SkillCred
by bluetick.ai

DATAWEEK 2024

Helping you get your first taste of real life data application!

Welcome onboard ☺

Welcome to Day one of *DATAWEEK – 2024*! It's great to have you on board 😃

Over the course of next dew days, you'll learn the most used tools by Analysts world over, take on the role of a data analyst and work with a real dataset to solve a business challenge.

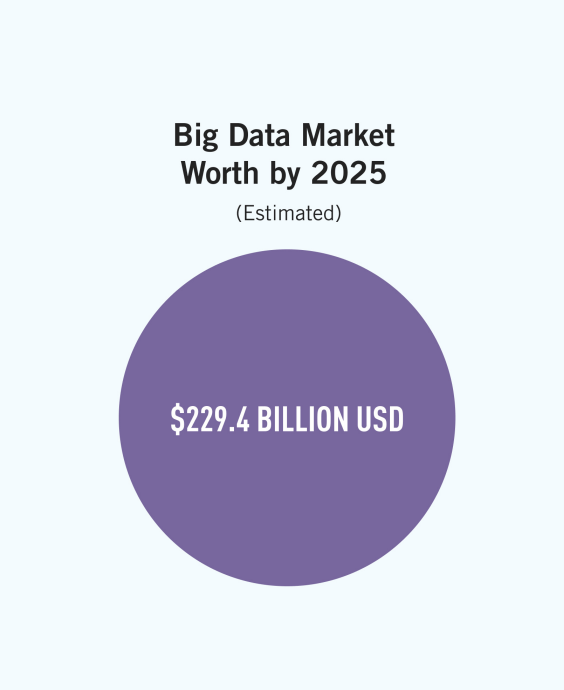
By the end of the week, you will:

- Be familiar with all the key steps in the data analysis process
- Understand, and be able to apply, some fundamental analysis techniques
- Have a first-hand glimpse of what it's like to work as a data analyst

But why a Career in data?

- Is it as big as everybody says it is ?
- What kinds of industries and companies might you work for?
- Is this really a secure career choice with high demand?

Let’s take a look 👀



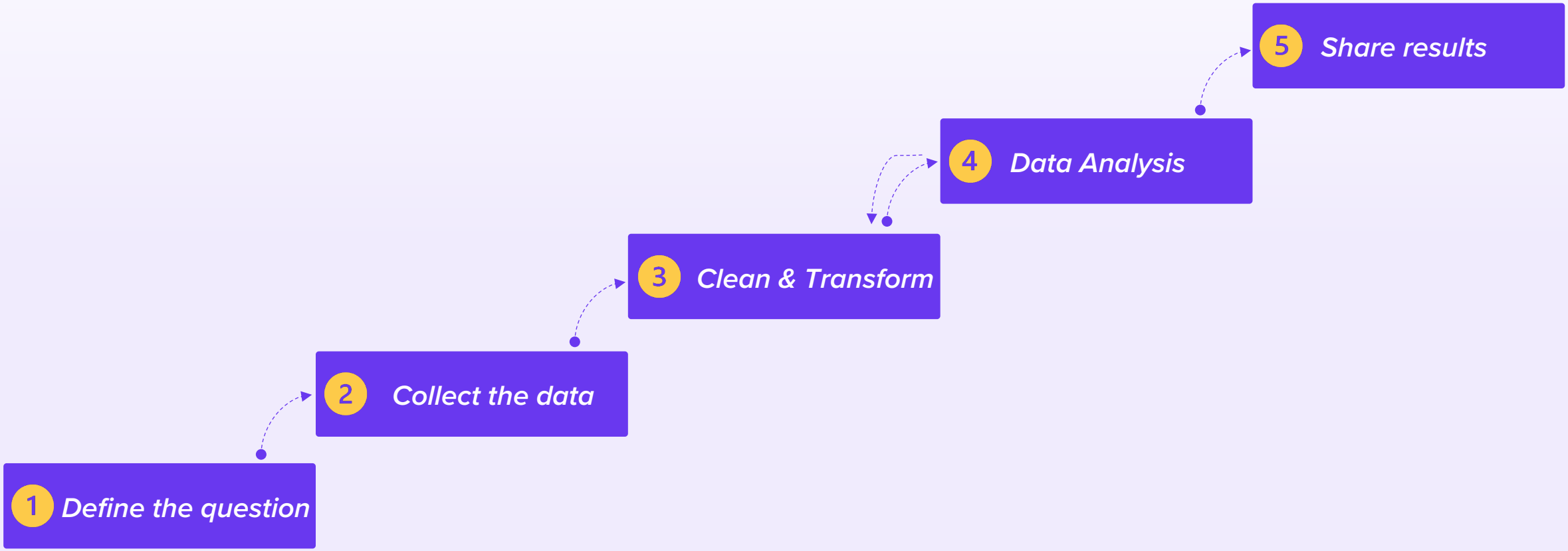
The Jobs of Tomorrow report published by the World Economic Forum in 2020 identifies data and artificial intelligence (AI) as one of seven high-growth emerging professions, showing the highest growth rate at 41% per year.

if you research the most in-demand tech skills for 2024 and beyond, you’ll find that data analytics crops up time and time again

