DATA SCIENCE

You can organize data in two different ways

Data is raw information. It might be facts, statistics, opinions—any kind of content that's recorded in some format. Numbers? Sure. Facts? Sure. Voices, photos, names, dance moves? Also data.

So when data scientists handle raw information to uncover its story, they start by organizing it into one of two forms: structured data or unstructured data. The difference between these forms changes how we work with them.

Structured data

Structured data is information that can be laid out in rows and columns. You might already have worked with structured data using a spreadsheet like Microsoft Excel. For complex information, data scientists use more powerful tools like SQL, Apache, or R, which can sort through vast amounts of data stored in many connected tables. Can you organize information within the data into groups based on specific characteristics? Those groups are structured data.

Here's a sample of structured data from your local hardware store.

Customer name	Last name	Phone number	Last order
00001	Ajay	(555) 678-9012	03/12/19
00002	Thompson	(555) 345-6789	08/14/18
00003	Smith	(555) 432-1098	08/01/19
00004	Kim	(555) 665-5443	11/16/18
00005	Gonzalez	(555) 912-9945	12/24/17
00006	Wangzi	(555) 212-3767	06/30/19
00007	Colbert	(555) 866-0922	05/21/19
00008	Jarrah	(555) 778-1845	05/22/18
00009	Magnusson	(555) 395-7677	01/02/19
00010	Cooper	(555) 550-5515	10/30/18

This table organizes customer information from the hardware store by characteristics such as customer number or name. Each row shows information related to a particular customer, while each column shows one customer characteristic that spans a group of customers.
As you can see, structured data tends to be well-organized, making it easier for data scientists to discover its treasure using common data analysis tools. Spreadsheets are based on tables like this, so they handle structured data very well.

Unstructured data

Then there's unstructured data, which is a fancy way of saying "everything else." We use this term when there's no built-in organization (or structure) to the data. Unstructured data can be a collection of audio files, or social media posts, or essay texts, or even song lyrics.
Here are two examples to help you see the difference:

1. Your Department of Motor Vehicles takes photographs of everyone who gets a driver's license. A collection of those images is unstructured data. (But the table of peoples' names, addresses, and licence numbers that indexes those photos is structured data.)
2. A downloadable library might offer text from thousands of different books. The catalog listing names, authors and dates of those books is structured data. (But the text of those books is unstructured data.)

Unstructured data can be harder to work with than structured data, but it's still useful! Suppose a video game company is getting a lot of email bug reports about a new release. The text of those emails is unstructured data. By examining those texts (and perhaps by converting some of their contents to structured data), a data scientist can figure out patterns and identify the problem so the company can fix it!

How do you become a data scientist?

Building a career

What's the path into this field?

In an earlier topic, you saw young professionals talk about data science. Did you notice that no one got into this field the same way? That's because data science is an emerging study that hasn't been defined clearly until recently. (And, in some areas, the definition is still changing.)

It helps to think of data science as creating knowledge from data, no matter what technique is used to analyze the data. It's like a treasure hunt. You start with raw text, or numbers, or graphics, or any other kind of data, and you explore that data to discover valuable patterns and insights.
No special certification or skillset qualifies you as a data scientist. You just start analyzing data in any of its many forms. If you use scientific techniques to derive information from data, you're a data scientist, no matter how you entered the field! But there are areas of study that can help you prepare to hunt for data treasure, and we'll look at them later in this course.

Could you be a data scientist?

Now that you know more about data science and what the work is like, is this a career you'd like to explore?

Find out by answering a few questions in the following text box.

