

Data Science-I

BCSE 0561**Data Science - I**

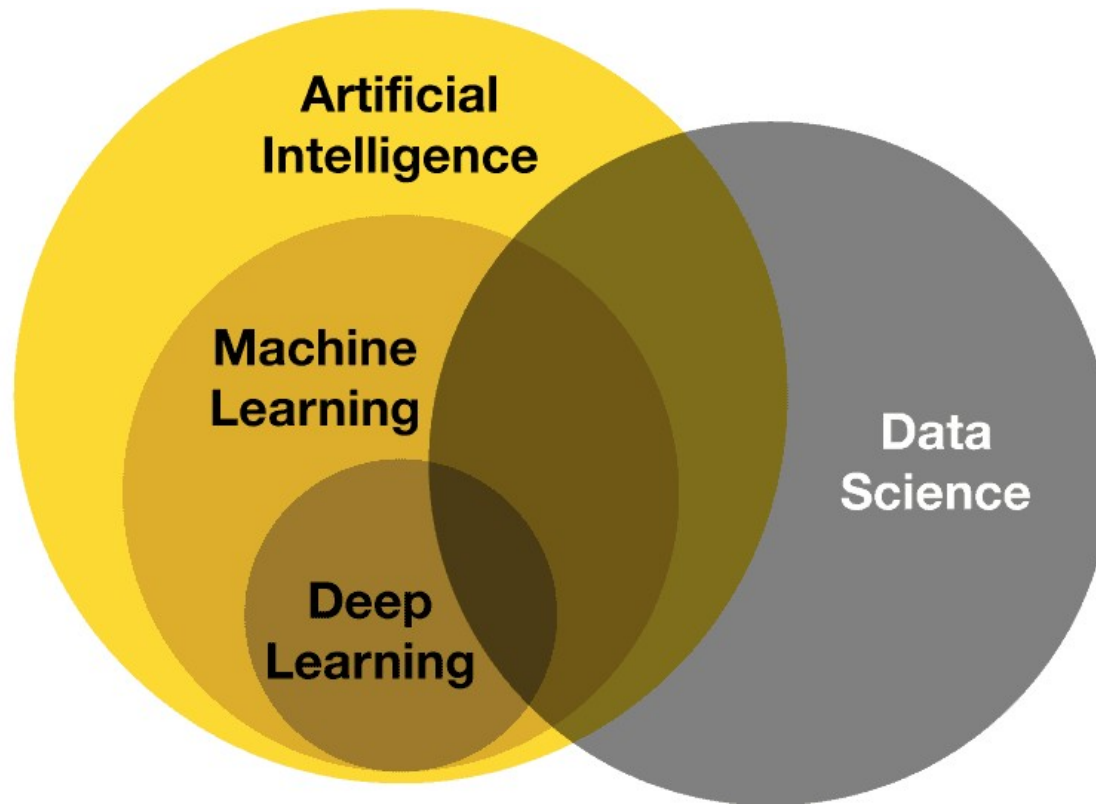
Objective: This course introduces and helps in understand and implement supervised learning techniques.

Credits:03

L-T-P-J:3-0-0-0

Module No.	Content	Teaching Hours
I	<p>Data Science: How to Sound Like a Data Scientist, What is data science?, Basic terminology, Why data science?, Example – Sigma Technologies, The data science Venn diagram, spawner-recruit models,</p> <p>Data science case studies: Automating government paper pushing, marketing dollars, what's in a job description?</p> <p>Types of Data: Structured versus unstructured data, Quantitative versus qualitative data</p> <p>The Five Steps of Data Science: Overview of the five steps, Exploring the data, A Data Scientist's Role in This Process, Thought Experiment: How Would You Simulate Chaos?, Case Study: RealDirect, How Does RealDirect Make Money?</p> <p>Statistical Inference: Statistical Inference, Populations and Samples, Populations and Samples of Big Data, Big Data Can Mean Big Assumptions, Modeling.</p>	20
II	<p>Exploratory Data Analysis, Philosophy of Exploratory Data Analysis, Exercise: EDA Exercise: Real Direct Data Strategy</p> <p>Statistical Learning (Supervised): What is it, why is it useful, what are the main challenges and applications?</p> <p>Regression: Linear regression, polynomial regression, ridge and lasso regression, logistic regression, etc. How to fit, evaluate and compare regression models. How to handle outliers, multicollinearity, overfitting and underfitting, Gradient descent for linear regression.</p> <p>Classification algorithms: Support vector machines (SVM), decision trees, random forests, k-nearest neighbors (kNN), naive Bayes classifier, etc. How to fit, evaluate and compare classification models. How to handle imbalanced data, feature selection, and performance metrics.</p>	20

AI vs ML vs DL vs DS



1. Artificial Intelligence

The term AI was first coined in 1956 by John McCarthy. Artificial Intelligence is a field of study in which machines are programmed and given a cognitive ability to think and mimic actions like humans.

Artificial Intelligence has two different levels:

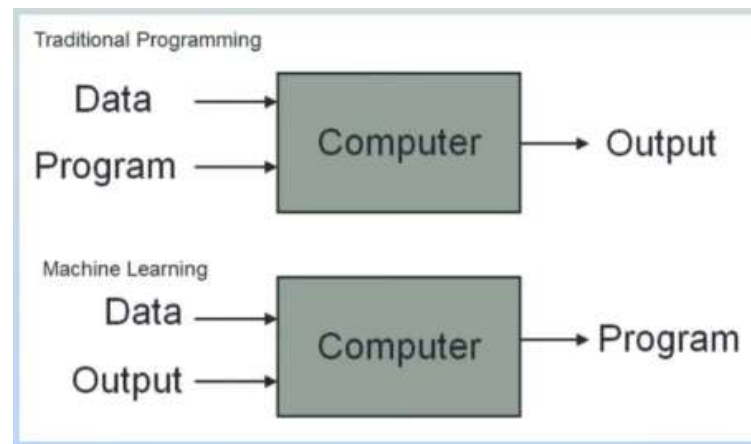
- **General Artificial Intelligence:** It can perform any intellectual task with the same accuracy level as a human would.
- **Narrow Artificial Intelligence:** It can perform a specific task better than a human.

2. Machine Learning (ML)

- The term ML was first coined in 1959 by Arthur Samuel. Machine learning is a subset of AI, here we try to provide ability to machine to learn by itself without getting explicitly programmed.
- In ML we use statistical algorithm such as Decision Tree, Support Vector Machine(SVM), KNN to predict/classify the output. There are three machine learning categories- Supervised Learning, Unsupervised Learning and Reinforcement Learning.

- Machine Learning helps machines to learn by itself without getting explicitly programmed.
- Now what does this “without getting explicitly programmed” means, to understand this we need to look at traditional programming approach where we used to explicitly program every thing and give it to machine, as per the program the machine behaves or gives output. Examples of traditional programming are – Calculator program, To find factorial, To determine a number is odd or even, etc. If we provide the machine with a logic of addition it will always perform addition, it would never learn to perform subtraction until explicitly programmed. Here the machine is not learning it is only giving output as programmed.