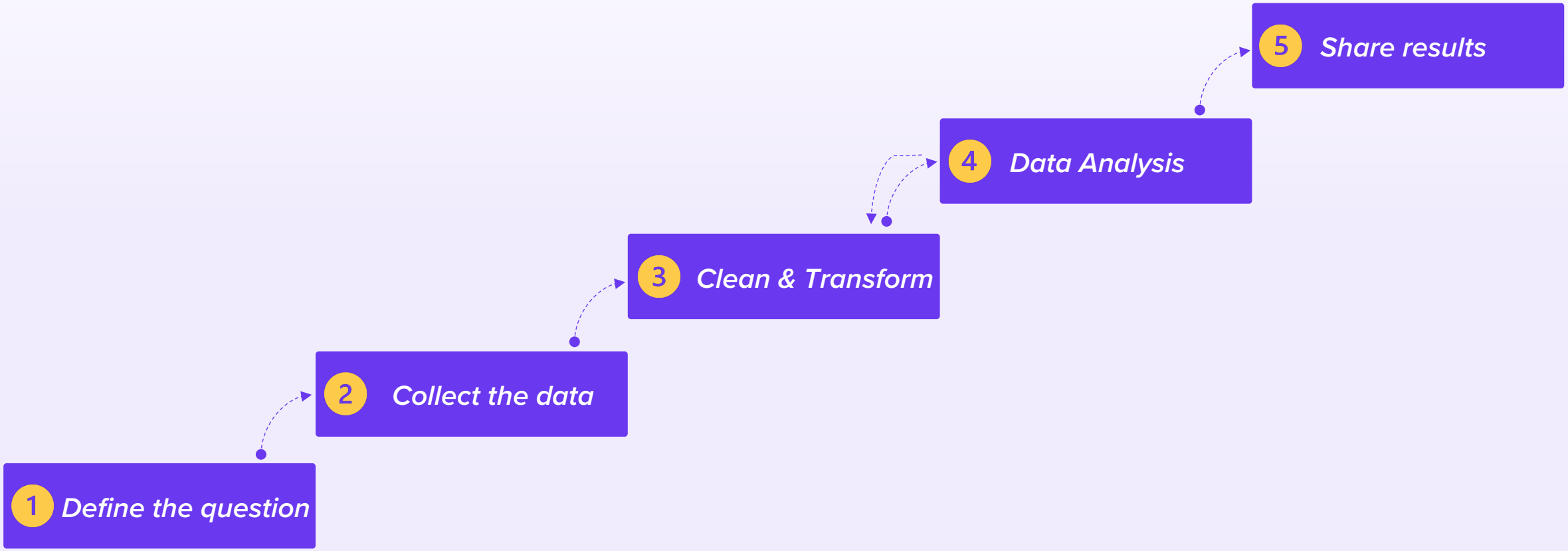
Introduction to Data Analytics

Real life analyst spend over 80% of their time cleaning and transforming data!

What is Data Analytics ?