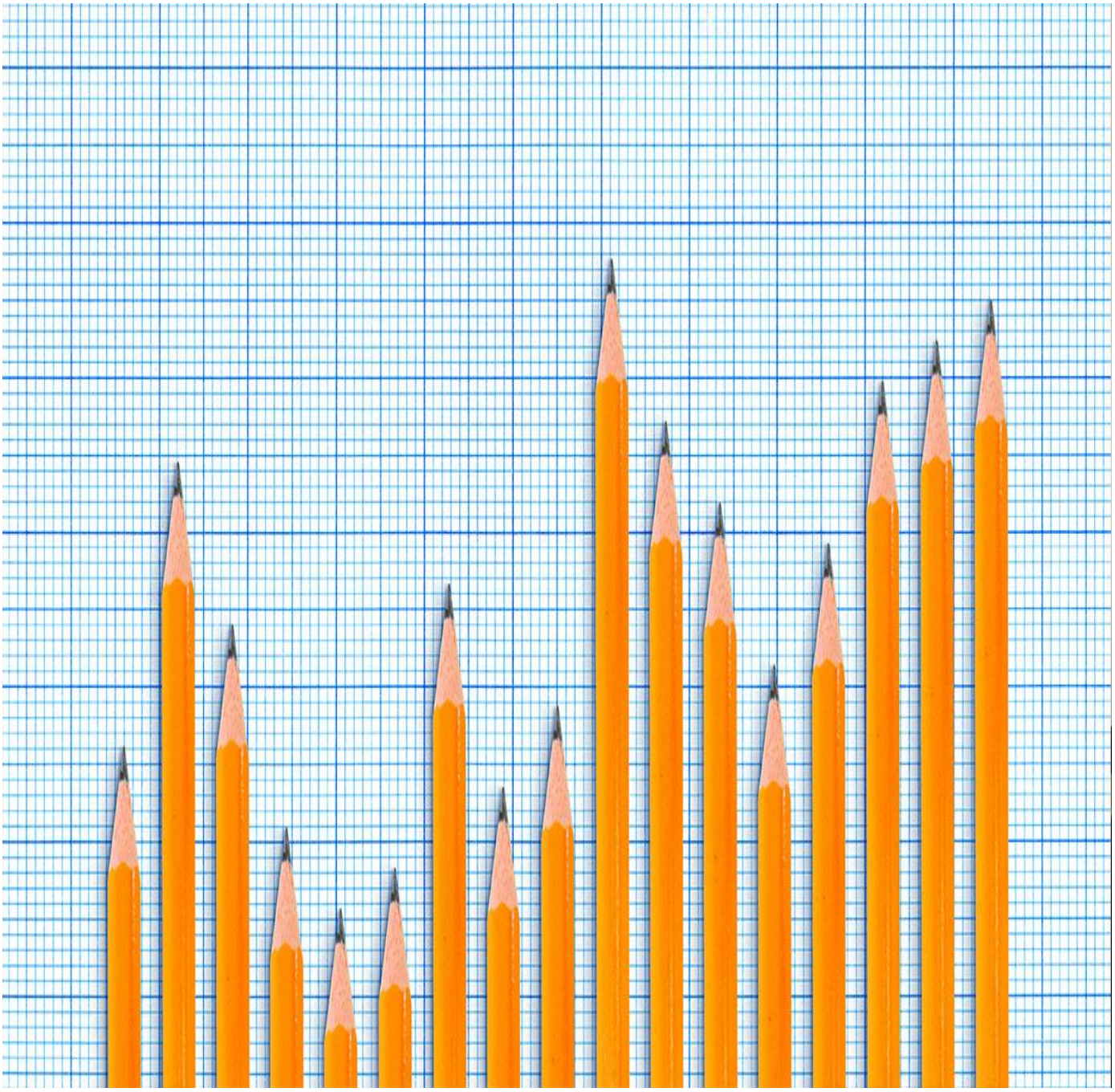
1. What are you passionate about? Write down a field of study, a hobby, or some other topic that you're interested in learning more about. Remember, from soccer to astrophysics to music to hurricane relief, there's data in everything.
2. What data might relate to your interests? Write down examples of what you think is out there.
3. Is that data mostly structured, so you could list it in neat rows and columns? Or is it mostly unstructured, like text or multimedia files?
4. Try asking a question (or two, or three!) that relates to your area of interest and the data that might describe it.

What Data Scientists Really Do, According to 35 Data Scientists

by

- Hugo Bowne-Anderson

August 15, 2018



burakpekakcan/Getty Images

Summary. What do data scientists do? According to interviews with more than 30 data scientists, data science is about infrastructure, testing, using machine learning for decision making, and data products. Data science is being used in numerous fields, but it's not...more

Modern data science emerged in tech, from optimizing Google search rankings and LinkedIn recommendations to influencing the headlines Buzzfeed editors run. But it's poised to transform all sectors, from retail, telecommunications, and agriculture to health, trucking, and the penal system. Yet the terms "data science" and "data scientist" aren't always easily understood, and are used to describe a wide range of data-related work.

What, exactly, is it that data scientists do? As the host of the DataCamp podcast *DataFramed*, I have had the pleasure of speaking with over 30 data scientists across a wide array of industries and academic disciplines. Among other things, I've asked them about what their jobs entail.

It's true that data science is a varied field. The data scientists I've interviewed approach our conversations from many angles. They describe a wide range of work, including the massive online experimental frameworks for product development at booking.com and Etsy, the methods Buzzfeed uses to implement a multi-armed bandit solution for headline optimization, and the impact machine learning has on business decisions at Airbnb. That last example came during my conversation with Airbnb data scientist Robert Chang. When Chang was at Twitter, that company was focused on growth. Now that he's at Airbnb, Chang works on productionized machine-learning models. Data science can be used in a number of different ways, depending not just on the industry but on the business and its goals.

But despite all the variety, a number of themes have emerged from these conversations. Here's what they are:

What data scientists do. We now know how data science works, at least in the tech industry. First, data scientists lay a solid data foundation in order to perform robust analytics. Then they use online experiments, among other methods, to achieve sustainable growth. Finally, they build machine learning pipelines and personalized data products to better understand their business and customers and to make better decisions. In other words, in tech, data science is about infrastructure, testing, machine learning for decision making, and data products.

Great strides are being made in industries other than tech. I spoke with Ben Skrainka, a data scientist at Convoy, about how that company is leveraging data science to revolutionize the North American trucking industry. Sandy Griffith of Flatiron Health told us about the impact data science has begun to have on cancer research. Drew Conway and I discussed his company Alluvium, which "uses machine learning and artificial intelligence to turn massive data streams produced by industrial operations into insights." Mike Tamir, now head of self-driving at Uber, discussed working with Takt to facilitate Fortune 500 companies' leveraging data science, including his work on Starbucks' recommendation systems. This non-exhaustive list illustrates data-science revolutions across a multitude of verticals.

It isn't all just the promise of self-driving cars and artificial general intelligence. Many of my guests are skeptical not only of the fetishization of artificial general intelligence by the mainstream media (including headlines such as VentureBeat's "An AI god will emerge by 2042 and write its own bible. Will you worship it?"), but also of the buzz around machine learning and deep learning. Sure, machine learning and deep learning are powerful techniques with important applications, but, as with all buzz terms, a healthy skepticism is in order. Nearly all of my guests understand that working data scientists make their daily bread and butter through data collection and data cleaning; building dashboards and reports; data visualization; statistical inference; communicating results to key stakeholders; and convincing decision makers of their results.