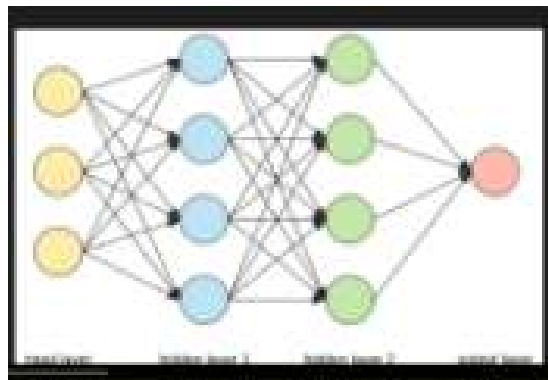
- Machine Learning is a field of Computer Science that uses Statistical techniques to give computer systems the ability to “learn” with data, without being explicitly programmed.



- To successfully implement machine learning we need two things :
Data and Algorithm.
- These are the very important requirements without which machine learning cannot be implemented. Appropriate data helps to generate more accurate models. Now what does this “appropriate” means, here for example if our objective is to classify students on basis of grades, then the appropriate data would be academic data of students which will help in better classifying the students based on grades. But instead of academic data if we provide students personal information data, the model so generated would not be helpful in fulfilling our objective.

3.Deep Learning (DL)

- Deep Learning is a subfield of Artificial Intelligence and Machine Learning that is **inspired by the structure of human brain**.
- Deep Learning Algorithms attempts to draw similar conclusions as humans would by continually analyzing data with a given logical structure called Neural Network.

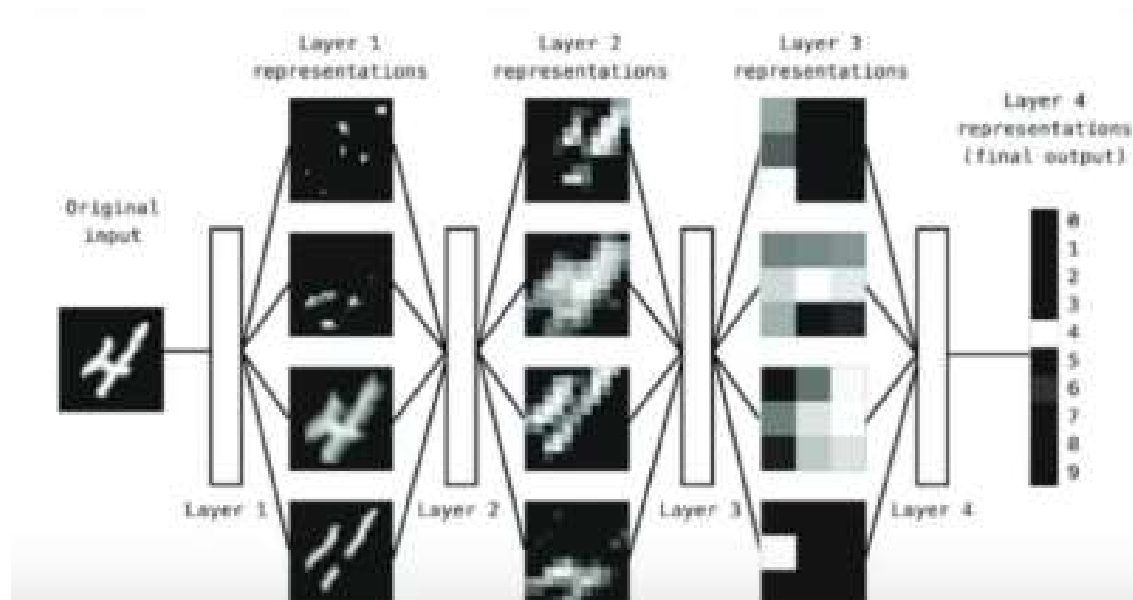


More technically

- DL is part of a broader family of ML methods based on ANN with **representation learning**.
- DL algorithms uses multiple layers to progressively **extract higher level features from the raw input**. For example, in Image Processing, lower layers may identify edges, while higher may identify the concepts relevant to a human such as digits or letters or faces.

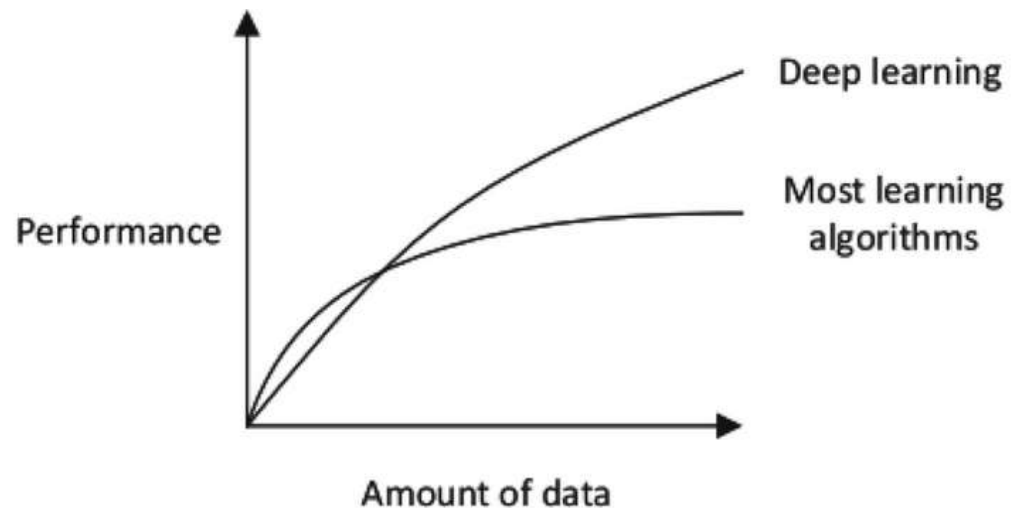
- Deep learning is subset of machine learning. Frank Rosenblatt is the father of DL.
- DL is the next evolution of machine learning. It works in a layered architecture and uses the artificial neural network, a concept inspired by the biological neural network. The human brain usually analyzes and converts the information it receives and tries to identify it from the past information the brain has stored. In a similar way, Deep Learning algorithms are trained to identify patterns and classify various types of information to give the desired output when it receives an input. We need to provide the features manually in Machine Learning. But in Deep Learning, it automatically extracts features for classification which in turn demands a huge amount of data for training DL algorithms. So, in Deep Learning, the accuracy of the output depends on the amount of data. In DL we use Artificial Neural Network(ANN), Convolutional Neural Network(CNN) and Recurrent Neural Network(RNN).

Deeper Layers of DL help in finding more exact patterns. Starting layers helps in finding edges etc., deeper edges find shapes etc.



ML vs DL

- In ML, we have to provide features while in DL, features are automatically extracted.
- DL performance becomes better if we have huge data while ML restrict to a certain limit.



4.Data Science

- Data science is **the study of data to extract meaningful insights for business.**
- It is a multidisciplinary approach that combines principles and practices from the fields of **mathematics**, **statistics**, artificial intelligence, and **computer engineering** to analyze large amounts of data.

Evolution starts from 1960, but becomes useful from 2010 because:-

- Non availability of Data
- Non availability of H/W and Internet
- Non availability of People and Community

Data Science Job Roles

- Data Engineer
- Data Analyst
- Data Scientist
- ML Engineer

1. Data Engineer- He gathers data. He brings the data from multiple sources by using different tools. His job is tough.

Responsibilities

1. Scrap data from the given sources
2. Move/Store the data in optimal servers/warehouses.
3. Build data pipelines/API's for easy access to the data.
4. Handle databases and data warehouses.