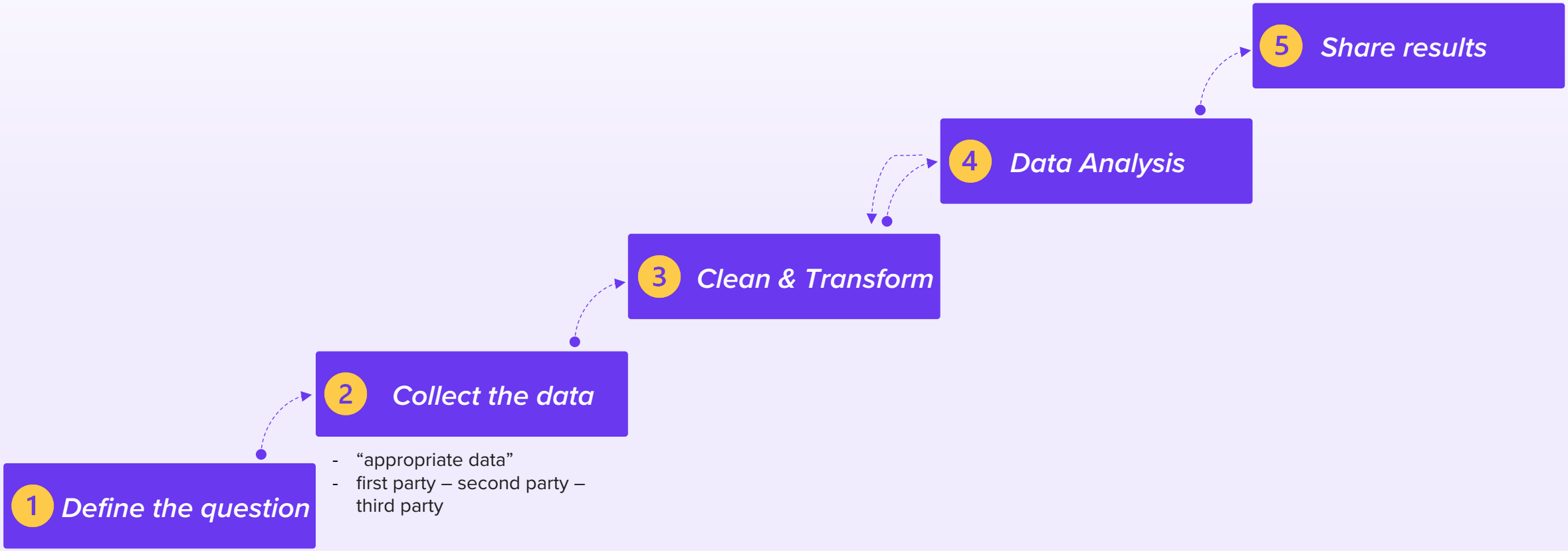


Data Analysis Process

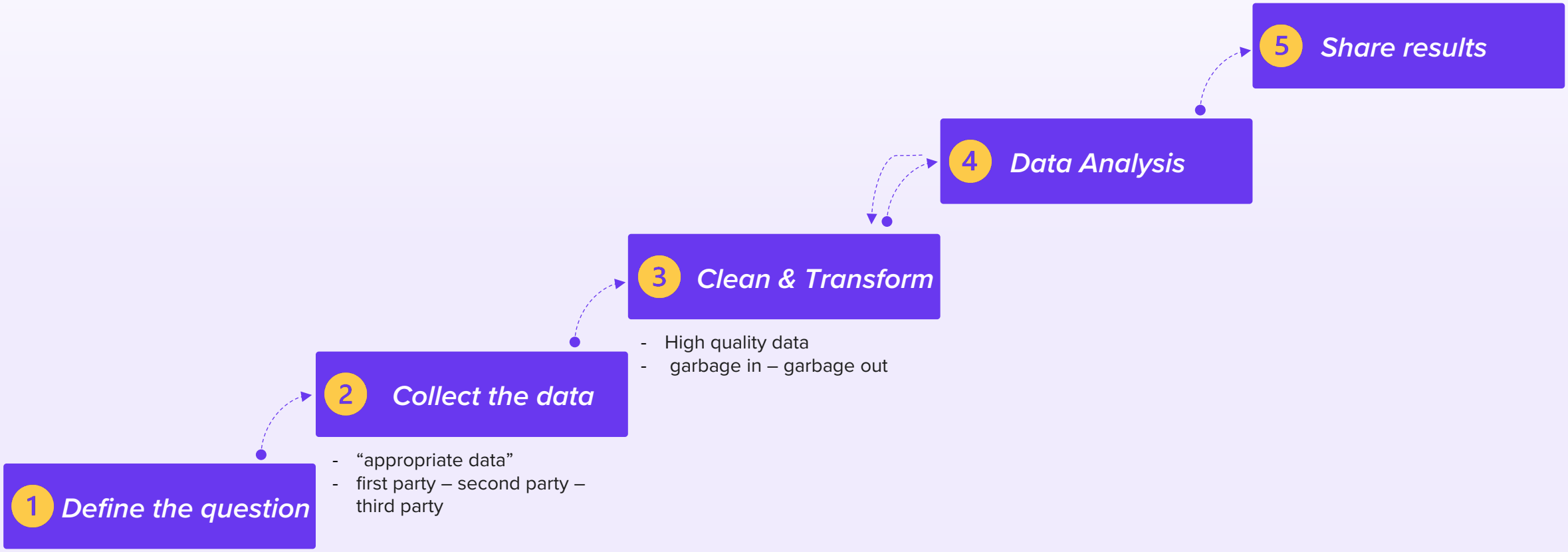


- the ‘problem statement
- Data needed to solve this
- Sometimes there is back & forth