The skills data scientists need are evolving (and experience with deep learning isn't the most important one). In a conversation with Jonathan Nolis, a data science leader in the Seattle area who helps Fortune 500 companies, we posed the question, "Which skill is more important for a data scientist: the ability to use the most sophisticated deep learning models, or the ability to make good PowerPoint slides?" He made a case for the latter, since communicating results remains a critical part of data work.

Another recurring theme is that these skills, so necessary today, are likely to change on a relatively short timescale. As we're seeing rapid developments in both the open-source ecosystem of tools available to do data science and in the commercial, productized data-science tools, we're also seeing increasing automation of a lot of data-science drudgery, such as data cleaning and data preparation. It has been a common trope that 80% of a data scientist's valuable time is spent simply finding, cleaning, and organizing data, leaving only 20% to actually perform analysis.

But this is unlikely to last. These days even a great deal of machine learning and deep learning is being automated, as we learned when we dedicated an episode to automated machine learning, and heard from Randal Olson, lead data scientist at Life Epigenetics.

One result of this rapid change is that the vast majority of my guests tell us that the key skills for data scientists are not the abilities to build and use deep-learning infrastructures. Instead they are the abilities to learn on the fly and to communicate well in order to answer business questions, explaining complex results to nontechnical stakeholders. Aspiring data scientists, then, should focus less on techniques than on questions. New techniques come and go, but critical thinking and quantitative, domain-specific skills will remain in demand.

Specialization is becoming more important. While there is no well-defined career path for data scientists, and little support for junior data scientists, we are starting to see some forms of specialization. Emily Robinson described the difference between Type A and Type B data scientists: "Type A is the analysis — sort of a traditional statistician — and Type B is building machine learning models."

Jonathan Nolis breaks data science down into three components: (1) business intelligence, which is essentially about "taking data that the company has and getting it in front of the right people" in the form of dashboards, reports, and emails; (2) decision science, which is about "taking data and using it to help a company make a decision"; and (3) machine learning, which is about "how can we take data science models and put them continuously into production." Although many working data scientists are currently generalists and do all three, we are seeing distinct career paths emerging, as in the case of machine learning engineers.

Ethics is among the field's biggest challenges. You may gather that the profession offers its practitioners a great deal of uncertainty. When I asked Hilary Mason in our first episode if any other major challenges face the data science community, she said, "Do you think that imprecise ethics, no standards of practice, and a lack of consistent vocabulary are not enough challenges for us today?"

All three are essential points, and the first two in particular are front of mind for nearly every *DataFramed* guest. At a time when so many of our interactions with the world are

dictated by algorithms developed by data scientists, what role does ethics play? As Omoju Miller, the senior machine learning data scientist at GitHub, said in our interview:

> We need to have that ethical understanding, we need to have that training, and we need to have something akin to a Hippocratic oath. And we need to actually have proper licenses so that if you actually do something unethical, perhaps you have some kind of penalty, or disbarment, or some kind of recourse, something to say this is not what we want to do as an industry, and then figure out ways to remediate people who go off the rails and do things because people just aren't trained and they don't know.

A recurring theme is the serious, harmful, and unethical consequences that data science can have, such as the COMPAS Recidivism Risk Score that has been "used across the country to predict future criminals" and is "biased against blacks," according to ProPublica.

We're approaching a consensus that ethical standards need to come from within data science itself, as well as from legislators, grassroots movements, and other stakeholders. Part of this movement involves a reemphasis on interpretability in models, as opposed to black-box models. That is, we need to build models that can explain why they make the predictions they make. Deep learning models are great at a lot of things, but they are infamously uninterpretable. Many dedicated, intelligent researchers, developers, and data scientists are making headway here with work such as Lime, a project aimed at explaining what machine learning models are doing.

The data science revolution across industries and society at large has just begun. Whether the title of data scientist will remain the "sexiest job of the 21st century," will become more specialized, or will become a set of skills that most working professionals are simply required to have is unclear. As Hilary Mason told me: "Will we even have data science in 10 years? I remember a world where we didn't, and it wouldn't surprise me if the title goes the way of 'webmaster.'"

Getting things done with data science

After learning about what data science is in the What is data science? module, now it's time to explore how it gets done.

In this module, you'll learn about some of the tools and techniques that data scientists use to find the insights that data holds. This includes learning about how to explore data in relational databases, how to communicate data visually, and how data science and machine learning come together to project the future.

Sort data to analyze it

What is a database?

Almost every data scientist will spend time working in a database, which is an organized collection of structured data in a computer system. (Remember, structured data is usually organized in a table format with rows and columns, like the following example.)

Last four digits of social security number	Last name	Age
6881	Marshall	23
0121	Rodriguez	19
5538	Cho	59
2972	Parker	33
3154	Sawyer	72

Most databases today are organized as relational databases, which are collections of multiple data sets or tables that link together. For example, one table might list names and addresses, while the other might list properties and their owners. If some of the owners also appear in the name-and-address table, the two tables can be linked, creating a relational database. Relational Database Management Systems (RDMSs) help data scientists correlate information from all the tables within the database, change related data, or add and delete data in the database without breaking its structure.

Use a database to store and sort data

Most database management systems use some form of Structured Query Language, or SQL, to ask questions or "query" the database or to modify its contents. SQL is incredibly useful, and it's something you can learn through a book, an online course, or a computer club. Let's see how it works.

Suppose you have a million rows of data in a table of registered voters in your state. (Many databases are this big, or even larger!) The table contains information about each voter's polling place, party affiliation, and age. If you wanted to just get the information for all the voters under the age of 21, you could write a query in SQL instead that performs a simple search. That would look like this:

Yes, a simple search might have answered this particular question. However, if you want to uncover voters who live in a particular geographic area so you can study how frequently people of different ages in that group show up at the polls, your query will overwhelm a simple search. But SQL can do the job!

