Introduction to Big Data

BIG DATA AND HOLLYWOOD: A LOVE STORY

CBC | MENU

Big Brother meets Big Data, in an office near you

The Atlantic

Sponsor Content: What's this?

THE WALL STREET JOURNAL.

CIO JOURNAL.

Carnival Strategy Chief Bets That Big Data Will Optimize Prices

Forbes / Tech

MAY 27, 2015 @ 10:20 AM 34,550

How Big Data And The Internet Of Things Improve Public Transport In London

The Little Black Book of Billior

npr SCIENCE

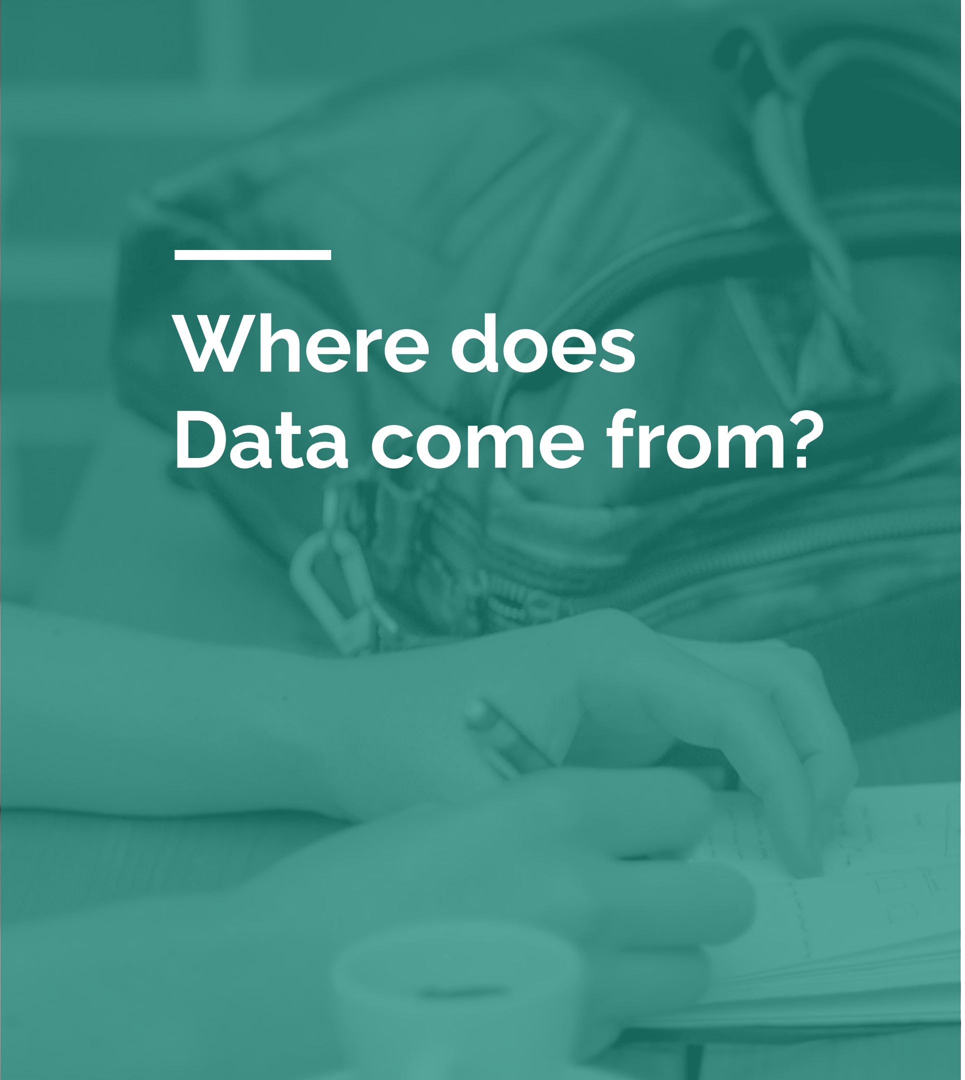
The Big Idea Behind Big Data

New York Times Adapts Data Science Tools for Advertisers

Team will help lure marketers with tools to predict which articles will resonate with certain read

Data Veracity is Critical for Insurers to Make Better Business Decisions, According to Accenture Report

Français



Where does Data come from?

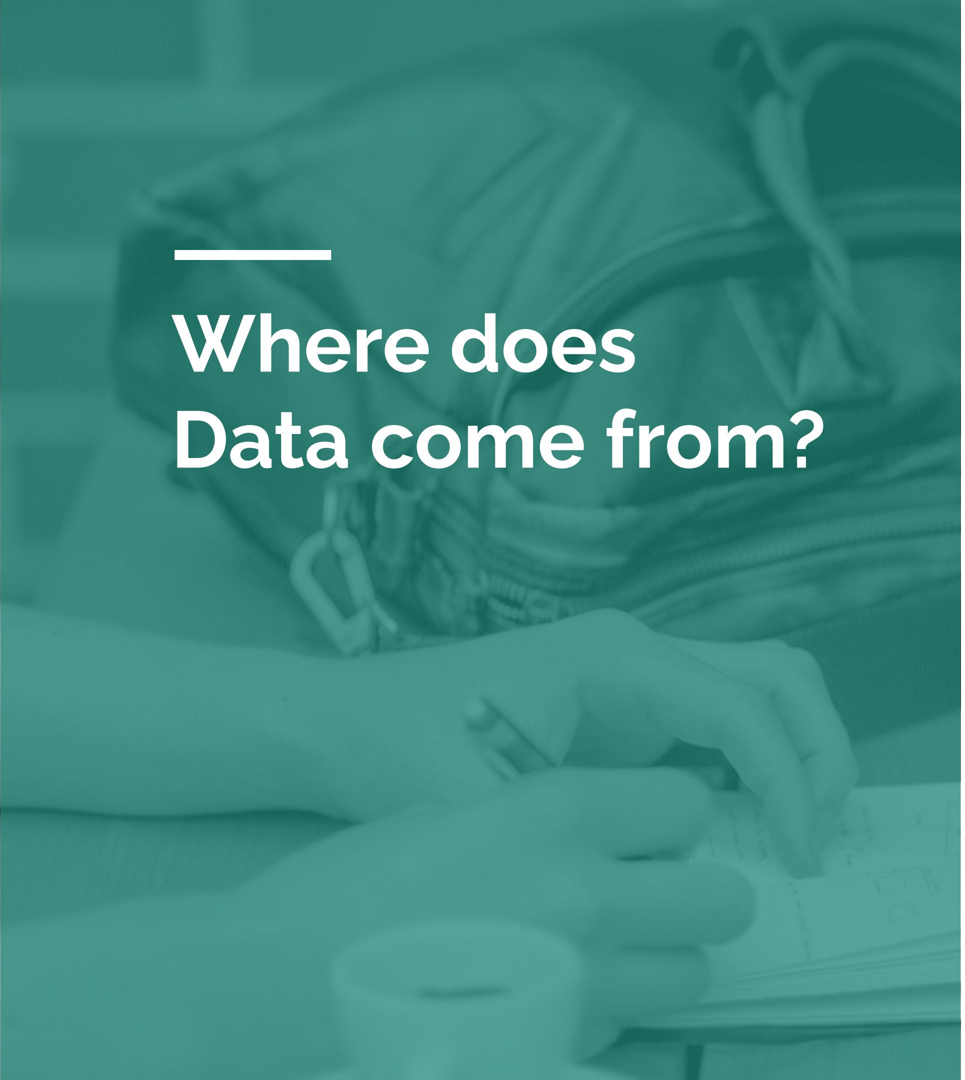
Not-So-Traditionally?