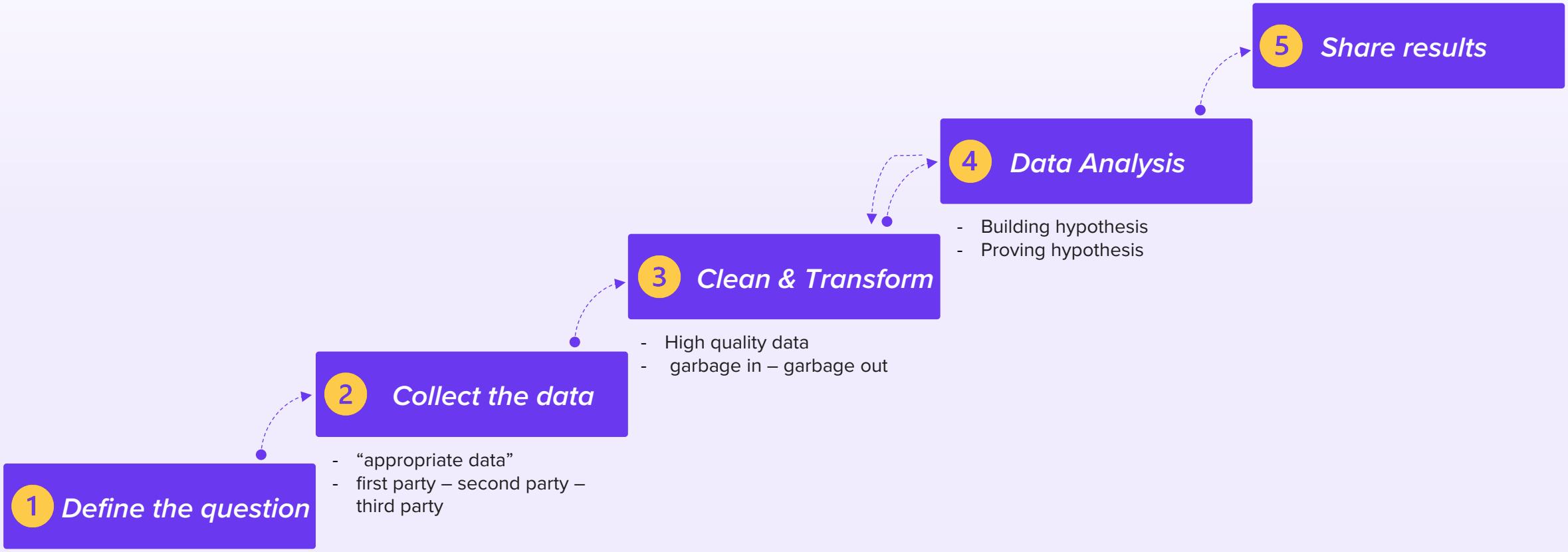
Data Analysis Process



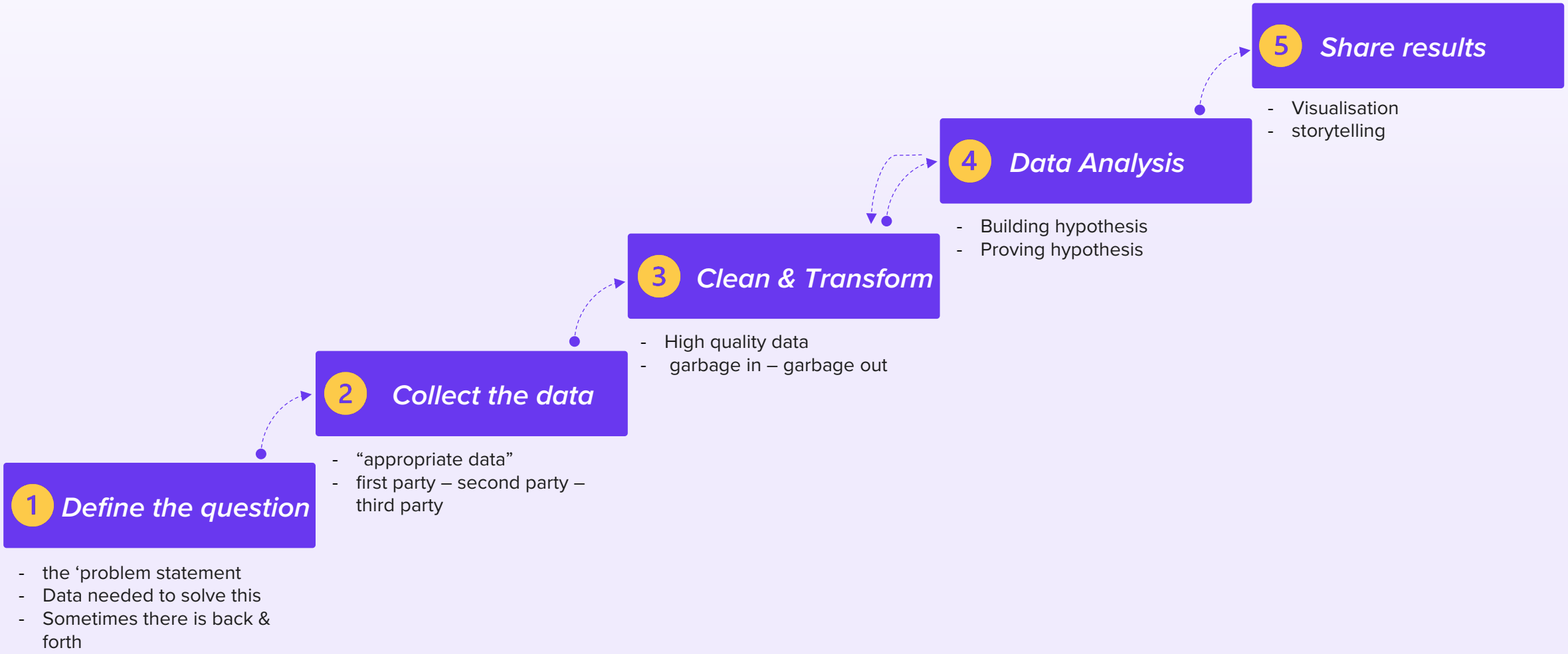
Data Analysis Process



Data Analysis Process



Data Analysis Process



Help *Cinamate compete with Netflix*

Cinemate, an open source online streaming platform boasts an extensive collection of TV shows and movies spanning over a century, offering diverse genres to the viewers. Any boyd with a valid rating can upload the content on Cinemate

You been tasked to extract valuable insights that will aid in understanding planning their operations and marketing activities for next few months.

Company is looking to understand their current library and plan for the operations and events based on what is being uploaded on the product for last few years

