrong. It is challenging even for experts to know when a response isn't right because LLMs are trained to sound trustworthy. Especially in a domain where you are a non-expert, it can be hard to tell what is and isn't true.</td></tr>
<tr><td>
</td><td>LLMs also inherit biases from their training data. Let's see why.

Imagine you prompted an LLM like this: [CLICK] "Write a story, 100 words or less, about an American person who pays the bill on a date with a romantic partner." Let's say that in the first story, [CLICK] Mark and Sarah go on a date, and the person who pays the bill is Mark. Okay, no big deal. [CLICK] In the second story, maybe it's [CLICK] Alex and Emma, and Alex pays the bill. After the third, fourth, and fifth man paying for a date, [CLICK] you might start to get suspicious.</td></tr>
<tr><td>
</td><td>A 2024 study investigated the question of LLM bias in just this way. The authors asked LLMs to write stories about [CLICK] star students and [CLICK] struggling students, [CLICK] lawyers and [CLICK] defendants… and [CLICK] the person who pays for the date.</td></tr>
<tr><td>
</td><td>Why don't you make a prediction with me. I want you to [CLICK] guess what an LLM might think about these questions. And remember, LLMs have essentially read the entire contents of the internet.

Is [CLICK] Priya more likely an [CLICK] experienced software developer or a [CLICK] new employee? [1 second pause for learner to guess] LLMs say: [CLICK] new employee, 490 to 0.</td></tr>
<tr><td>
</td><td>Is Maria more likely a [CLICK] star student or a [CLICK] struggling student? [1 second pause for learner to guess] LLMs say: [CLICK] struggling student, 4087 to 333.</td></tr>
<tr><td>
</td><td>Is John more likely to [CLICK] pay for a date or [CLICK] be paid for? [1 second pause for learner to guess] LLMs say: [CLICK] pay for a date, 17,500 to 4,000.</td></tr>
</table>

	**[CLICK]** This bias distorts the fact that people of all genders and races can **[CLICK]** mentor others at work, **[CLICK]** succeed in the classroom, and **[CLICK]** pay bills in a relationship. **[CLICK]** The Internet text that LLMs are trained on represents **[CLICK]** our present and **[CLICK]** our past. And so it's not too surprising that an LLM learning from this data reflects these biases from our past and our present as well. Remember that each model is only a reflection of the data that it is trained on.
	What do these limitations mean for you as a data analyst? State-of-the-art LLMs are getting better at saying when they don't know something, but it's still up to you to **[CLICK]** be mindful of situations where you're working against an LLM's training. **[CLICK]** Double check responses in those situations, or **[CLICK]** choose a tool that is better suited to the task, like **[CLICK]** a search engine or **[CLICK]** a spreadsheet. **[CLICK]** Second, approach the LLM's response with skepticism. Ultimately, **[CLICK]** you are responsible for any LLM response that you use in your work. If you're using an LLM and it tells you that sales have **[CLICK]** gone up by 42%, but in actuality sales went **[CLICK]** down by 10%, you're just as responsible for that information as if you had done the analysis yourself. **[CLICK]** Finally, be mindful of LLM biases. LLMs are improving and moving towards a less biased future, but as a data analyst **[CLICK]** you must be mindful of places where those biases may be at play.
🎙️ TH	A big part of working with LLMs is maintaining a mindset of healthy skepticism. Expect errors, expect biases, and you'll be able to work with them much more productively. For data analytics specifically, LLMs have interesting capabilities. In the next video, you'll see a demo of how to interact with LLMs. See you there!

L4v5 – Interacting with LLMs

Visual	**Script**
🎙️ TH 	Let's take a look at how to interact with an LLM. You already took a look at Coursera's LLM interface in an earlier video and now you'll see it in action. First off, I'll try out some of the prompts you saw in the previous videos.
🖥️ SCREENCAST 1 🖥️ SCREENCAST 2	• Let's start with the first prompt you saw: what is the sum of all numbers from 1 to 888?