Digital Transactions: Online purchases, website visits, social media interactions, financial transactions.

Sensor Data: Internet of Things (IoT) devices, GPS tracking, wearable technology.

Scientific Monitoring: Satellite imagery, climate data.

Text and Media: Online news articles, blog posts, social media streams, videos, images.



Where does Data come from?

Traditionally?

Government Records: Census data, tax records, vital records (births, deaths, marriages), trade statistics, legal documents.

Historical Documents: Newspapers, personal diaries, business records, scientific reports.

Cultural Artifacts: Literature, artwork, archaeological findings.

Data is not new!

CBC | MENU ⌄

Big Brother meets Big Data, in an office near you

Forbes / Tech

MAY 27, 2015 @ 10:20 AM 34,550

How Big Data And Th
Transport In Lo

The Atlantic

Sponsor Content:
What's this?

EET JOURNAL.

The Little Black Book of Billior

mprove Public

DATA AND

The Big Idea Behind Big Data

HOLLYWOOD: A LOVE STORY

THE

CIO JOURNAL.

Carnival Strategy
ptimize Prices

New York Times Adapts Data Science Tools for Advertisers

Team will help lure marketers with tools to predict which articles will resonate with certain read

Data Veracity is Critical for Insurers to Make Better Business Decisions, According to Accenture Report

Français

Data is not new!

CBC | MENU

Big Brother meets Big Data, in an office near you

The Atlantic Sponsor Content: What's this?

Forbes / Tech

MAY 27, 2015 @ 10:20 AM 34,550

How Big Data And Th Transport In Lo

The Little Black Book of Billion

mprove Public

THE WALL STREET JOURNAL.

AND HOLLYWOOD: A LOVE STORY

The Big Idea Behind Big Data

THE

CIO JOURNAL.

Carnival Strategy timize Prices

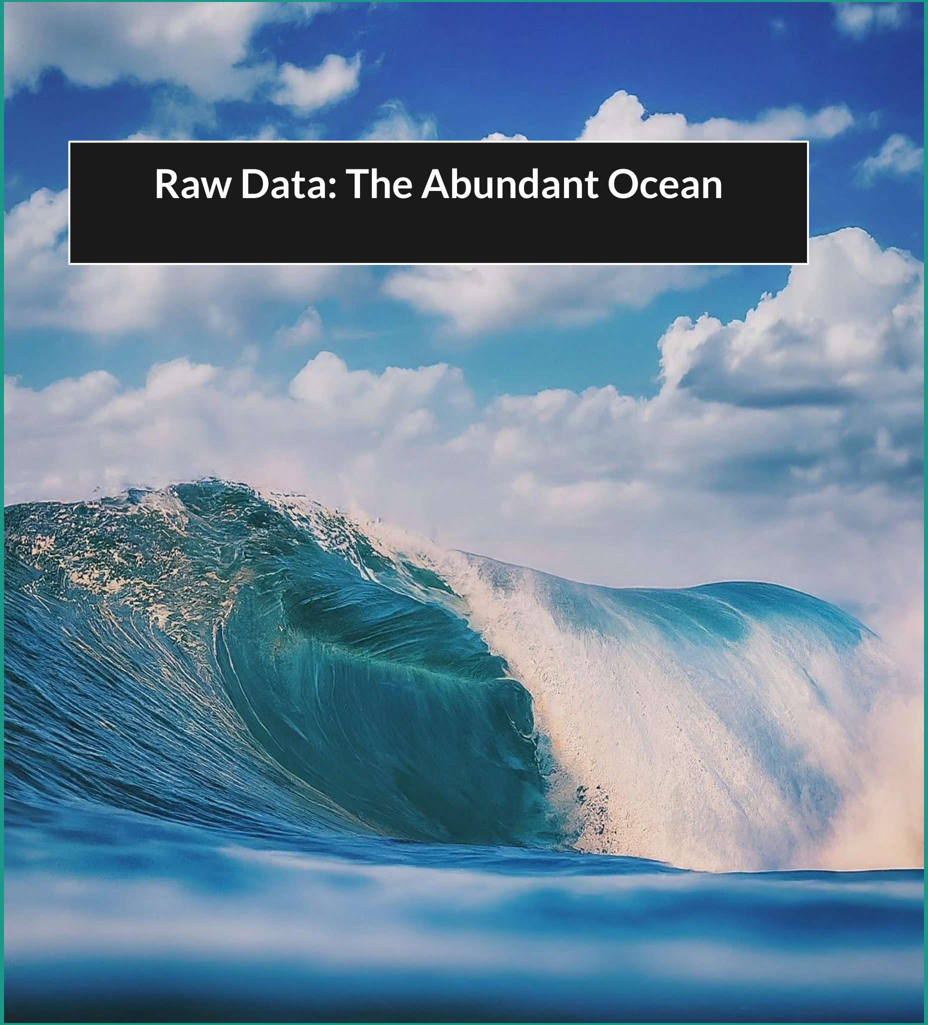
New York Times Adapts Data Science Tools for Advertisers