show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in
s1	Movie	Dick Johnson	Kirsten Johns	United State	9/25/2021	2020	PG-13	90 min	Documentaries
s3	TV Show	Ganglands	Julien Lecler	France	9/24/2021	2021	TV-MA	1 Season	Crime TV Shows,International TV Shows,TV Action & Adventu
s6	TV Show	Midnight Ma	Mike Flanaga	United State	9/24/2021	2021	TV-MA	1 Season	TV Dramas,TV Horror,TV Mysteries
s14	Movie	Confessions	Bruno Garott	Brazil	9/22/2021	2021	TV-PG	91 min	Children & Family Movies,Comedies
s8	Movie	Sankofa	Haile Gerim	United State	9/24/2021	1993	TV-MA	125 min	Dramas,Independent Movies,International Movies
s9	TV Show	The Great Bri	Andy Devons	United Kingd	9/24/2021	2021	TV-14	9 Seasons	British TV Shows,Reality TV
s10	Movie	The Starling	Theodore Me	United State	9/24/2021	2021	PG-13	104 min	Comedies,Dramas
s939	Movie	Motu Patlu in	Suhas Kadav	India	05/01/21	2019	TV-Y7	87 min	Children & Family Movies,Comedies,Music & Musicals
s13	Movie	Je Suis Karl	Christian Sc	Germany	9/23/2021	2021	TV-MA	127 min	Dramas,International Movies
s940	Movie	Motu Patlu in	Suhas Kadav	India	05/01/21	2013	TV-Y7	76 min	Children & Family Movies,Music & Musicals
s941	Movie	Motu Patlu: D	Suhas Kadav	India	05/01/21	2014	TV-Y7	76 min	Children & Family Movies,Comedies
s942	Movie	Motu Patlu: N	Suhas Kadav	India	05/01/21	2013	TV-Y7	71 min	Children & Family Movies,Comedies
s852	Movie	99 Songs (Tamil)		Pakistan	5/21/2021	2021	TV-14	131 min	Dramas,International Movies,Music & Musicals
s471	Movie	Bridgerton - T	Krysia Plonka	United State	7/13/2021	2021	TV-14	39 min	Movies
s730	Movie	Bling Empire	Krysia Plonka	United State	06/12/21	2021	TV-MA	36 min	Movies
s731	Movie	Cobra Kai - Tl	Krysia Plonka	United State	06/12/21	2021	TV-MA	34 min	Movies
s913	Movie	The Circle - T	Krysia Plonka	United State	05/07/21	2021	TV-14	35 min	Comedies
s4	TV Show	Jailbirds New	Orleans	Pakistan	9/24/2021	2021	TV-MA	1 Season	Docuseries,Reality TV
s15	TV Show	Crime Stories:	India Detec	Pakistan	9/22/2021	2021	TV-MA	1 Season	British TV Shows,Crime TV Shows,Docuseries
s3232	Movie	True: Winter '	Mark Thornto	United State	11/26/2019	2019	TV-Y	46 min	Children & Family Movies
s4832	TV Show	True: Magica	Mark Thornto	United State	6/15/2018	2018	TV-Y	1 Season	Kids' TV
s4833	TV Show	True: Wonde	Mark Thornto	United State	6/15/2018	2018	TV-Y	1 Season	Kids' TV
s4857	TV Show	Dance & Sing	Mark Thornto	United State	5/18/2018	2018	TV-Y	1 Season	Kids' TV
s7	Movie	My Little Pon	Robert Cullen, Jos	√É¬© Lu	9/24/2021	2021	PG	91 min	Children & Family Movies
s12	TV Show	Bangkok Bre	Kongkiat Komesiri		9/23/2021	2021	TV-MA	1 Season	Crime TV Shows,International TV Shows,TV Action & Adventu
s17	Movie	Europe's Mos	Pedro de Echave Garc	√É¬	9/22/2021	2020	TV-MA	67 min	Documentaries,International Movies
s7930	Movie	Samudri Loot	Anirban Majumder		6/18/2019	2018	TV-Y	65 min	Children & Family Movies
s21	TV Show	Monsters Ins	Olivier Megat	United State	9/22/2021	2021	TV-14	1 Season	Crime TV Shows,Docuseries,International TV Shows
s24	Movie	Go! Go! Cory	Alex Woo, Sta	United State	9/21/2021	2021	TV-Y	61 min	Children & Family Movies
s25	Movie	Jeans	S. Shankar	India	9/21/2021	1998	TV-14	166 min	Comedies,International Movies,Romantic Movies
s28	Movie	Grown Ups	Dennis Duga	United State	9/20/2021	2010	PG-13	103 min	Comedies
s29	Movie	Dark Skies	Scott Stewar	United State	9/19/2021	2013	PG-13	97 min	Horror Movies,Sci-Fi & Fantasy
s30	Movie	Paranoia	Robert Luket	United State	9/19/2021	2013	PG-13	106 min	Thrillers
s20	TV Show	Jaguar		Pakistan	9/22/2021	2021	TV-MA	1 Season	International TV Shows,Spanish-Language TV Shows,TV Acti
s32	TV Show	Chicago Party Aunt		Pakistan	9/17/2021	2021	TV-MA	1 Season	TV Comedies
s34	TV Show	Squid Game		Pakistan	9/17/2021	2021	TV-MA	1 Season	International TV Shows,TV Dramas,TV Thrillers
s35	TV Show	Tayo and Little Wizards		Pakistan	9/17/2021	2020	TV-Y7	1 Season	Kids' TV
s75	TV Show	The World's Most Amazing		Pakistan	9/14/2021	2021	TV-PG	2 Seasons	Reality TV
s84	TV Show	Metal Shop Masters		Pakistan	09/10/21	2021	TV-MA	1 Season	Reality TV
s86	TV Show	Pok√É¬©mon Master Jourr		Pakistan	09/10/21	2021	TV-Y7	1 Season	Anime Series,Kids' TV
s88	TV Show	Titipo Titipo		Pakistan	09/10/21	2019	TV-Y	2 Seasons	Kids' TV,Korean TV Shows
s90	TV Show	Mighty Raju		Pakistan	09/09/21	2017	TV-Y7	4 Seasons	Kids' TV
s101	TV Show	Tobot Galaxy Detectives		Pakistan	09/07/21	2019	TV-Y7	2 Seasons	Kids' TV
s122	TV Show	Hotel Del Luna		Pakistan	09/02/21	2019	TV-14	1 Season	International TV Shows,Romantic TV Shows,TV Comedies
s133	TV Show	Brave Animated Series		Pakistan	09/01/21	2021	TV-MA	1 Season	International TV Shows,TV Action & Adventure,TV Comedies
s148	TV Show	How to Be a Cowboy		Pakistan	09/01/21	2021	TV-PG	1 Season	Reality TV
s166	TV Show	Oldsters		Pakistan	09/01/21	2019	TV-MA	1 Season	Crime TV Shows,International TV Shows,Spanish-Language
s190	TV Show	Bread Barbershop		Pakistan	8/28/2021	2020	TV-Y	2 Seasons	Kids' TV,TV Comedies
s182	TV Show	Turning Point: 9/11 and the		Pakistan	09/01/21	2021	TV-14	1 Season	Docuseries

Data Cleanup & Transformation

Real life analyst spend over 60% of their time cleaning and transforming data!