While SQL is the underlying language that drives most work done in relational databases, there are many RDBMSs in which you can do that work. As you venture into this field, you'll run into names like these:

- MySQL
- Microsoft Access
- PostgreSQL
- Oracle
- IBM DB2
- MongoDB

Choose the right tools to manage data

Where do you begin? There are dozens of useful data science tools and platforms! Here's a list of some popular and open source platforms that you can use to begin your own data science journey.

Click the following sections to learn more about tools to manage data.

R is a good place to start

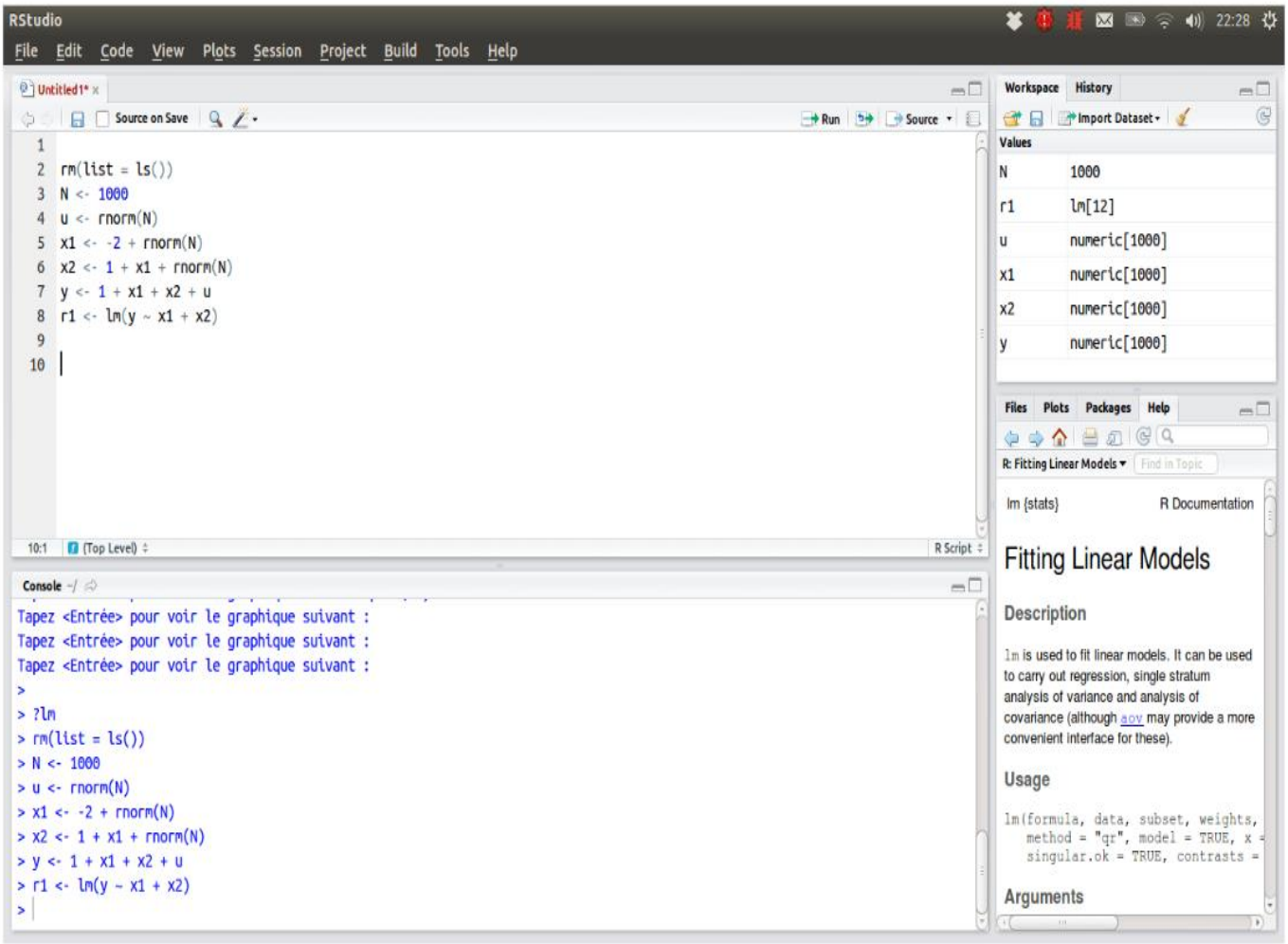
R is a programming language and free software environment often used for statistical analysis and data science. Many would-be data scientists start with this tool or with one of the popular R interfaces, and there are hundreds of useful packages in R that help with data visualization such as ggplot2.

Here's what RStudio looks like. It's a popular interface for working in R.

Python works for general purposes

Python is a popular, general-purpose programming language that can also be used for data science. Pair it with a library like pandas library and with a useful interface, and Python can help you create new insights and data visualizations.

Here's what RStudio looks like. It's a popular interface for working in R.



Here's what Python looks like in a notebook interface.