Team will help lure marketers with tools to predict which articles will resonate with certain read better target advertising

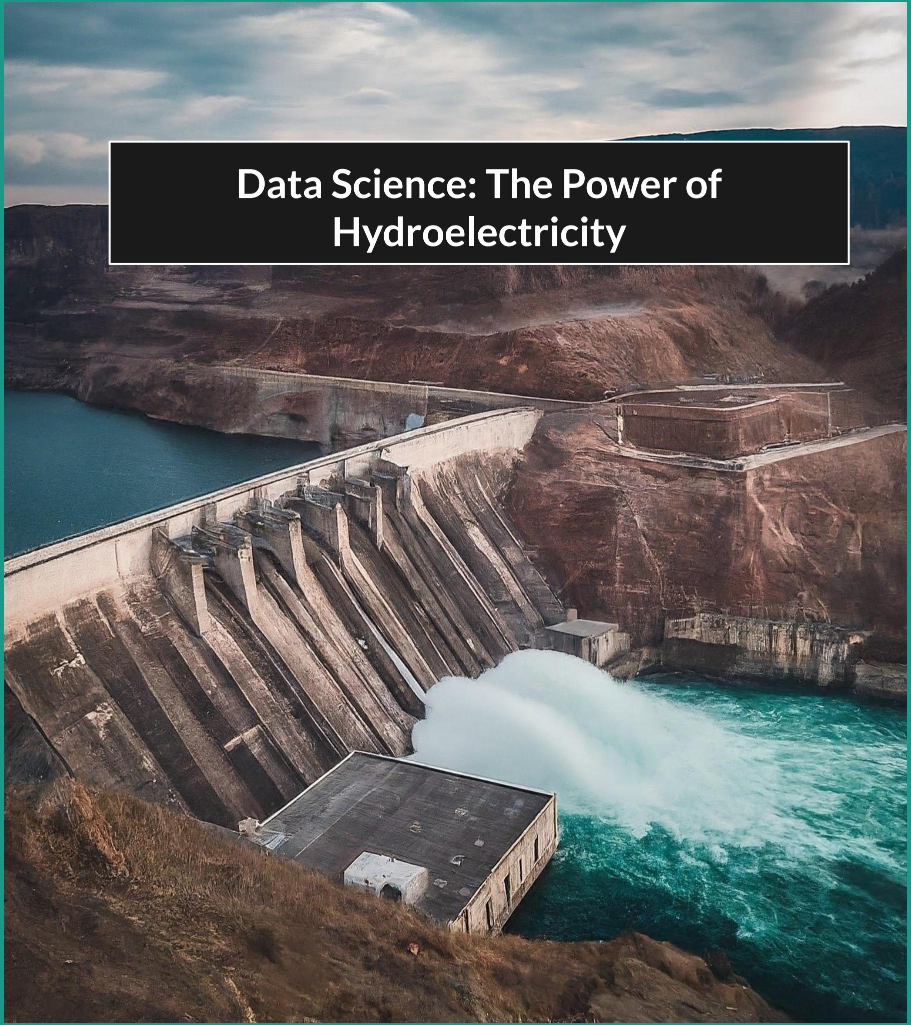
Data Veracity is Critical for Insurers to Make Better Business Decisions, According to Accenture Report

Français

Raw Data: The Abundant Ocean



Data Science: The Power of Hydroelectricity



Introduction to Data Science

What is Data Science?

Data science is like the science and engineering behind hydroelectricity.

Capturing and Channeling: Just like dams and reservoirs manage water flow.

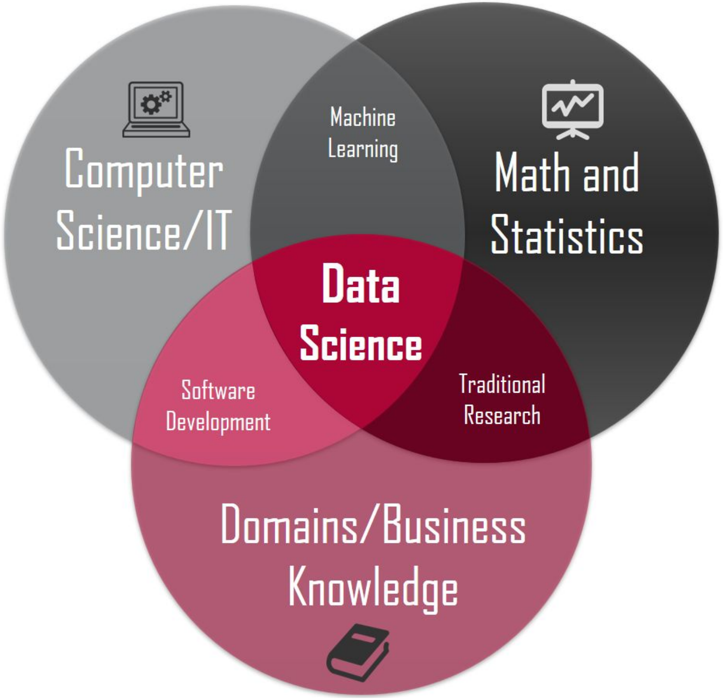
Filtering and Purification: Much like water treatment plants ensure water quality.

Generating Insights (Energy): Reveal hidden patterns – equivalent of turbines converting water flow into usable energy.