What is Data Cleanup ?

- *Irrelevant Data*
- *Structural Errors*
- *Duplicates*
- *Missing Data*
- *Outliers*

What is Data Cleanup ?

- *Irrelevant Data*
- *Structural Errors*
- *Duplicates*
- *Missing Data*
- *Outliers*

Remove distraction and noise → Make sure that the data you're including really needs to be there.

For example, if you are collecting data on women between the ages of 18-35, there is no reason for a 60-year-old man to appear in your data set.

- Personally identifiable (PII) data
- URLs
- HTML tags
- Boilerplate text (such as in emails)
- Tracking codes
- Excessive blank space between text

What is Data Cleanup ?

- *Irrelevant Data*
- ***Structural Errors***
- *Duplicates*
- *Missing Data*
- *Outliers*

Structural errors in your data include things like typos, inconsistent formatting, incorrect capitalization, and any spelling issues or formatting that might confuse a machine learning model

- Typos like spelling out a date rather than using a number
- Standardizing date and time formats or units of measurement.
- Standardizing capitalisation
- Numbers as texts
- unnecessary punctuation in data such as email addresses

What is Data Cleanup ?

- *Irrelevant Data*
- *Structural Errors*
- ***Duplicates***
- *Missing Data*
- *Outliers*

When you collect or scrape data from various sources, there's a good chance you'll end up with duplicate items. These duplicates could result from human error, such as an error committed by the individual entering data or when filling out a form.

Duplicates will significantly alter your data and/or cause confusion in your results.

They can also make data difficult to interpret when you try to visualize it, so it's preferable to get rid of them as soon as possible.

What is Data Cleanup ?

- *Irrelevant Data*
- *Structural Errors*
- *Duplicates*
- ***Missing Data***
- *Outliers*

3 possibilities when it comes to incomplete data:

- Remove all observations with missing values.
- Fill in the blanks.
- Leave blanks as-is.

What you do depends on your analytical aims and what you want from the data!

What is Data Cleanup ?

- *Irrelevant Data*
- *Structural Errors*
- *Duplicates*
- *Missing Data*
- *Outliers*

An outlier is a minority data point that varies greatly from the majority of the other data.

Outliers are not incorrect, but they may give an inaccurate representation of your data if you take them into account.

We discuss this more during Exploratory data analysis!

Data Cleanup with Excel

1. Import the data from an external data source.
2. Create a backup copy of the original
3. Ensure that the data is in a tabular format of rows and columns with: similar data in each column, all columns and rows visible, and no blank rows within the range. For best results, use an Excel table.
4. Do tasks that don't require column manipulation first, such as spell-checking or using the Find and Replace dialog box.
5. Next, do tasks that do require column manipulation :
   - Insert a new column (B) next to the original column (A) that needs cleaning.
   - Add a formula that will transform the data at the top of the new column (B).
   - Fill down the formula in the new column (B). In an Excel table, a calculated column is automatically created with values filled down.
   - Select the new column (B), copy it, and then paste as values into the new column (B).
   - Remove the original column (A), which converts the new column from B to A.

Spell checking
Removing duplicate rows
Finding and replacing text
Changing the case of text
Removing spaces and nonprinting characters from text
Fixing numbers and number signs
Fixing dates and times
Merging and splitting columns
Transforming and rearranging columns and rows
Reconciling table data by joining or matching
Third-party providers

Let’s clean this data

Check each column one by one and make sure to understand what is happening.

What do you see odd and why ?

Plan for things before you start making changes.

1. Import the data from an external data source.
2. Create a backup copy of the original
3. Ensure that the data is in a tabular format of rows and columns with: similar data in each column, all columns and rows visible, and no blank rows within the range. For best results, use an Excel table.
4. Do tasks that don't require column manipulation first, such as spell-checking or using the Find and Replace dialog box.
5. Next, do tasks that do require column manipulation :
   - Insert a new column (B) next to the original column (A) that needs cleaning.
   - Add a formula that will transform the data at the top of the new column (B).
   - Fill down the formula in the new column (B). In an Excel table, a calculated column is automatically created with values filled down.
   - Select the new column (B), copy it, and then paste as values into the new column (B).
   - Remove the original column (A), which converts the new column from B to A.

→ *Irrelevant Data*

→ *Structural Errors*

→ *Duplicates*

→ *Missing Data*

→ *Outliers*

Data Transformation : Organizing / Shaping data

- *Smoothing*
- *Attribute Construction*
- *Generalization*
- *Aggregation*
- *Normalization*
- *Discretization*

Data Transformation

Real life analyst spend over 60% of their time cleaning and transforming data!

Too Advanced for now!

Data Transformation : Smoothing

Smoothing is a technique where you apply an algorithm in order to remove noise from your dataset when trying to identify a trend. Noise can have a bad effect on your data and by eliminating or reducing it you can extract better insights or identify patterns that you wouldn't see otherwise.