IP[y]: Notebook Modulation Last Checkpoint: Jan 05 11:01 (autosaved)

File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

An angle modulated signal generally can be written as

$u(t) = A_c cos(2\pi f_c t + \phi(t))$

In a phase modulated (PM) system, the phase is proportional to the message

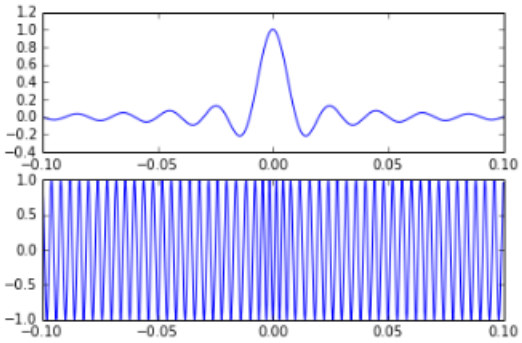
$\phi(t) = k_p m(t)$

In a frequency modulated (FM) system, instatantaneous frequency deviation is proportional to the message

$$f_i(t) - f_c = k_f m(t) = \frac{1}{2\pi}\frac{d}{dt}\phi(t)$$

In [12]:
```
from numpy.fft import fft,fftfreq
t = arange(-0.1,0.1,0.0001)
m = sinc(100*t)
int_m = empty(len(t))
for k in range(len(t)):
    int_m[k] = trapz(m[0:k],t[0:k])
u = cos(2*pi*250*t + 2*pi*100*int_m)
subplot(211)
plot(t,m)
subplot(212)
plot(t,u)
```

Out[12]: [<matplotlib.lines.Line2D at 0xd3a490c>]



MATLAB helps crunch numbers

MATLAB was built to focus on numerical computing. It is often used in higher education.

Apache Spark supports big data and machine learning

Apache Spark is a proprietary general-purpose framework that can be especially useful for extremely large data sets and the machine learning that uses them.

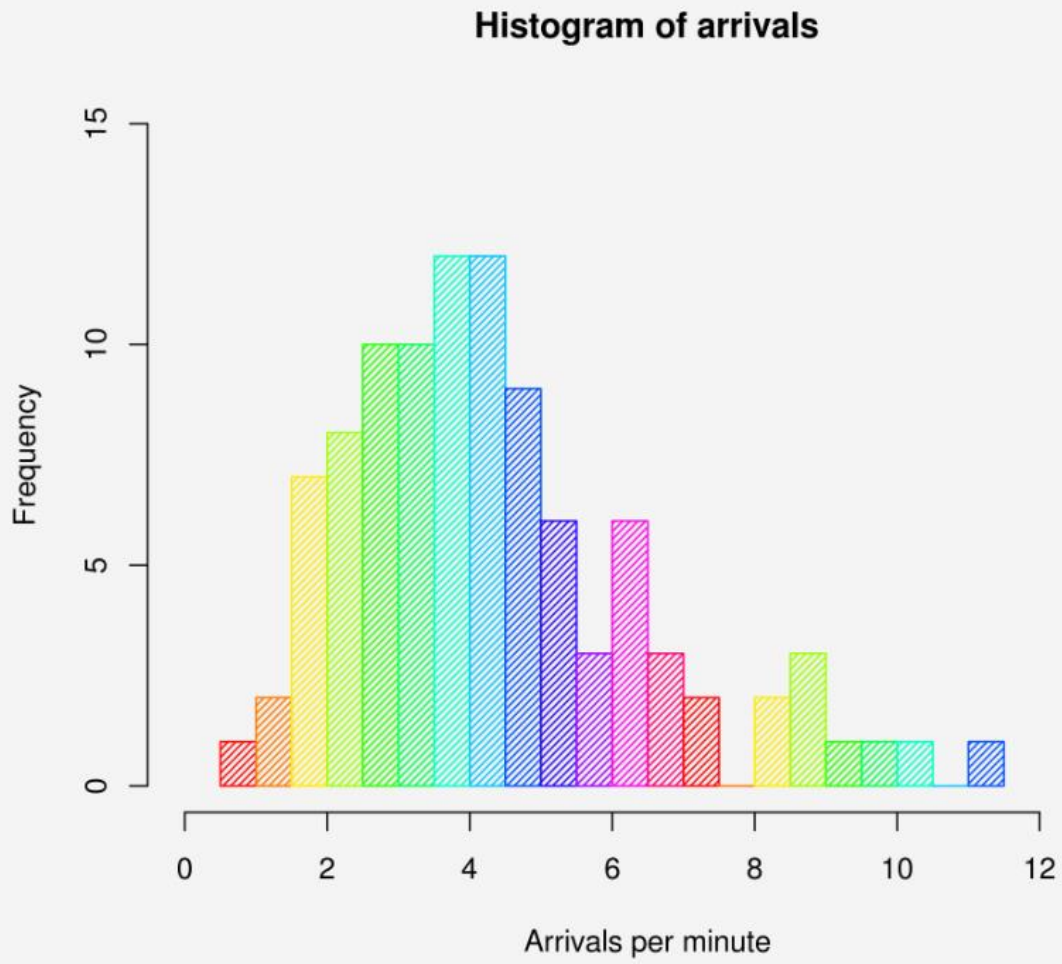
Visualize your data

What good is data if you can't communicate it in a way that people understand? Data scientists use visualizations like graphs or maps to help people grasp data's meaning. Done right, these tools can bring clarity and simplicity to complicated issues.

Use charts for statistics

Graphs and charts are a great way to shine a light on messy statistics or information in tables.

There are many types of graphs and charts, such as the histogram displayed above, and including scatter plots, bar charts, box plots, and many more.



There are many types of graphs and charts, such as the histogram displayed above, and including scatter plots, bar charts, box plots, and many more.

Use maps for relationships

Maps are a great way to express data that can be laid out as areas in two dimensions. They work well not only for geography, but also for relationships like brand popularity, music preferences, or even municipal power grids.