Problem-Solving Applications: Inform decision-making, just like hydroelectricity provides light, powers industries.

What is Data Science?

Formally -



Computer Science/IT

Machine Learning

Math and Statistics

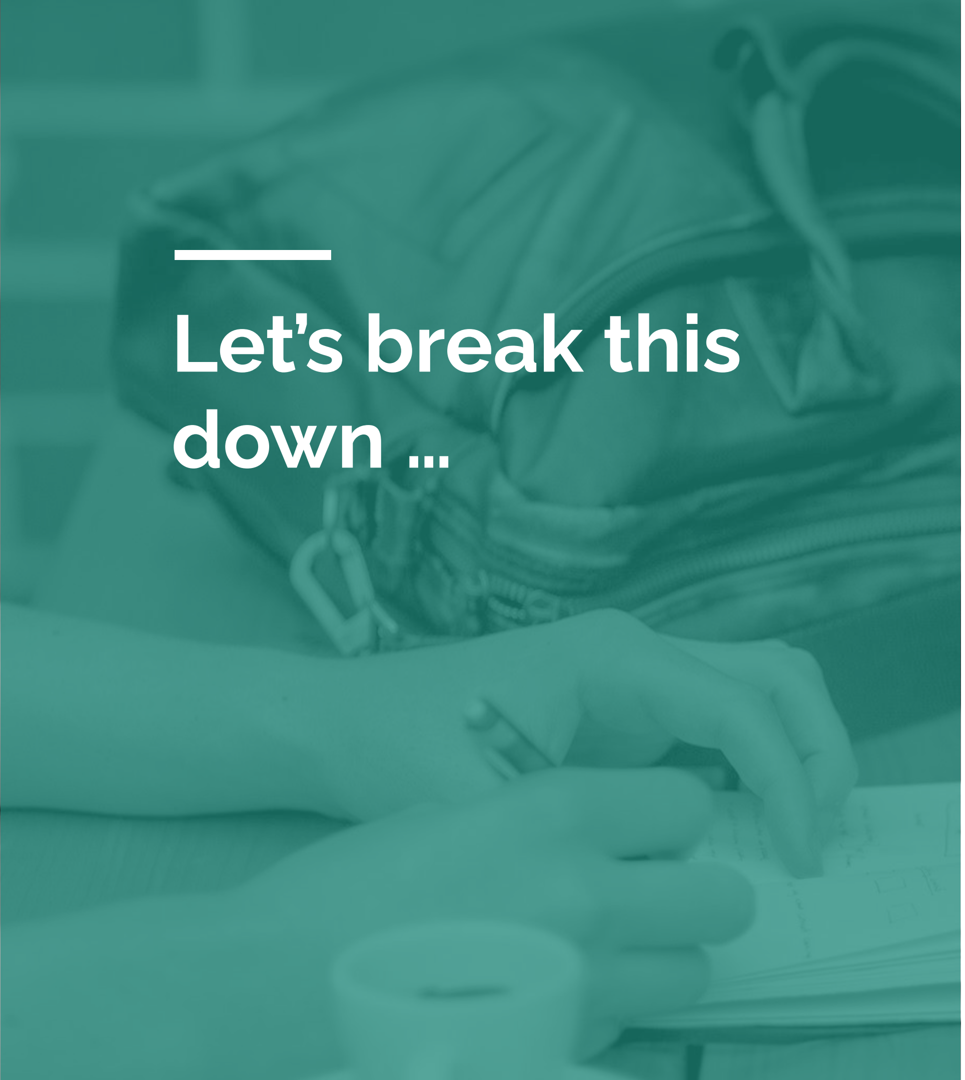
Data Science

Software Development

Traditional Research

Domains/Business Knowledge

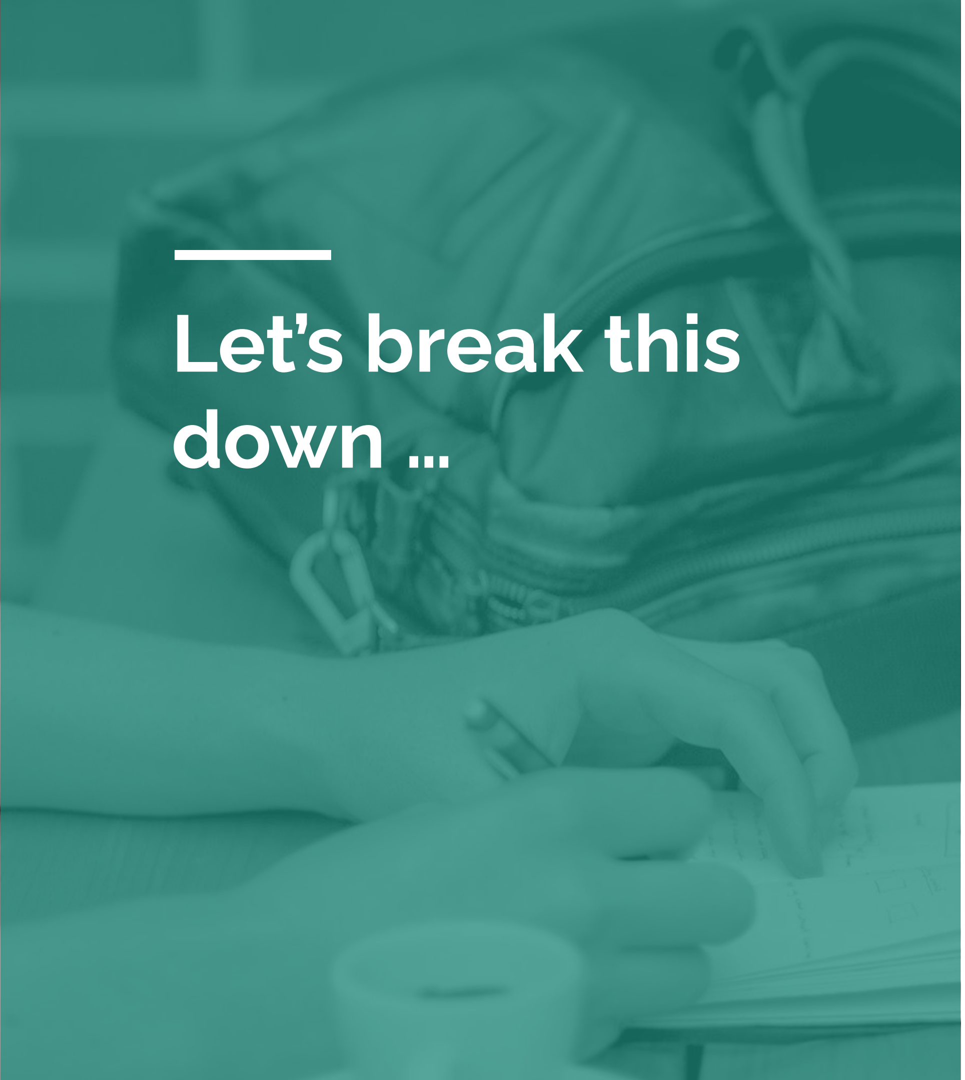
Pic Credits:
https://medium.com/@anuraggandhi29/what-is-datascience-6ac639f830c2