Skills Required

1. Strong grasp of algorithms and data structures.
2. Programming Languages(Java/R/Python/Scala) and script writing.
3. Advanced DBMS
4. Big Data Tools(Apache Spark, Hadoop, Apache Kafka, Apache Hive)
5. Cloud platforms (Amazon Web Services, Google Cloud Platform)
6. Distributed Systems
7. Data Pipelines

2. Data Analyst-We already have data that is provided by Data Engineer. Now he analyze that data eg. why profit is less? etc. He looks past to understand why it happened? Analyzing why it happened.

Responsibilities
1.Cleaning and organizing raw data
2.Analyzing data to derive insights
3.Creating data Visualizations
4.Producing and maintaining reports
5.Collaborating with teams/colleagues based on the insight gained
6. Optimizing data collection procedures

Skills Required
1.Statistical Programming
2.Programming languages(R/Python)
3.Creative and Analytical Thinking
4.Business Acumen- Medium to High preferred
5.Strong communications skills
6.Data Mining, Cleaning and Munging
7.Data Visualization
8.Data Story Telling
9.SQL
10.Advanced MS-Excel

3. Data Scientist

- A Data Scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.
- He looks for future. He helps in better prediction.
- Data Scientist=Data Engineer + Data Analyst

4. ML Engineer-His job starts when Model is successfully created.
He deploys/integrate the model to the website

Responsibilities

1. Deploying ML models to production ready environment
2. Scaling and Optimizing the model for production
3. Monitoring and maintenance of deployed models

Skills Required

1. Mathematics
2. Programming Languages (R/Python/Java/Scala)
3. Distributed Systems
4. Data Model and evaluation
5. ML models
6. S/W Engg. & System Design

Comparison among different Job Roles

	Analytical Skills	Business Acumen	Data Story Telling	Soft Skills	S/W Skills
Data Analyst	High	Medium to High	High	Medium to High	Medium
Data Engineer	Medium	Low	Low	Medium	High
Data Scientist	High	High	High	High	Medium
ML Engineer	Medium to High	Medium	Low	High	High

Why is data science important?

- In this data age, **it's clear that we have a surplus of data**. But why should that necessitate an entire new set of vocabulary? What was wrong with our previous forms of analysis? For one, the sheer volume of data makes it literally impossible for a human to parse it in a reasonable time. Data is collected in various forms and from different sources, and often comes in **very unorganized**.
- Data science is important because it combines tools, methods, and technology **to generate meaning from data**.
- Modern organizations are inundated with data; there is a proliferation of **devices that can automatically collect and store information**. Online systems and payment portals capture more data in the fields of e-commerce, medicine, finance, and every other aspect of human life. We have **text, audio, video, and image data available in vast quantities**.

Why is data science important? continues.....

Example-Sigma Technologies

Ben Runkle, CEO, Sigma Technologies, is trying to resolve a huge problem. The company is consistently **losing long-time customers**. He does not know why they are leaving, but he must do something fast. He is convinced that in order to reduce his churn, he must create new products and features, and consolidate existing technologies. To be safe, he calls in his chief data scientist, Dr. Jessie Hughan.

However, she is not convinced that new products and features alone will save the company. Instead, she turns to the transcripts of recent customer service tickets. She shows Runkle the most recent transcripts and finds something surprising:-

- "... Not sure how to export this; are you?"
- "Where is the button that makes a new list?"
- "Wait, do you even know where the slider is?"
- "If I can't figure this out today, it's a real problem..."

It is clear that customers were having problems with the existing UI/UX, and weren't upset due to a lack of features. Runkle and Hughan organized a mass UI/UX overhaul and their sales have never been better

Of course, the science used in the last example was minimal, but it makes a point. We tend to call people like Runkle, a driver. Today's common stick-to-your-gut CEO wants to make all decisions quickly and iterate over solutions until something works. Dr. Haghun is much more analytical. She wants to solve the problem just as much as Runkle, but she turns to user-generated data instead of her gut feeling for answers. Data science is about applying the skills of the analytical mind and using them as a driver would.

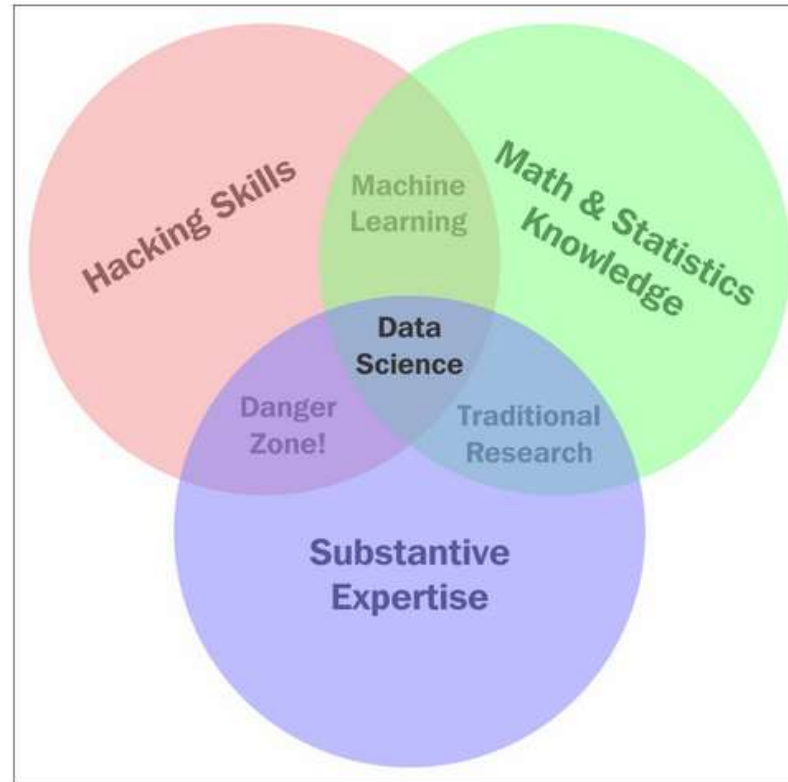
Both of these mentalities have their place in today's enterprises; however, it is Haghun's way of thinking that dominates the ideas of data science—using data generated by the company as her source of information rather than just picking up a solution and going with it.

Data Science Venn Diagram

It is a common misconception that only those with a PhD or geniuses can understand the math/programming behind data science. This is absolutely false. Understanding data science begins with three basic areas:-

- **Math/statistics:** This is the use of equations and formulas to perform analysis.
- **Computer programming:** This is the ability to use code to create outcomes on computer.
- **Domain knowledge:** This refers to understanding the problem domain (medicine, finance, social science, so on).

The following Venn diagram provides a visual representation of how these three areas of data science intersect:



The Venn diagram of data science

- Those with **hacking skills** can conceptualize and program complicated algorithms using computer languages.
- Having a **math and statistics background** allows you to theorize and evaluate algorithms and tweak the existing procedures to fit specific situations.
- Having **substantive expertise (domain expertise)** allows you to apply concepts and results in a meaningful and effective way.

- While having only two of these three qualities can make you intelligent, it will also leave a gap. Let's say that you are **very skilled in coding** and have formal training in day trading(**domain knowledge**). You might create an automated system to trade in your place, but **lack the math skills** to evaluate your algorithms. **This will mean that you end up losing money** in the long run. **It is only when you boost your skills in coding, math, and domain knowledge that you can truly perform data science.**
- The quality that was probably a surprise for you was domain knowledge. It is really just knowledge of the area you are working in. If a financial analyst started analyzing data about heart attacks, they might need the help of a cardiologist to make sense of a lot of the numbers.