There are 3 algorithm types that help with data smoothing:

- **Clustering**: Where you can group similar values together to form a cluster while labeling any value out of the cluster as an outlier.

- **Binning**: Using an algorithm for binning will help you split the data into bins and smooth the data value within each bin.

- **Regression**: Regression algorithms are used to identify the relation between two dependent attributes and help you predict an attribute based on the value of the other.

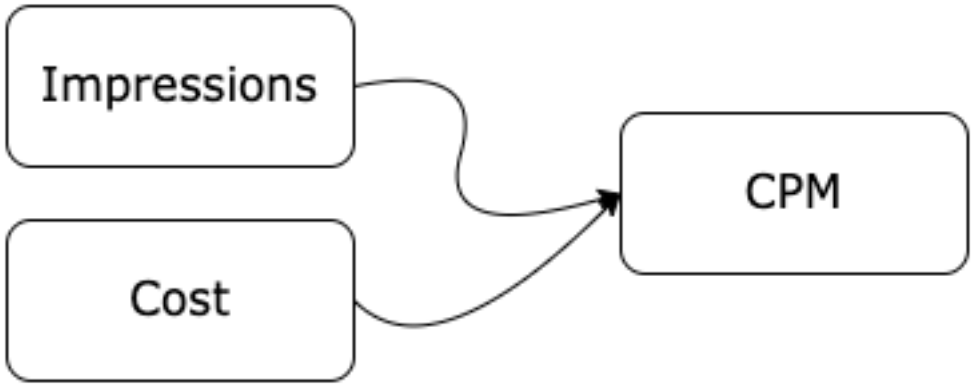
Data Transformation : Attribute Construction

Attribution construction is one of the most common techniques in data transformation pipelines.

Attribution construction or feature construction is the process of creating new features from a set of existing features/attributes in the dataset.

Imagine working in marketing and trying to analyze the performance of a campaign. You have all the impressions that your campaign generated and the total cost for the given time frame.

Instead of trying to compare these two metrics across all of your campaigns, you can construct another metric to calculate the cost per million impressions or CPM.
This will make your data mining and analysis process a lot easier, as you'll be able to compare the campaign performance on a single metric rather than two separate metrics.



Data Transformation : Data Generalization

Data generalization refers to the process of transforming low-level attributes into high-level ones by using the concept of hierarchy.

Data generalization is applied to categorical data where they have a finite but large number of distinct values.

This is something that we, as people, are already doing without noticing and it helps us get a clearer picture of the data.

For ex. Address is divided into 4 categorical attributes :

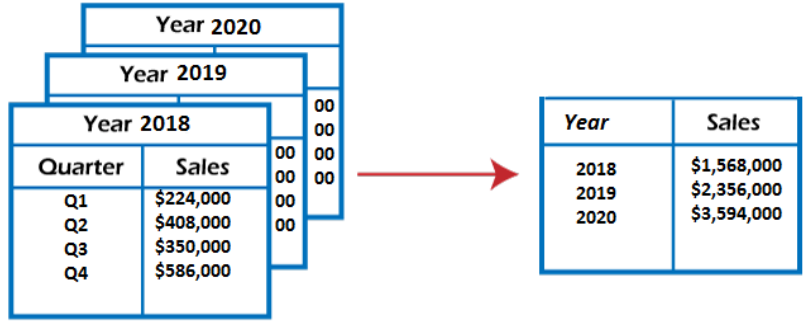
- City
- Street
- Country
- State/province.

Data Transformation : Aggregation

Data aggregation is possibly one of the most popular techniques in data transformation. When you're applying data aggregation to your raw data you are essentially storing and presenting data in a summary format.

This is ideal when you want to perform statistical analysis of your data as you might want to aggregate your data over a specific time period and provide statistics such as average, sum, minimum, and maximum

	A	B
1	**Date**	**Temperature, C**
2	1981-01-01	20.7
3	1981-01-02	17.9
4	1981-01-03	18.8
5	1981-01-04	14.6
6	1981-01-05	15.8
7	1981-01-06	15.8
8	1981-01-07	15.8
9	1981-01-08	17.4
10	1981-01-09	21.8
11	1981-01-10	20
12	1981-01-11	16.2
13	1981-01-12	13.3
14	1981-01-13	16.7
15	1981-01-14	21.5
16	1981-01-15	25
17	1981-01-16	20.7



Date - Month	AVERAGE of Temprature
Jan	17.14032258
Feb	16.8
Mar	14.21774194
Apr	11.91333333
May	9.537096774
Jun	6.456666667
Jul	6.109677419
Aug	7.290566038
Sep	10.14333333
Oct	10.08709677
Nov	11.89
Dec	13.68064516

Too Advanced for now!

Data Transformation : Normalization

process of scaling the data to a much smaller range, without losing information to help minimize or exclude duplicated data and improve algorithm efficiency and data extraction performance.

There are three methods to normalize an attribute:

- **Min-max normalization**: Where you perform a linear transformation on the original data.
- **Z-score normalization**: In z-score normalization (or zero-mean normalization) you are normalizing the value for attribute A using the mean and standard deviation.
- **Decimal scaling**: Where you can normalize the value of attribute A by moving the decimal point in the value.

Normalization methods are frequently used when you have values that skew your dataset and you find it hard to extract valuable insights.