Let’s break this down ...

Computer Science

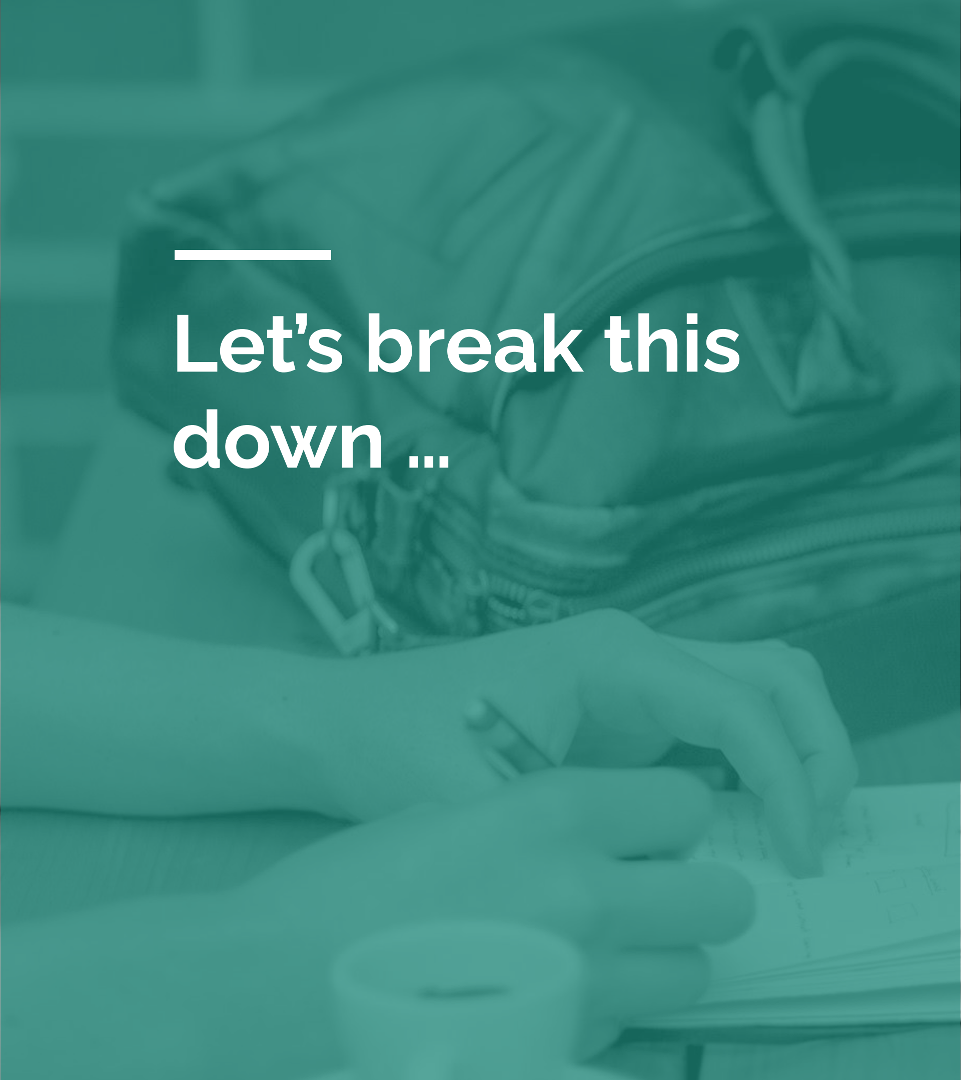
- Algorithms and Data Structures
- Programming
- Databases
- Machine Learning



Let’s break this down …

Mathematics & Statistics

- Statistics
- Probability
- Linear Algebra
- Calculus



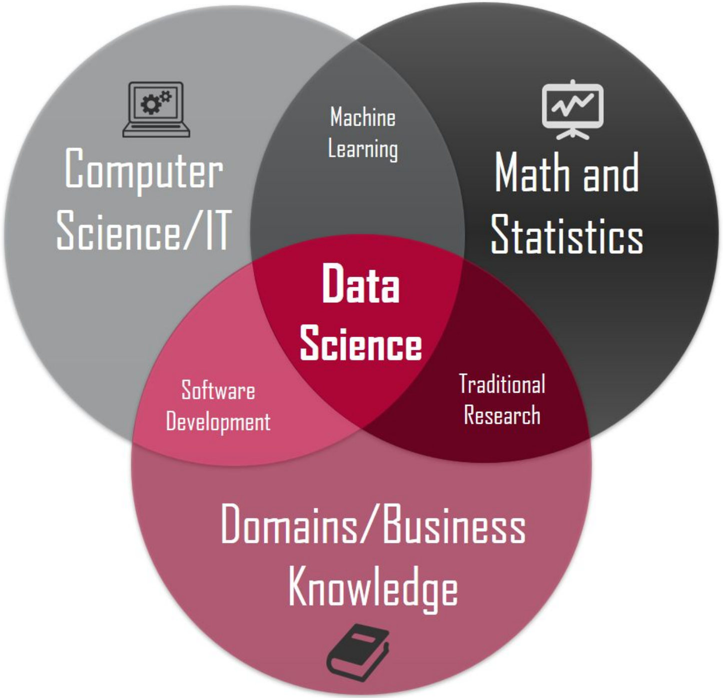
Let’s break this down …

Domain Knowledge

- Understanding the Problem
- Feature Engineering
- Data Storytelling

What is Data Science?