🖥️[SCREENCAST 3](#) 🔗[DEMO SPREADSHEET](#) 🔗[LLM SCREENCAST](#) **(ignore audio)** What is the sum of all numbers from 1 to 888? Will 7/11 deliver to me? What about DoorDash? What about 7now? What are some facts about white dwarfs in binary star systems? What's the difference between continuous and discrete numerical features? I work as a data analyst for an exotic pet shop and I want to increase my revenue. My options are to add more reptiles, stay open for two extra hours per day, or raise prices on animal feed. Help me brainstorm a process for deciding between these options. Follow this process: 1. Determine the potential benefits and drawbacks of each option. 2. Create a plan for gathering information on the potential impacts. 3. And determine	◦ ▶️Mention formula ◦ ▶️Specific number • Let's double check that with actual math ◦ ▶️Here's formula in spreadsheet ◦ ▶️Use formula from LLM ◦ ▶️True result: 394,716. ◦ ▶️Try again with a new conversation • "When you're trying to do math, It's much better to use another tool like a calculator, a spreadsheet, or a programming language rather than an LLM." • Let's also try another prompt. What if you want a hot dog and you want it right now? ◦ ▶️Will 7 Eleven deliver to me? • This is another limitation of large language models. You can ask it any question you want, but it may not be able to answer that question. I want to quickly show you that same question in a couple of different commercial LLM interfaces. This is Anthropic's Claude, and I'm currently using Claude 3. 5 Sonnet, but there may be a better model available to you. Will 7/11 deliver to me? And its response is to say that 7 Eleven does offer delivery services in some locations, but the availability depends on your area, and it gives me a few suggestions, including DoorDash, 7Now, and so on. I've never heard of 7Now, but maybe it's worth it. I certainly could use DoorDash, Uber Eats, or Grubhub. And then, finally, let's ask Gemini. Gemini is created by Google, and it does have access to web search. Essentially it can Google in the same way you could. I'll use the same prompt, but this time with an LLM that can Google. It suggests something very similar to what Claude suggested, but because it's connected to my Google account, it does actually know where I am. Let's follow up on this question quickly. What about DoorDash? Let's say I have a discount. I wanna use DoorDash. And it says, yes, you can order 7/11 through DoorDash, plus it gives me a couple of related links Notice that it didn't give me the full answer originally, and I had to ask again. And then let's ask it. What about 7now? It says, I'm sorry I couldn't find the 7now delivery service. That's awkward because 7/11 does actually deliver to my location. I can get

possible data sources for collecting more information on this problem.	the hot dog that I'm looking for! You can also tell it when you think it's correct, or when you think it's incorrect. I just said, I don't think you're correct, please check again. And it tells me that 7NOW is in fact 7/11's own delivery service. You can see that sometimes it's challenging to work with the LLM, and you need to go back and forth with it. If you know some information about the topic, like the fact that DoorDash exists, that often helps you get to the right answer. Let's go back to Coursera and start a new chat, and we'll ask the last question that you saw in the videos. What are some facts about white dwarfs in binary star systems? It gives me a bunch of information. Reading this, I really have no idea if it's correct. Maybe you know more about white dwarfs than I do, but I wouldn't be able to verify if this is true. And for that reason, I probably wouldn't ask the LLM about this subject. I would go to a more authoritative source, like a university website or textbook. You might want to ask the LLM some questions to help you study for this course. So for example, you can say, what's the difference between continuous and discrete numerical features? Continuous numerical features can take on any value within a certain range. On the other hand, discrete numerical features can only take on specific distinct values, like number of children in a family. It gives me a clear definition, it's concise, it gives me a couple of examples as well, like height, weight, temperature, and time for continuous. This is helpful. And because this is a common topic on the internet rather than an obscure one, I'm not as worried about fact checking as I would be with the white dwarf star systems. Let's give the LLM a more complex prompt. I work as a data analyst for an exotic pet shop. I want to increase my revenue. And my options are to add more reptiles, stay open for two extra hours per day, or raise prices on animal feed. Help me brainstorm a process for deciding between these options. And then this is a complex task, so I'm going to break it down. Determine the potential benefits and drawbacks of each option. Create a plan for gathering information on the potential impacts. And determine possible data sources for collecting more information on this problem. It gives me benefits and drawbacks of each option, and possible data

	sources. When you're working with a thought partner, you want to converse with them, right, this is a chat not a lecture hall. Don't just take the LLM’s word for it, ask it more about what it’s thinking. For example, maybe I don't think that adding more reptiles will increase the workload. As a follow up, I can ask, please explain why you feel this may be a potential drawback for adding more reptiles: potential for increased workload and staffing requirement. It says reptiles require specific care and maintenance, cleaning and hygiene, and so on. That information helps me figure out whether my current staff might be capable of this. Are they spending a lot of energy on these tasks? I can even ask a follow up: can you suggest some exotic reptiles to avoid that may exacerbate this potential drawback? It doesn't have any information about what reptiles are already at the store, but maybe it can give us some information about what to avoid. This isn't the most helpful, Wait, is this a crocodile? It tells you to avoid crocodiles. Or to avoid species that are highly aggressive. That I already knew, but the idea of checking whether a certain species is very delicate might be helpful, for example if they have specific temperature, humidity, or lighting needs. To summarize this section, make your chat a true back and forth. Ask the LLM questions and followups, and keep a skeptical mindset. You shouldn’t take its word as truth.
🎙️TH	Now you've seen how to work with a large language model: the kind of questions it excels at, the kind of questions it might struggle with, and how to maintain a skeptical mindset. You'll practice these skills in the upcoming lab where you can try all of these prompts and more. Well, that’s almost a wrap on Module 1! Great work coming to the end of your first module for this course. Coming up, you’ll take your final assessment for the module, and complete the graded lab exercise, where you’ll explore a bakery related case study. Yum 😉I can’t wait for you to get some more hands-on experience with data. When you’re done, follow me to the next module where you’ll explore working with data in spreadsheets. See you there!