- Data science is the intersection of the three key areas mentioned earlier.
- In order to gain knowledge from data, we must be able to utilize computer programming to access the data, understand the mathematics behind the models we derive, and, above all, understand our analyses' place in the domain we are in. This includes the presentation of data.
- If we are creating a model to predict heart attacks in patients, is it better to create a PDF of information, or an app where you can type in numbers and get a quick prediction? All these decisions must be made by the data scientist.

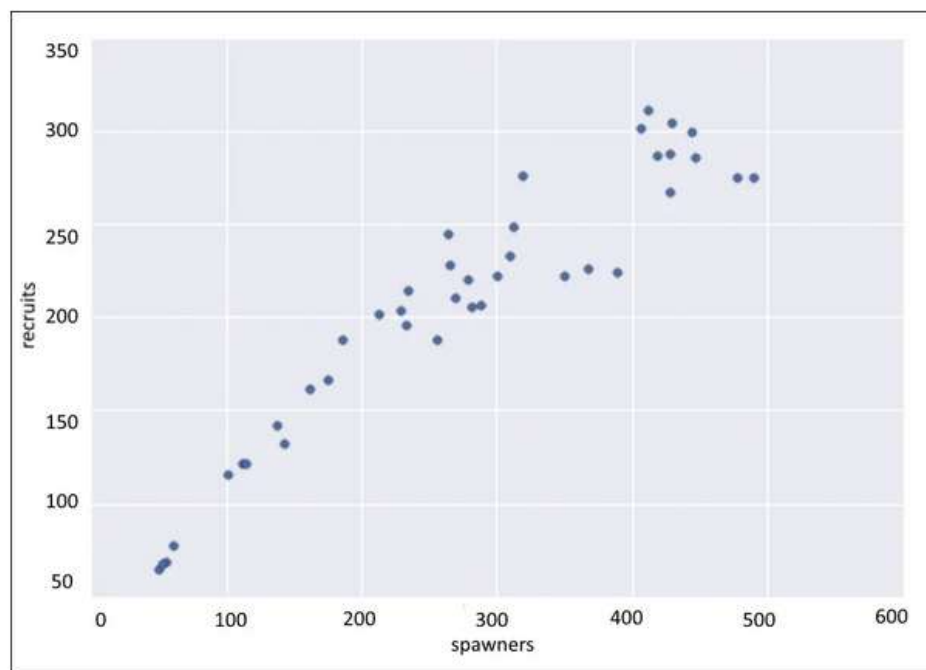
Example:-Spawner-Recruit model

- In biology, we use a model known as the **Spawner-Recruit** model to judge the biological health of a species.
- It is a basic relationship between the number of healthy parental units of a species and the number of new units in the group of animals.
- In a public dataset of the number of salmon spawners and recruits, the following graph was formed to visualize the relationship between the two. We can see that there definitely is some sort of positive relationship (as one goes up, so does the other).

- Essentially, models allow us to plug in one variable to get the other.

$$\text{Recruits} = 0.5 * \text{Spawners} + 60$$

- Let's say we knew that a group of salmon had 1.15 (in thousands) spawners. Then, we would have $\text{Recruits} = 0.5 * 1.15 + 60 = 60.575$
- This result can be very beneficial to estimate how the health of a population is changing. If we can create these models, we can visually observe **how the relationship between the two variables can change.**
- There are many types of data models, including probabilistic and statistical models. Both of these are subsets of a larger paradigm, called machine learning. The essential idea behind these topics is that we use data in order to come up with the best model possible. We no longer rely on human instincts—rather, we rely on data, such as that displayed in the following graph:



The spawner-recruit model visualized

- The purpose of this example is to show how we can define relationships between data elements using mathematical equations.

Data science case studies

Data science case studies

- The combination of math, computer programming, and domain knowledge is what makes data science so powerful. Oftentimes, it is difficult for a single person to master all three of these areas. That's why it's very common for companies to hire teams of data scientists instead of a single person. Let's look at a few powerful examples of data science in action and their outcomes.

Case study –1. Automating government paper pushing

- Social security claims are known to be a major hassle for both the agent reading it and the person who wrote the claim. Some claims take over two years to get resolved in their entirety, and that's absurd! Let's look at the following diagram, which shows what goes into a claim:

B. To be completed by the claimant

PLEASE PRINT

Please Answer the Following Questions:

(1) Have you been treated or examined by a doctor (other than a doctor at a hospital) since the above date? _____ ➔

☐ Yes

☐ No

(If yes, please list the names, addresses and telephone numbers of doctors who have treated or examined you since the above date. Also list the dates of treatment or examination. If possible, send updated reports from these doctors to the Administrative Law Judge before the date of your hearing.)

DOCTORS NAME(S)	ADDRESS(ES) & TELEPHONE NO.(S)	DATE(S)

(2) What have these doctors told you about your condition?

(3) Have you been hospitalized since the above date? _____ ➔

☐ Yes

☐ No

(If yes, please list the name and address of the hospital. Also, explain why you were hospitalized and what treatment you received.)

- Not bad. It's mostly just text, though. Fill this in, then that, then this, and so on. You can see how it would be difficult for an agent to read these all day, form after form. There must be a better way!
- Well, there is. Elder Research Inc. parsed this unorganized data and was able to automate 20% of all disability social security forms. This means that a computer could look at 20% of these written forms and give its opinion on the approval.
- Not only that—the third-party company that is hired to rate the approvals of the forms actually gave the machine-graded forms a higher grade than the human forms. So, not only did the computer handle 20% of the load on average, it also did better than a human.

Fire all humans, right?

- Before I get a load of angry emails claiming that data science is bringing about the end of human workers, keep in mind that the computer was only able to handle 20% of the load. This means that it probably performed terribly on 80% of the forms! This is because the computer was probably great at simple forms. The claims that would have taken a human minutes to compute took the computer seconds. But these minutes add up, and before you know it, each human is being saved over an hour a day!
- Forms that might be easy for a human to read are also likely easy for the computer. It's when the forms are very terse, or when the writer starts deviating from the usual grammar, that the computer starts to fail. This model is great because it lets the humans spend more time on those difficult claims and gives them more attention without getting distracted by the sheer volume of papers.

Case study –2. Marketing dollars

- A dataset shows the relationships between TV, radio, and newspaper sales. The goal is to analyze the relationships between the three different marketing mediums and how they affect the sale of a product. In this case, our data is displayed in the form of a table. Each row represents a sales region, and the columns tell us how much money was spent on each medium, as well as the profit that was gained in that region. For example, from the following table, we can see that in the third region, we spent \$17,200 on TV advertising and sold 9,300 widgets:

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

Advertising budgets' data

- Usually, the data scientist must ask for **units** and the **scale**. In this case, I will tell you that the TV, radio, and newspaper categories are measured in "thousands of dollars" and the sales in "thousands of widgets sold." This means that in the first region, \$230,100 was spent on TV advertising, \$37,800 on radio advertising, and \$69,200 on newspaper advertising. In the same region, 22,100 items were sold.