Formally -



Pic Credits:
https://medium.com/@anuraggandhi29/what-is-datascience-6ac639f830c2

The 5 P’s of Data Science

What does it take to go from raw data to useful products?

What does it take to go from raw data to useful products?

What does it take for a chef to bake a cake?

Ingredients!

What does it take to go from raw data to useful products?

What does it take for a chef to bake a cake?

The 5 P's

Purpose - Aims to bake a specific cake

People - The chef himself!

Process - A recipe.

Platforms - Kitchenware, tools (oven, mixer, whisk) etc.

Programmability - Skill - chef's mastery of techniques.

The 5 P’s

Purpose - A clearly defined problem.

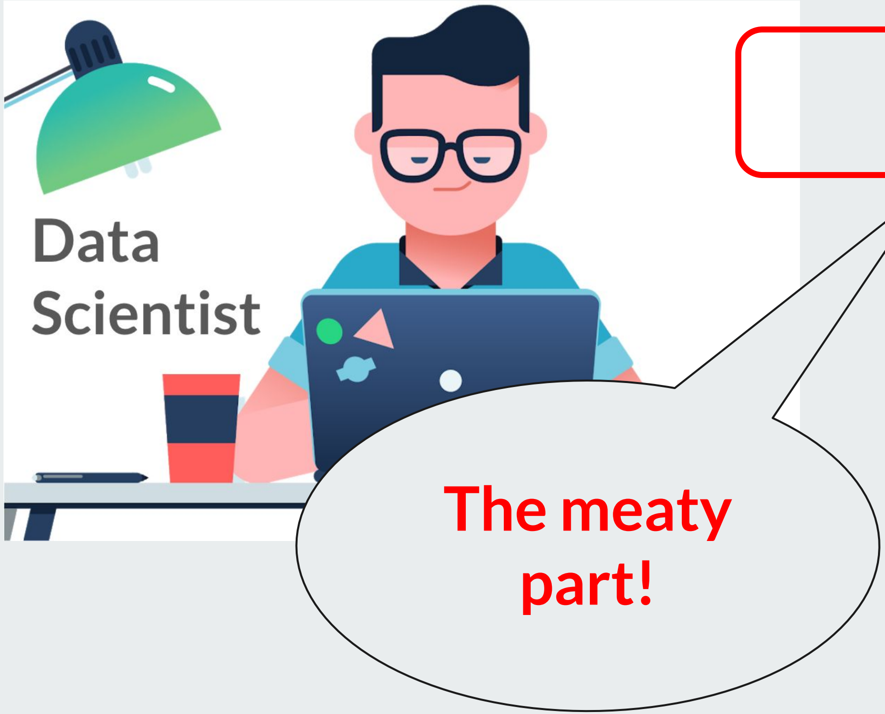
People - A qualified team.

Process - A well-defined workflow.

Platforms - The right platforms & tools to transform data.

Programmability - The ability to code & automate.

The 5 P's



The meaty part!

Purpose - A clearly defined problem.

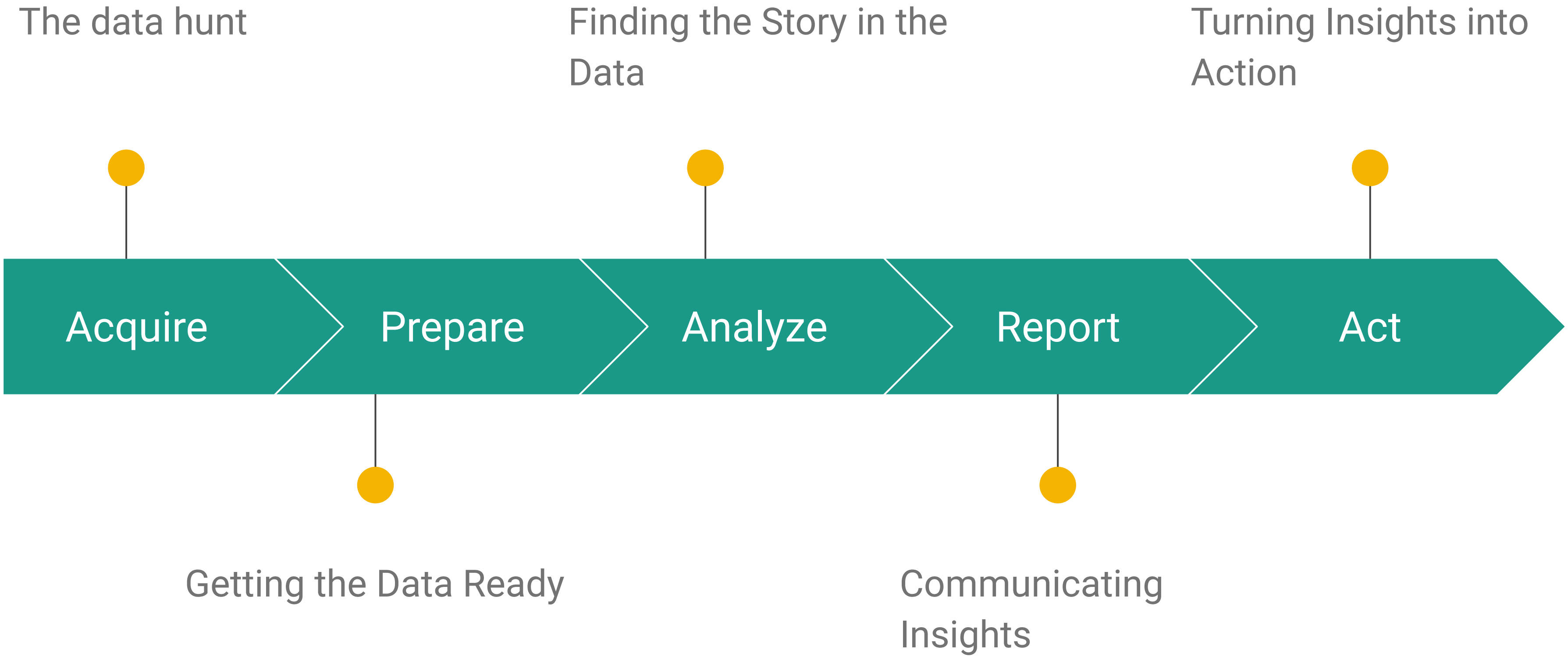
People - A qualified team.