Choose the right tool for data visualization

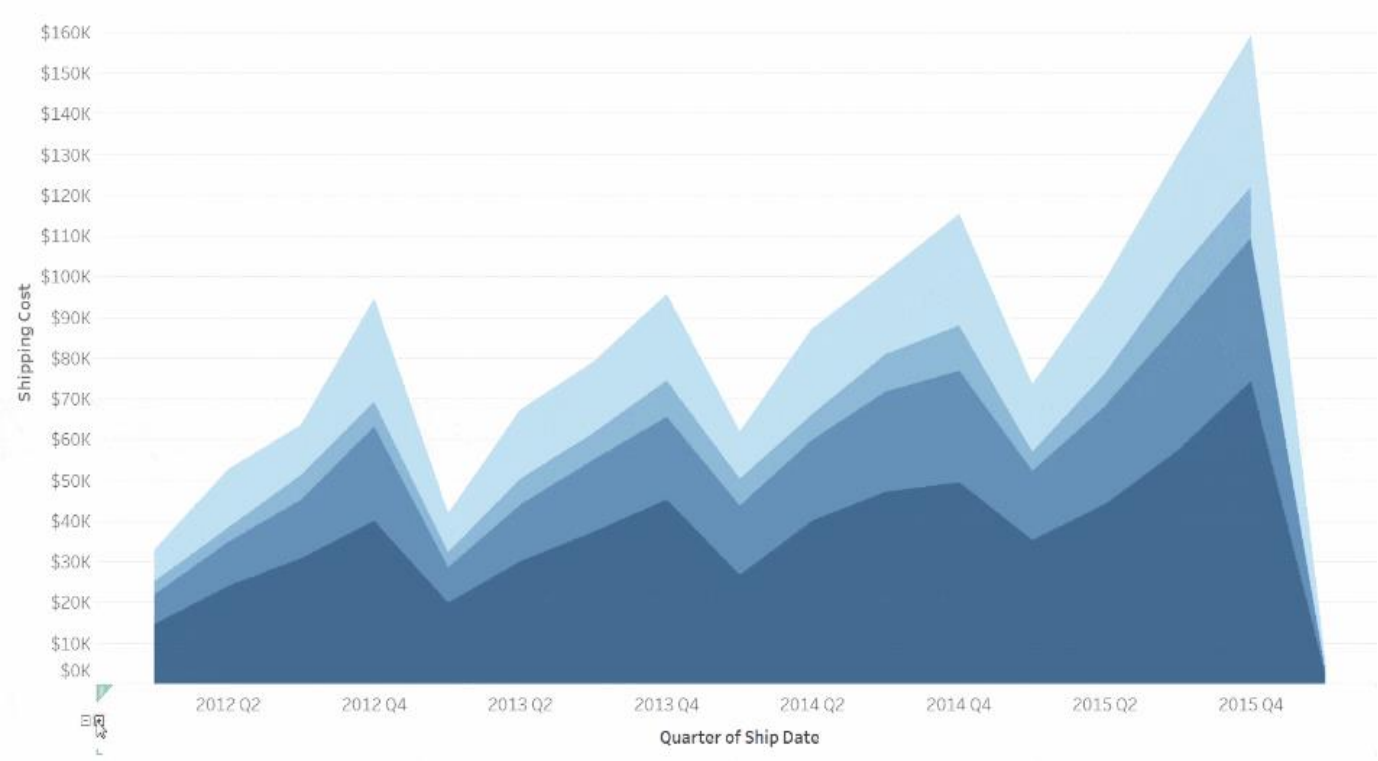
You can create simple data visualizations like charts or graphs in most spreadsheet programs. But as the data gets more complex, you'll turn to other tools. Some are sold as stand-alone visualization products, while others are add-ons to data management systems.

You'll also find great books, in print and online, that help you make your visualizations more useful. Here's a hint: Check out the work of Edward Tufte, a visualization master!

Click the following sections to learn more about tools to visualize data.