If we plot each variable against the sales, we get the following graph:

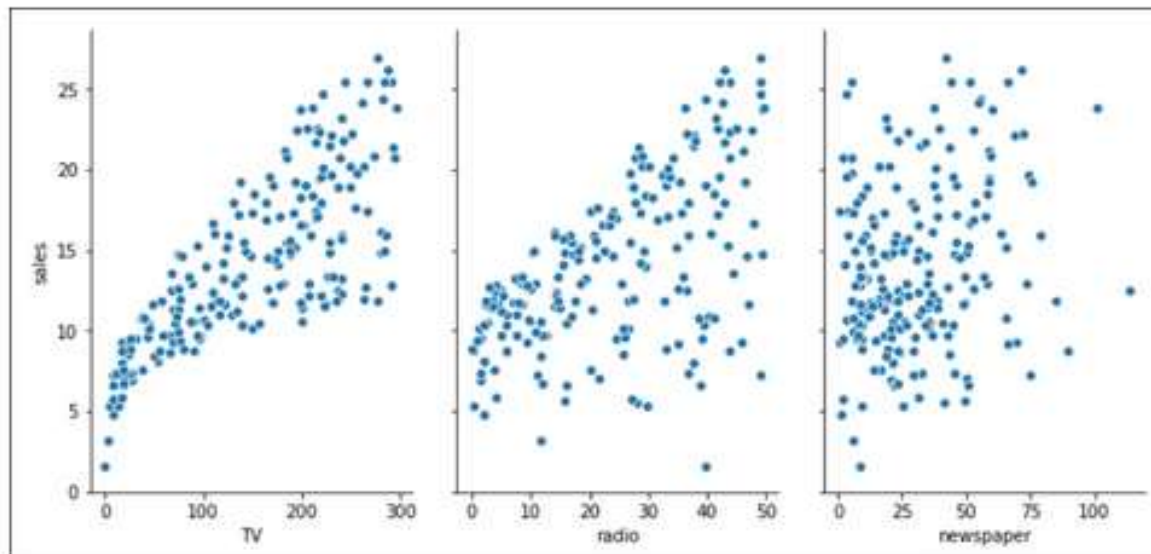
```
import pandas as pd
```

```
import seaborn as sns
```

```
data = pd.read_csv('http://www.bcf.usc.edu/~gareth/ISL/Advertising.csv', index_col=0)
```

```
data.head()
```

```
sns.pairplot(data, x_vars=['TV', 'radio', 'newspaper'], y_vars='sales', height=4.5, aspect=0.7)
```

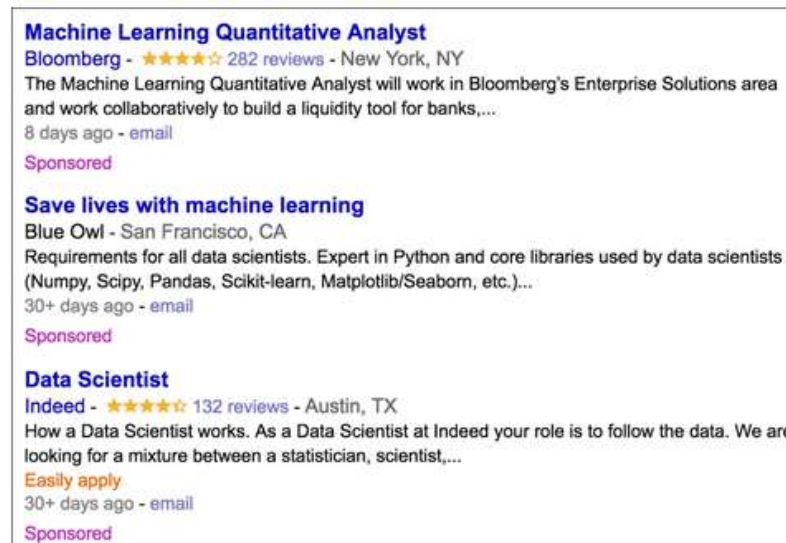


Results – Graphs of advertising budgets

- Note how none of these variables form a very strong line, and that therefore they might not work well in predicting sales on their own. TV comes closest in forming an obvious relationship, but even that isn't great. In this case, we will have to create a more complex model than the one we used in the spawner-recruiter model and combine all three variables in order to model sales.
- This type of problem is very common in data science. In this example, we are attempting to identify key features that are associated with the sales of a product. If we can isolate these key features, then we can exploit these relationships and change how much we spend on advertising in different places with the hope of increasing our sales.

Case study-3. what's in a job description?

- Looking for a job in data science? Great! Let me help. In this case study, I have "scraped" (taken from the web) 1,000 job descriptions for companies that are actively hiring data scientists. The goal here is to look at some of the most common keywords that people use in their job descriptions, as shown in the following screenshot:



An example of data scientist job listings

Types of Data

- Consider an example where we are looking at election results for a country. In the dataset of people, there is a "race" column that is denoted via an identifying number to save space. For example perhaps caucasian is denoted by 7 while Asian American is 2. Without understanding that these numbers are not actually ordered numbers like we think about them (where 7 is greater than 2 and therefore Caucasian is "greater than" Asian American) we will make terrible mistakes in our analysis.

Structured vs unstructured data
or
Organized vs Unorganized data

- **Structured (organized) data:** This is data that can be thought of as observations and characteristics. It is usually organized using a table method(rows and columns).
- **Unstructured (unorganized) data:** This data exists as a free entity and does not follow any standard organization hierarchy.

Here are a few examples that could help you differentiate between the two:

- Most data that exists in text form, including server logs and Facebook posts, is *unstructured*
- Scientific observations, as recorded by careful scientists, are kept in a very neat and organized (*structured*) format
- A genetic sequence of chemical nucleotides (for example, ACGTATTGCA) is *unstructured* even if the order of the nucleotides matters as we cannot form descriptors of the sequence using a row/column format without taking a further look

- Structured data is generally thought of as being much easier to work with and analyze. Most statistical and machine learning models were built with structured data in mind and cannot work on the loose interpretation of unstructured data.
- The natural row and column structure is easy to digest for human and machine eyes. So why even talk about unstructured data? Because it is so common! Most estimates place unstructured data as 80-90% of the world's data. This data exists in many forms and for the most part, goes unnoticed by humans as a potential source of data. Tweets, e-mails, literature, and server logs are generally unstructured forms of data.

- While a data scientist likely prefers structured data, they must be able to deal with the world's massive amounts of unstructured data. If 90% of the world's data is unstructured, that implies that about 90% of the world's information is trapped in a difficult format.
- So, with most of our data existing in this free-form format, we must turn to pre-analysis techniques, called *preprocessing*, in order to apply structure to at least a part of the data for further analysis.
- we attempt to apply transformations to convert unstructured data into a structured counterpart.

Example of data preprocessing

When looking at text data (which is almost always considered unstructured), we have many options to transform the set into a structured format. We may do this by applying new characteristics that describe the data. A few such characteristics are as follows:-


- Word/phrase count
- The existence of certain special characters
- The relative length of text
- Picking out topics

Word/phrase count

- Example of any tweet:-*This Wednesday morn, are you early to rise? Then look East. The Crescent Moon joins Venus & Saturn. Afloat in the dawn skies.*
- We can find **Word/phrase counts** using **sklearn CountVectorizer()**

	this	wednesday	morn	are	this wednesday
Word Count	1	1	1	1	1

✓ [1] `from sklearn.feature_extraction.text import CountVectorizer`

✓  L=["One Geek helps Two Geeks",
"Two Geeks help Four Geeks",
"Each Geek helps many other Geeks at GeeksforGeeks"]

✓ [3] `vectorizer=CountVectorizer()`

✓ [4] `result=vectorizer.fit(L)`


✓ [5] `result`


 [Show hidden output](#)

✓ [6] `result.vocabulary_`

 [Show hidden output](#)

✓ [7] `result1=vectorizer.transform(L)`

✓  `result1.toarray()`

 `array([[0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1],
[0, 0, 1, 0, 2, 0, 1, 0, 0, 0, 0, 1],
[1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0]])`