Process - A well-defined workflow.

Platforms - The right platforms & tools to transform data.

Programmability - The ability to code & automate.

Steps in the Data Science Process



The data hunt

Finding the Story in the Data

Turning Insights into Action

Acquire

Prepare

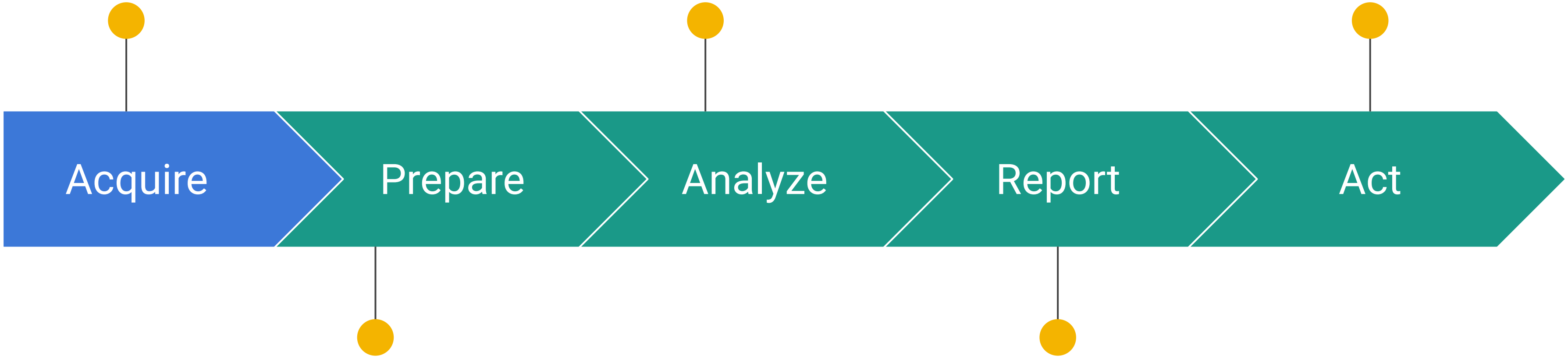
Analyze

Report

Act

Getting the Data Ready

Communicating Insights



Acquire

The data hunt

Prepare

Getting the Data Ready

Analyze

Finding the Story in the Data

Report

Communicating Insights

Act

Turning Insights into Action

Acquire

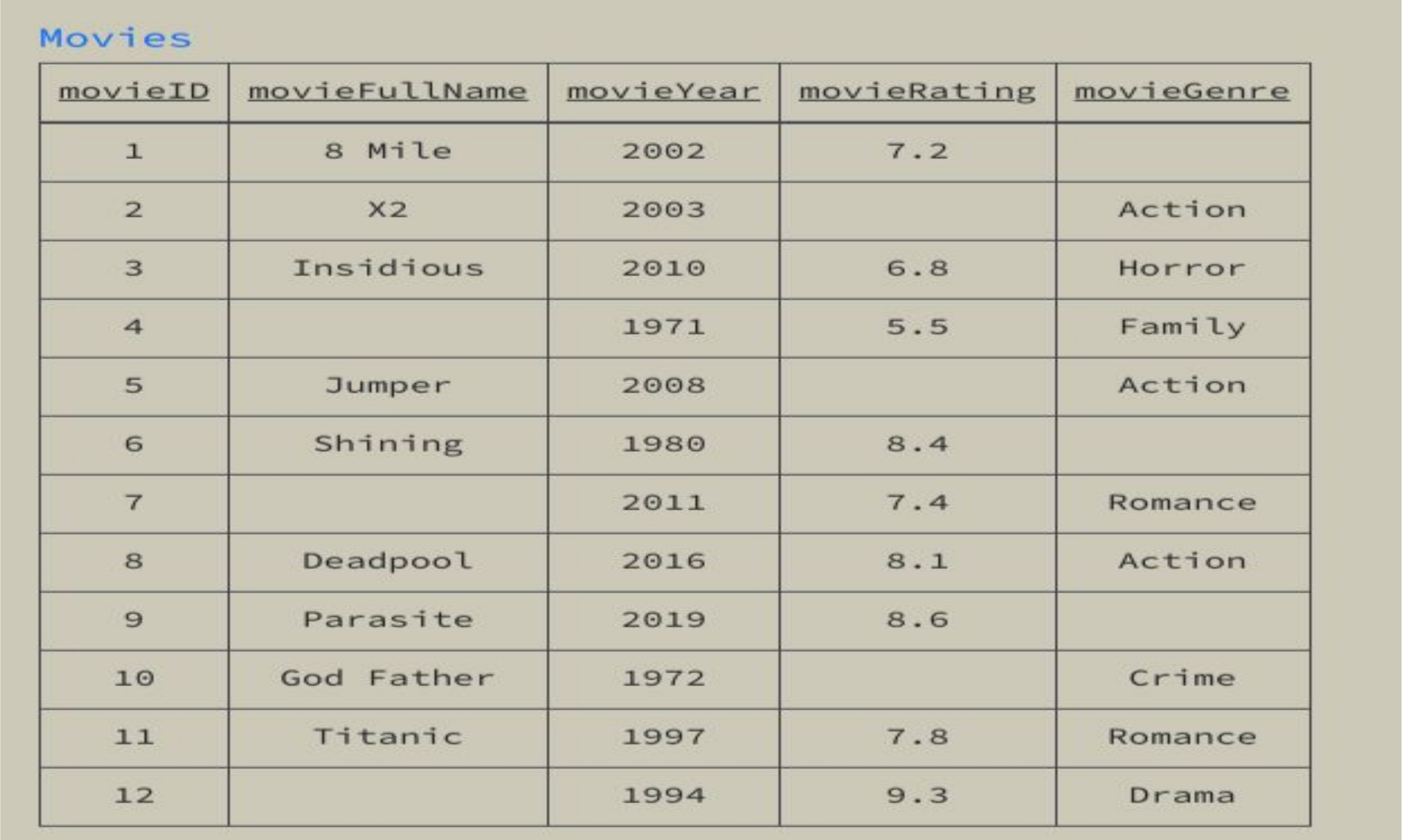
The Data Hunt

- Process of obtaining the data needed
- Variety of sources:
  - **Relational Databases**
  - **NoSql databases**