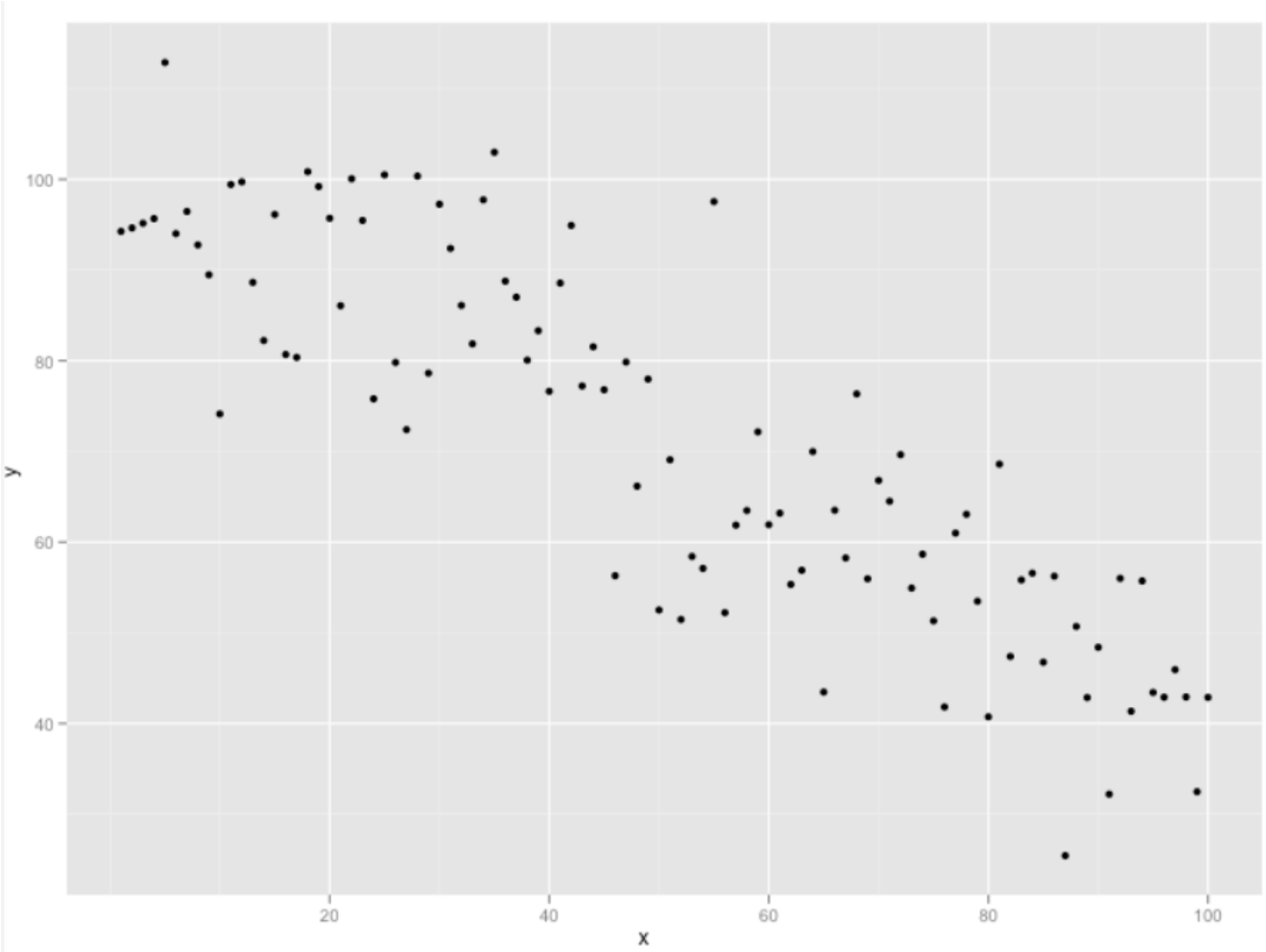
Tableau creates interactive visualizations

Tableau can help you create interactive visualizations without writing code. Many online organizations use it, making it a great tool for you to explore.