The existence of certain special characters

- The appearance of these characters might imply certain ideas about the data that are otherwise difficult to know. The fact that this tweet contains a question mark might strongly imply that this tweet contains a question for the reader.

	this	wednesday	morn	are	this wednesday	?
Word Count	1	1	1	1	1	1

The relative length of text

- `len("text")`
- The average tweet, as discovered by analysts, is about 30 characters in length. So, we might impose a new characteristic, called **relative length**, (which is the length of the tweet divided by the average length), telling us the length of this tweet as compared to the average tweet. This tweet is actually 4.03 times longer than the average tweet, as shown:

$$\frac{121}{30} = 4.03$$

	this	wednesday	morn	are	this wednesday	?	Relative length
Word Count	1	1	1	1	1	1	4.03

Picking out topics

- We can pick out some topics of the tweet to add as columns. This tweet is about astronomy, so we can add another column, as illustrated:

	this	wednesday	morn	are	this wednesday	?	Relative length	Topic
Word Count	1	1	1	1	1	1	4.03	astronomy

- Topic is the only extracted feature we looked at that is not automatically derivable from the tweet. Looking at word count and tweet length in Python is easy; however, more advanced models (called topic models) are able to derive and predict topics of natural text as well.

Quantitative versus Qualitative data

- **Quantitative data:** This data can be described using numbers, and basic mathematical procedures, including addition, are possible on the set.
- **Qualitative data:** This data cannot be described using numbers and basic mathematics. This data is generally thought of as being described using "natural" categories and language.

Example – coffee shop data

- Name of coffee shop
- Revenue (in thousands of dollars)
- Zip code
- Average monthly customers
- Country of coffee origin

Each of these characteristics can be classified as either quantitative or qualitative and that simple distinction can change everything.

- Name of coffee shop – Qualitative

The name of a coffee shop is not expressed as a number and we cannot perform math on the name of the shop.

- Revenue – Quantitative

How much money a cafe brings in can definitely be described using a number. Also, we can do basic operations such as adding up the revenue for 12 months to get a year's worth of revenue.

- Zip code – Qualitative

This one is tricky. A zip code is always represented using numbers, but what makes it qualitative is that it does not fit the second part of the definition of quantitative—we cannot perform basic mathematical operations on a zip code. If we add together two zip codes, it is a nonsensical measurement. We don't necessarily get a new zip code and we definitely don't get "double the zip code".

- Average monthly customers – Quantitative

Again, describing this factor using numbers and addition makes sense. Add up all of your monthly customers and you get your yearly customers.

- Country of coffee origin – Qualitative

We will assume this is a very small café with coffee from a single origin. This country is described using a name (Ethiopian, Colombian), and not numbers.

If you are having trouble identifying which is which, basically, when trying to decide whether or not the data is qualitative or quantitative, ask yourself a few basic questions about the data characteristics:

- Can you describe it using numbers?
 - °° No? It is **qualitative**.
 - °° Yes? Move on to next question.
- Does it still makes sense after you add them together?
 - °° No? They are **qualitative**.
 - °° Yes? You probably have **quantitative** data.

- For a quantitative column, you may ask questions such as the following:
 - What is the average value?
 - Does this quantity increase or decrease over time (if time is a factor)?
 - Is there a threshold that if this number grew above or be too low would signal trouble for the company?
- For a qualitative column, none of the preceding questions can be answered; however, the following questions *only* apply to qualitative values:
 - Which value occurs the most and the least?
 - How many unique values are there?
 - What are these unique values?

	country	beer_servings	spirit_servings	wine_servings	total_litres_of_pure_alcohol	continent
0	Afghanistan	0	0	0	0.0	AS
1	Albania	89	132	54	4.9	EU
2	Algeria	25	0	14	0.7	AF
3	Andorra	245	138	312	12.4	EU
4	Angola	217	57	45	5.9	AF

We have six different columns that we are working with in this example:

- country: Qualitative
- beer_servings: Quantitative
- spirit_servings: Quantitative
- wine_servings: Quantitative
- total_litres_of_pure_alcohol: Quantitative
- continent: Qualitative

describe() method on Qualitative data

```
drinks['continent'].describe()
```

```
>> count 193
```

```
>> unique 5
```

```
>> top AF
```

```
>> freq 53
```

describe() method on Quantitative data

```
drinks['beer_servings'].describe()
```

```
>> mean 106.160622
```

```
>> min 0.000000
```

```
>> max 376.000000
```

Now we can look at the mean (average)

Five Steps of Data Science

The five essential steps to perform data science are as follows:

1. Asking an interesting question
2. Obtaining the data
3. Exploring the data
4. Modeling the data
5. Communicating and visualizing the results

Explore the data

- It involves the ability to recognize the different types of data, transform data types, and use code to systemically improve the quality of the entire dataset to prepare it for the modeling stage.
- In order to best represent and teach the art of exploration, we will present several different datasets and use the python package pandas to explore the data.
- There are three basic questions we should ask ourselves when dealing with a new dataset that we may not have seen before. Keep in mind that these questions are not the beginning and the end of data science; they are some guidelines that should be followed when exploring a newly obtained set of data.

Basic questions for data exploration

- Is the data organized or not?
- What does each row represent?
- What does each column represent?
- Are there any missing data points?
- Do we need to perform any transformations on the columns?

Dataset-1: yelp.csv

- The first dataset we will look at is a public dataset made available by the restaurant review site, Yelp.

```
import pandas as pd  
yelp_raw_data = pd.read_csv("yelp.csv")  
yelp_raw_data.head()  
yelp_raw_data.shape
```

	business_id	date	review_id	stars	text	type	user_id	cool	useful	funny
0	9yKzy9PApeiPPOUJEtrvkg	2011-01-26	fWKvX83p0-ka4JS3dc6E5A	5	My wife took me here on my birthday for breakf...	review	rL1i8ZkDX5vH5nAx9C3q5Q	2	5	0
1	ZRJwVLyzEJq1VAihDhYiow	2011-07-27	ljZ33sJrzXqU-0X6U8NwyA	5	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbluQ	0	0	0
2	6oRAC4uyJCsJl1X0WZpVSA	2012-06-14	IESLBzqUCLdSzSqmdCsxQ	4	love the gyro plate. Rice is so good and I als...	review	0hT2KtfllobPvh6cDC8JQg	0	1	0
3	_1QQZuf4zZOyFCvXc0o6Vg	2010-05-27	G-WvGalSbqqaMHInnByodA	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	review	uZetl9T0NcROGOyFfughhg	1	2	0
4	6ozycU1RpktNG2-1BroVtw	2012-01-05	1uJFq2r5QfJG_6ExMRCaGw	5	General Manager Scott Petello is a good egg!!!...	review	vYmM4KTsC8ZfQBg-j5MWkw	0	0	0

- Is the data organized or not?
- What does each row represent?
- What does each column represent?

business_id: **Nominal level**

date: **Ordinal level**

review_id: **Nominal level**

stars: **Ordinal level**

text: **Nominal level**

type: **Nominal level**

user_id: **Nominal level**

- Are there any missing data points?

df.isnull().sum() which will show the number of missing values in each column.

- Do we need to perform any transformations on the columns?

At this point, we are looking for a few things. For example, will we need to change the scale of some of the quantitative data, or do we need to create dummy variables for the qualitative variables? As this dataset has only qualitative columns, we can only focus on transformations at the ordinal and nominal scale.