Data Transformation : Discretization

Data discretization refers to the process of transforming continuous data into a set of data intervals. This is an especially useful technique that can help you make the data easier to study and analyze and improve the efficiency of any applied algorithm.

Imagine having tens of thousands of rows representing people in a survey providing their first name, last name, age, and gender.

Age is a numerical attribute that can have a lot of different values. To make our life easier we can divide the range of this continuous attribute into intervals.

Mapping this attribute to a higher-level concept, like youth, middle-aged, and senior, can help a lot with the efficiency of the task and improve the speed of the algorithms applied.

Make this clean data useful now!

Now that you have clean data, let us see how can we make it more useful.

You know what to do!

show_id	sho_id_cl	type	title	director	country	date_added_cl	release_year	rating	duration_cl	listed_in_unclean	listed_in_cl1	listed_in_cl2	listed_in_cl3
s1	1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	25/09/21	2020	PG-13	90	Documentaries	Documentaries		
s3	3	TV Show	Ganglands	Julien Leclercq	France	24/09/21	2021	TV-MA	1 Season	Crime TV Shows,Interna	Crime TV Shows	International TV Shows	TV Action & Adventure
s6	6	TV Show	Midnight Mass	Mike Flanagan	United States	24/09/21	2021	TV-MA	1 Season	TV Dramas,TV Horror,TV	TV Dramas	TV Horror	TV Mysteries
s14	14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	22/09/21	2021	TV-PG	91	Children & Family Movie	Children & Family Movies	Comedies	
s8	8	Movie	Sankofa	Haile Gerima	United States	24/09/21	1993	TV-MA	125	Dramas,Independent M	Dramas	Independent Movies	International Movies
s9	9	TV Show	The Great British Baking Show	Andy Devonshire	United Kingdom	24/09/21	2021	TV-14	9 Seasons	British TV Shows,Reality	British TV Shows	Reality TV	
s10	10	Movie	The Starling	Theodore Melfi	United States	24/09/21	2021	PG-13	104	Comedies,Dramas	Comedies	Dramas	
s939	939	Movie	Motu Patlu in the Game of Zones	Suhas Kadav	India	01/05/21	2019	TV-Y7	87	Children & Family Movie	Children & Family Movies	Comedies	Music & Musicals
s13	13	Movie	Je Suis Karl	Christian Schwochow	Germany	23/09/21	2021	TV-MA	127	Dramas,International M	Dramas	International Movies	
s940	940	Movie	Motu Patlu in Wonderland	Suhas Kadav	India	01/05/21	2013	TV-Y7	76	Children & Family Movie	Children & Family Movies	Music & Musicals	
s941	941	Movie	Motu Patlu: Deep Se										
s942	942	Movie	Motu Patlu: Mission										
s852	852	Movie	99 Songs (Tamil)										Music & Musicals
s471	471	Movie	Bridgerton - The Afte										
s730	730	Movie	Bling Empire - The Af										
s731	731	Movie	Cobra Kai - The After										
s913	913	Movie	The Circle - The After										
s4	4	TV Show	Jailbirds New Orlean										
s15	15	TV Show	Crime Stories: India										Docuseries
s3232	3232	Movie	True: Winter Wishes										
s4832	4832	TV Show	True: Magical Friend										
s4833	4833	TV Show	True: Wonderful Wish										
s4857	4857	TV Show	Dance & Sing with True	Mark Thornton, Todd Kauffman	United States	18/05/18	2018	TV-Y	1 Season	Kids' TV	Kids' TV		
s7	7	Movie	My Little Pony: A New Generation	Robert Cullen, Jos√É¬© Luis Ucha	Not Available	24/09/21	2021	PG	91	Children & Family Movie	Children & Family Movies		
s12	12	TV Show	Bangkok Breaking	Kongkiat Komesiri	Not Available	23/09/21	2021	TV-MA	1 Season	Crime TV Shows,Interna	Crime TV Shows	International TV Shows	TV Action & Adventure
s17	17	Movie	Europe's Most Dange										
s7930	7930	Movie	Samudri Lootere										
s21	21	TV Show	Monsters Inside: The										International TV Shows
s24	24	Movie	Go! Go! Cory Carson										
s25	25	Movie	Jeans										Romantic Movies
s28	28	Movie	Grown Ups										
s29	29	Movie	Dark Skies										
s30	30	Movie	Paranoia										
s20	20	TV Show	Jaguar									hows	TV Action & Adventure
s32	32	TV Show	Chicago Party Aunt										
s34	34	TV Show	Squid Game										TV Thrillers
s35	35	TV Show	Tayo and Little Wizards	Not Available	Pakistan	17/09/21	2020	TV-Y7	1 Season	Kids' TV	Kids' TV		
s75	75	TV Show	The World's Most Amazing Vacation Rentals	Not Available	Pakistan	14/09/21	2021	TV-PG	2 Seasons	Reality TV	Reality TV		
s84	84	TV Show	Metal Shop Masters	Not Available	Pakistan	10/09/21	2021	TV-MA	1 Season	Reality TV	Reality TV		
s86	86	TV Show	Pok√É¬©mon Master Journeys: The Series	Not Available	Pakistan	10/09/21	2021	TV-Y7	1 Season	Anime Series,Kids' TV	Anime Series	Kids' TV	

Smoothing

Attribute Construction

Generalization

Aggregation

Normalization

Discretization

CONNECT WITH US

+91 93217 48851

mm@skillcred.co

Please connect with us for detailed references and learner feedback.

SkillCred