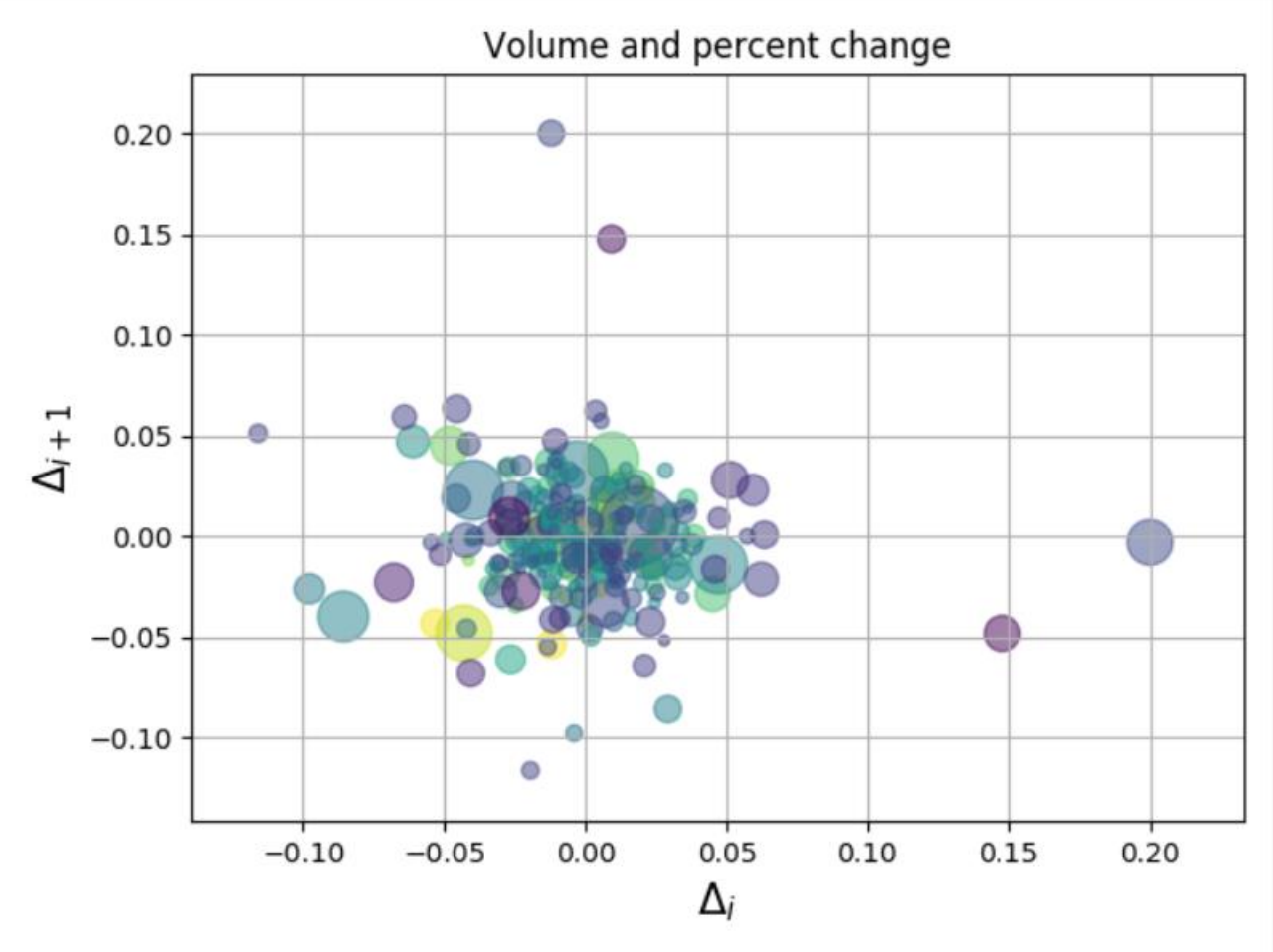
ggplot2 helps with complex data

One of the most popular R workspace tools is ggplot2, which can help visualize data that's too complex for other, less robust programs.



Matplotlib works well with Python

Matplotlib is a popular visualization tool that works with the Python programming language. It helps programmers build charts, graphs, and maps in many different formats.



Can data visualization describe your own passion?

In the What is data science? module, you explored the idea of something you are passionate about, and how data could be applied to that field. Now, try going out on the web and finding a visualization of data that applies to your passion.

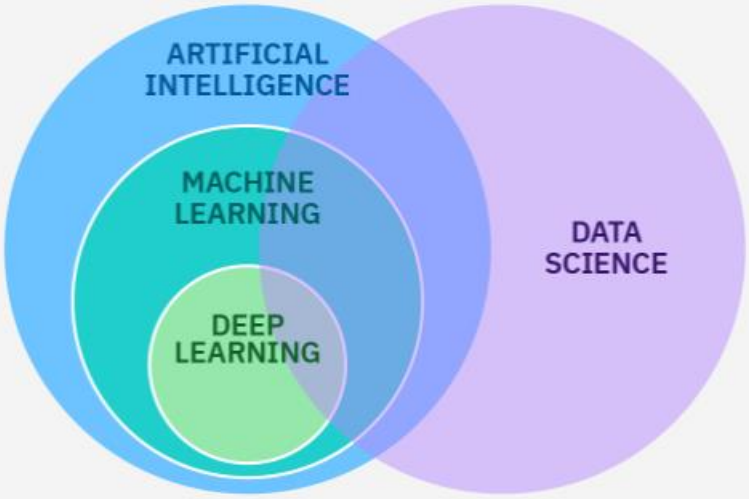
This could be a map of some kind, a chart, a graph, or some other image that helps express data in a way that's different than a simple table. After searching, take some time to note what kind of data visualization you've found. In the following text box, write one thing that you learned from the data visualization.

Enter your response in the text box. (Writing an answer is a good way to process your thoughts. These answers are saved to your computer for your use only.)

Data science drives machine learning

Wondering why you're reading about AI in a data science course? Because machine learning requires huge amounts of data and the ability to extract meaning from that data. That's data science on a very large scale! Today's data scientists use machine learning algorithms and techniques to draw insights from the massive stores of data in the world.

Here are two examples of what we might learn when machine learning breaks down large amounts of data:



- Regression explores how one set of facts or number will change when other related factors or numbers change. For example, machine learning can estimate the appropriate sale price of a new home based on the sales of other similar homes in the area.
- Classification identifies groups hidden within seemingly random sets of information. For example, machine learning can look at video from thousands of street-corner cameras and track the paths taken by people of one particular gender, race, or appearance.

Applications like these would be impossible without the power of modern data science.

Working with data

A day in the life

What's it like to work with data?

Let's explore what careers in data science look like and what the job opportunities are. We'll look at some real-world applications of data science. Then we'll dig into data science jobs and ways that companies are using data. Essentially, the job of a data scientist is to analyze large amounts of raw information to find patterns. That might mean preprocessing data, building models to analyze that data, or presenting information using data visualization techniques. Data science is always about extracting valuable insights from the data.

How do we apply data science?

From simple things like Google search, gaming, and product development to futuristic applications like augmented reality and predictive advertising, data science offers benefits to almost every industry.

Where do I start?

In a previous module, you thought about what you're passionate about and how you might ask data science questions to explore that passion. You'll see your answers below in the text box. Has anything changed based on what you've learned in this course?

Some activities to explore data science further

Try the following activities to see how data science might apply to your life and interests.

Ask a question!

Think of an interesting question that might be answered with data. Here are some examples:

- How many teenagers live in my city?
- What's the average number of assists per game for point guards in the NBA?
- What pop song has been listened to the most over the past six months?

The question can be anything you like, so long as it interests you and you might be able to answer it if you have the right data.

Gather some data!

Sometimes a question comes first. Other times, data scientists find a data set that appears interesting, and then they explore it. If you don't have the perfect question, don't worry! Your next step might be to find some interesting data about your passion. That could spark several great questions.

Lots of industries have publicly available data on the web. Whether you like sports, music, fashion, art, health, gaming, or anything else, start looking for tables that contain structured data that you can explore.

Try a tool!

Perhaps a particular tool spiked your interest? Maybe you would like to start creating graphs using R, or you always wanted to learn more about Python? Or maybe you have your question and your data, and you're ready to start using advanced techniques to pull out insights?
Go ahead and experiment using one of the publicly available data science tools that we talked about in the previous module. Import a table into a tool, build a chart, or take another course on a piece of software. You'll learn (and do) more while you try working on something!

Read a blog!

IBM bloggers are a great resource for a deeper perspective on data science. Check back from time to time to see topics about big data and analytics in action across the globe on IBM's Cloud Computing News site.
Big Data (ibm.com)