A Data Scientist's Role in This Process

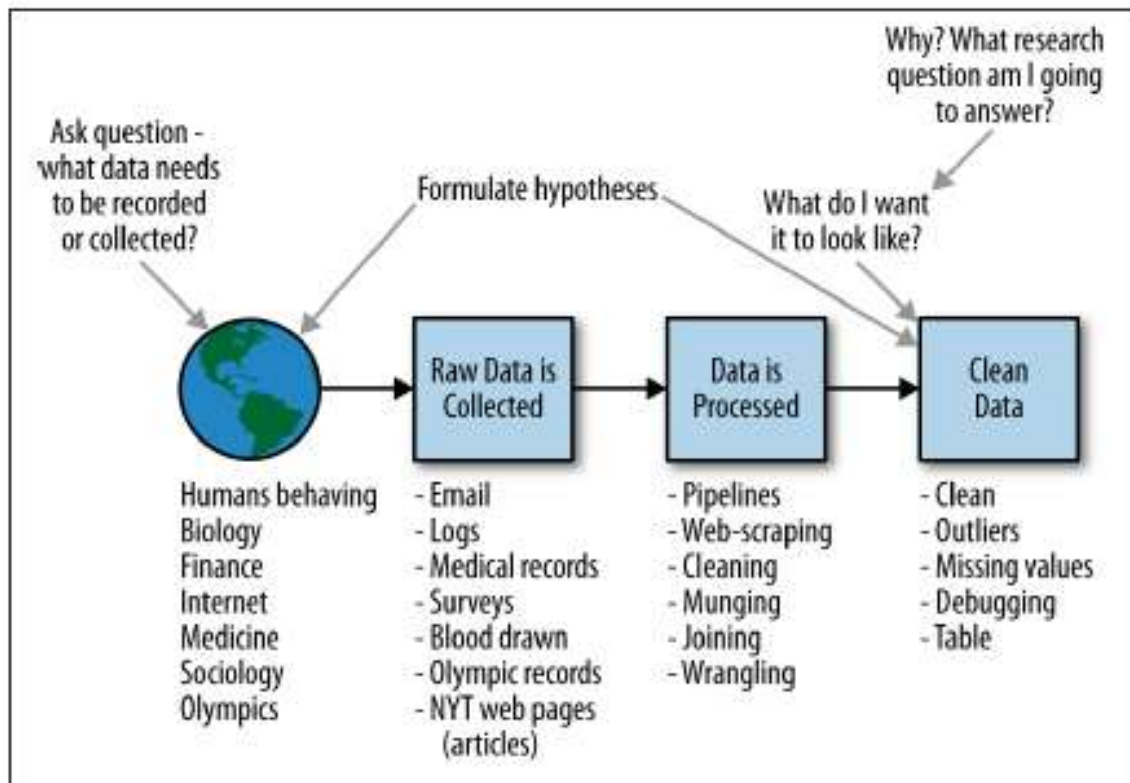


Figure 2-3. The data scientist is involved in every part of this process

Connection to the Scientific Method

We can think of the data science process as an extension of or variation of the scientific method:

- Ask a question.
- Do background research.
- Construct a hypothesis.
- Test your hypothesis by doing an experiment.
- Analyze your data and draw a conclusion.
- Communicate your results.

In both the data science process and the scientific method, not every problem requires one to go through all the steps, but almost all problems can be solved with some combination of the stages. For example, if your end goal is a data visualization (which itself could be thought of as a data product), it's possible you might not do any machine learning or statistical modeling, but you'd want to get all the way to a clean dataset, do some exploratory analysis, and then create the visualization.

Thought Experiment: How Would You Simulate Chaos?

- Most data problems start out with a certain amount of dirty data, ill defined questions, and urgency. As data scientists we are, in a sense, attempting to create order from chaos.

Data scientists might be communicating with domain experts who don't really understand what "logistic regression" means, say, but will pretend to know because they don't want to appear stupid, or because they think they ought to know, and therefore don't ask. But then the whole conversation is not really a successful communication if the two people talking don't really understand what they're talking about. Similarly, the data scientists ought to be asking questions to make sure they understand the terminology the domain expert is using (be it an astrophysicist, a social networking expert, or a climatologist). There's nothing wrong with not knowing what a word means, but there is something wrong with not asking! You will likely find that asking clarifying questions about vocabulary gets you even more insight into the underlying data problem.

Case Study: RealDirect

Doug Perlson, the CEO of RealDirect, has a background in real estate law, startups, and online advertising. His goal with RealDirect is to use all the data he can access about real estate to improve the way people sell and buy houses.

Normally, people sell their homes about once every seven years, and they do so with the help of professional brokers and current data. **But there's a problem both with the broker system and the data quality. RealDirect addresses both of them.**

First, the brokers. They are typically “free agents” operating on their own—think of them as home sales consultants. This means that they guard their data aggressively, and the really good ones have lots of experience. But in the grand scheme of things, that really means they have only slightly more data than the inexperienced brokers.

- RealDirect is addressing this problem by hiring a team of licensed real estate agents who work together and pool their knowledge. To accomplish this, it built an interface for sellers, giving them useful data driven tips on how to sell their house. It also uses interaction data to give real-time recommendations on what to do next.
- Another problem with publicly available data is that it's old news—there's a three-month lag between a sale and when the data about that sale is available. RealDirect is working on real-time feeds on things like when people start searching for a home, what the initial offer is, the time between offer and close, and how people search for a home online.
- Ultimately, good information helps both the buyer and the seller. Atleast if they're honest.

Statistical Inference

- Statistical inference is the discipline that concerns itself with the development of procedures, methods, and theorems that allow us to extract meaning and information from data that has been generated by stochastic (random) processes.

Populations and Samples

- When we take a *sample*, we take a subset of the units of size n in order to examine the observations to draw conclusions and make inferences about the population.
- There are different ways you might go about getting this subset of data, and you want to be aware of this sampling mechanism because it can introduce *biases* into the data, and distort it, so that the subset is not a “mini-me” shrunk-down version of the population. Once that happens, any conclusions you draw will simply be wrong and distorted.

- In the BigCorp email example, you could make a list of all the employees and select 1/10th of those people *at random* and take all the email they ever sent, and that would be your sample. Alternatively, you could sample 1/10th of all email sent each day at random, and that would be your sample. Both these methods are reasonable, and both methods yield the same sample size.

Populations and Samples of Big Data

why would we need to take a sample?

- *Sampling solves some engineering challenges*

In the current popular discussion of Big Data, the focus on enterprise solutions such as Hadoop to handle engineering and computational challenges caused by too much data overlooks sampling as a legitimate solution. At Google, for example, software engineers, data scientists, and statisticians sample all the time.

- How much data you need at hand really depends on what your goal is: for analysis or inference purposes, you typically don't need to store all the data all the time.

New kinds of data

Gone are the days when data is just a bunch of numbers and categorical variables. A strong data scientist needs to be versatile and comfortable with dealing a variety of types of data, including:

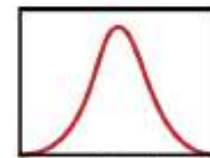
- Traditional: numerical, categorical, or binary
- Text: emails, tweets, *New York Times* articles
- Records: user-level data, timestamped event data, JSON formatted log files
- Geo-based location data
- Network
- Sensor data
- Images

Statistical Modelling

- Statistical modeling is like a formal depiction of a theory. It is typically described as the mathematical relationship between random and non-random variables.
- Statistical modeling is an important process in the field of [data science](#). It involves identifying the best statistical model to identify a relationship in a given dataset, such as census data, public health data, or a company's user data. Think of statistical modeling as a framework. You'll use different frameworks to find different relationships within different datasets.

Distributions

- Normal Distribution
- Gamma Distribution
- Exponential Distribution
- Log Normal Distribution



Normal Distribution



Gamma Distribution

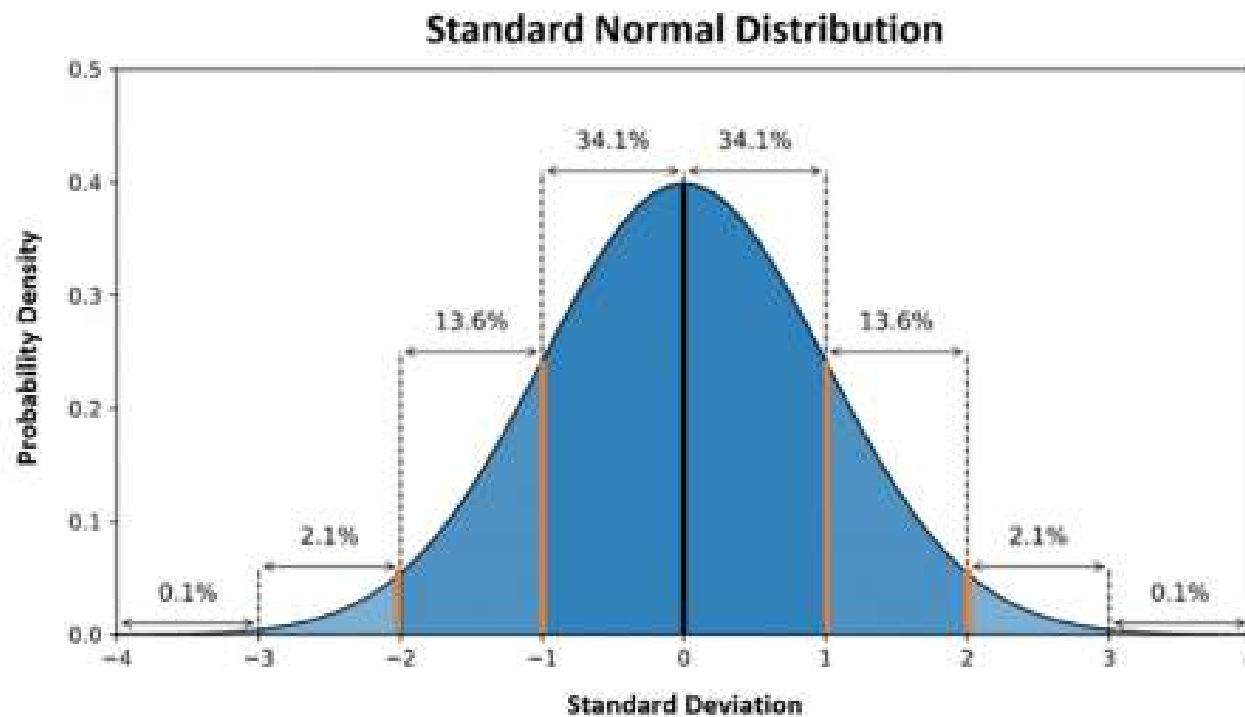


Exponential Distribution

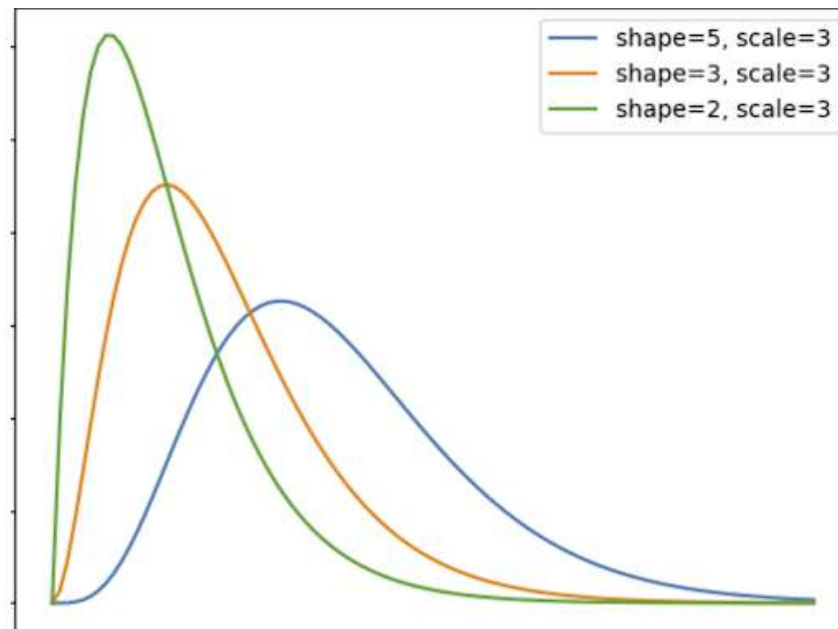


Lognormal Distribution

Normal or Guassian Distribution or Bell curve



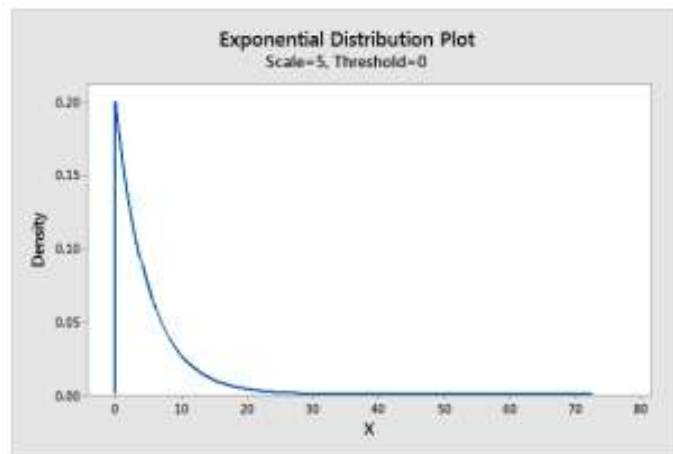
Gamma Distribution



- The Gamma distribution is used to measure continuous variables that possess positive and skewed distributions.
- As a result, this distribution is ideal for modeling the time between events since time is a continuous variable.
- Common variables the Gamma distribution can model include rainfall, wait times, failure times, COVID infection.

Exponential Distribution

- The exponential distribution is a right-skewed continuous probability distribution that models variables in which small values occur more frequently than higher values.
- It is a [unimodal distribution](#) where small values have relatively high probabilities, which consistently decline as data values increase.



Log Normal Distribution

- A distribution of a variable whose logarithm is normally distributed.

Distribution	Parameters	Support	Shape Characteristics
Normal	μ, σ	$(-\infty, \infty)$	Symmetric
Gamma	k, θ or k, β	$X \geq 0$	Depends on k (shape)
Exponential	λ (or θ)	$X \geq 0$	Right-skewed
Log-Normal	μ, σ of $\ln(X)$	$X > 0$	Positively skewed

- **Normal Distribution:**

Shape: Symmetric around the mean.

- **Gamma Distribution:**

Shape: Depends on k

- $k < 1$: Right-skewed.
- $k = 1$: Exponential distribution.
- $k > 1$: Approaches normality as k increases.

- **Exponential Distribution:**

Shape: Always right-skewed.

- **Log Normal Distribution:**
right tail.

Shape: Positively skewed; has a long