  - **Text files**
  - **Websites**

Acquire
The Data Hunt

- Process of obtaining the data needed
- Variety of sources:
  - ○ **Relational Databases -** Use SQL
  - ○ **NoSql databases -** Use API & Web Services
  - ○ **Text files -** Scripting languages (js, python, perl, php)
  - ○ **Websites -** Web Services for remote data

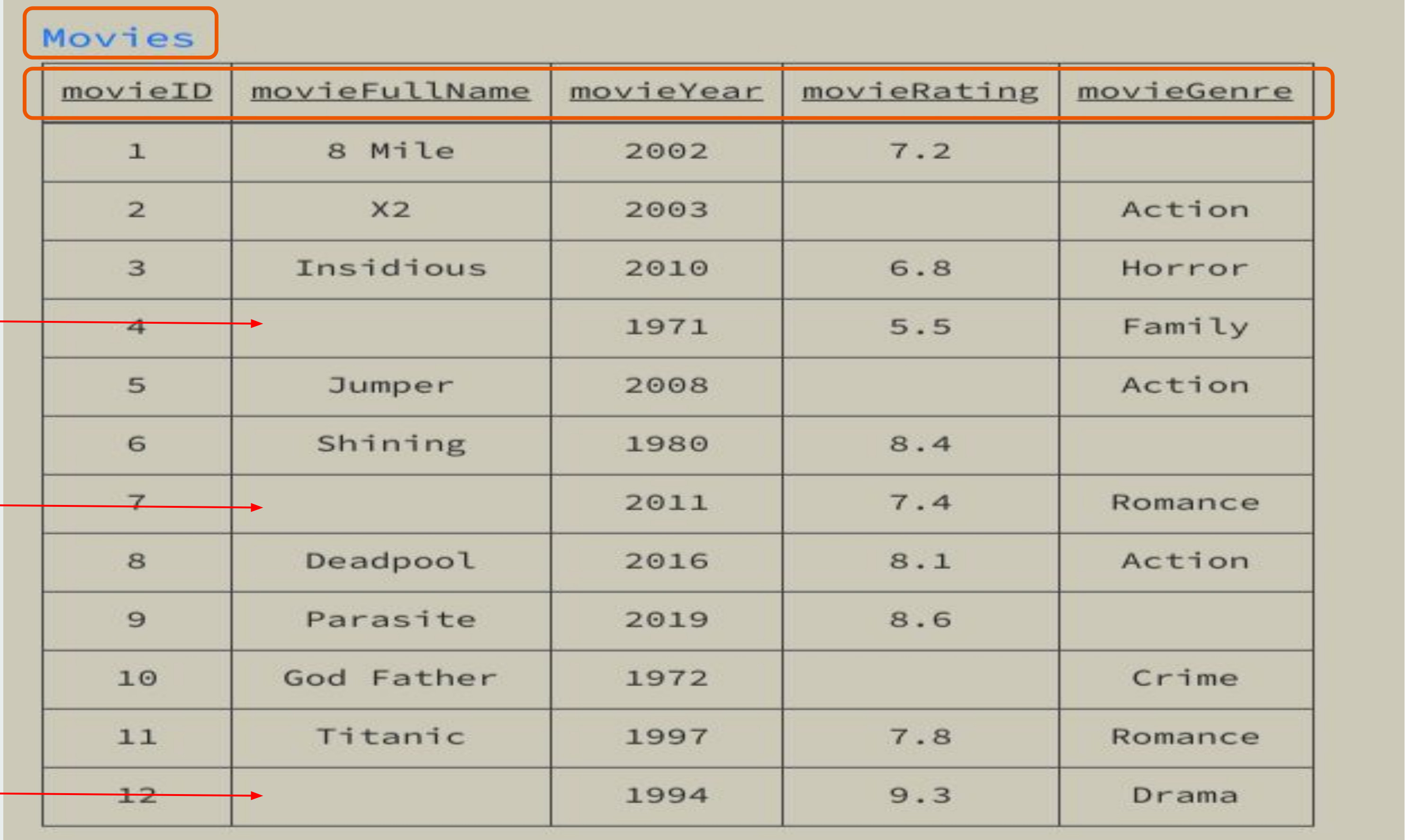
Acquire - The Data Hunt



Movies

movieID	movieFullName	movieYear	movieRating	movieGenre
1	8 Mile	2002	7.2	
2	X2	2003		Action
3	Insidious	2010	6.8	Horror
4		1971	5.5	Family
5	Jumper	2008		Action
6	Shining	1980	8.4	
7		2011	7.4	Romance
8	Deadpool	2016	8.1	Action
9	Parasite	2019	8.6	
10	God Father	1972		Crime
11	Titanic	1997	7.8	Romance
12		1994	9.3	Drama

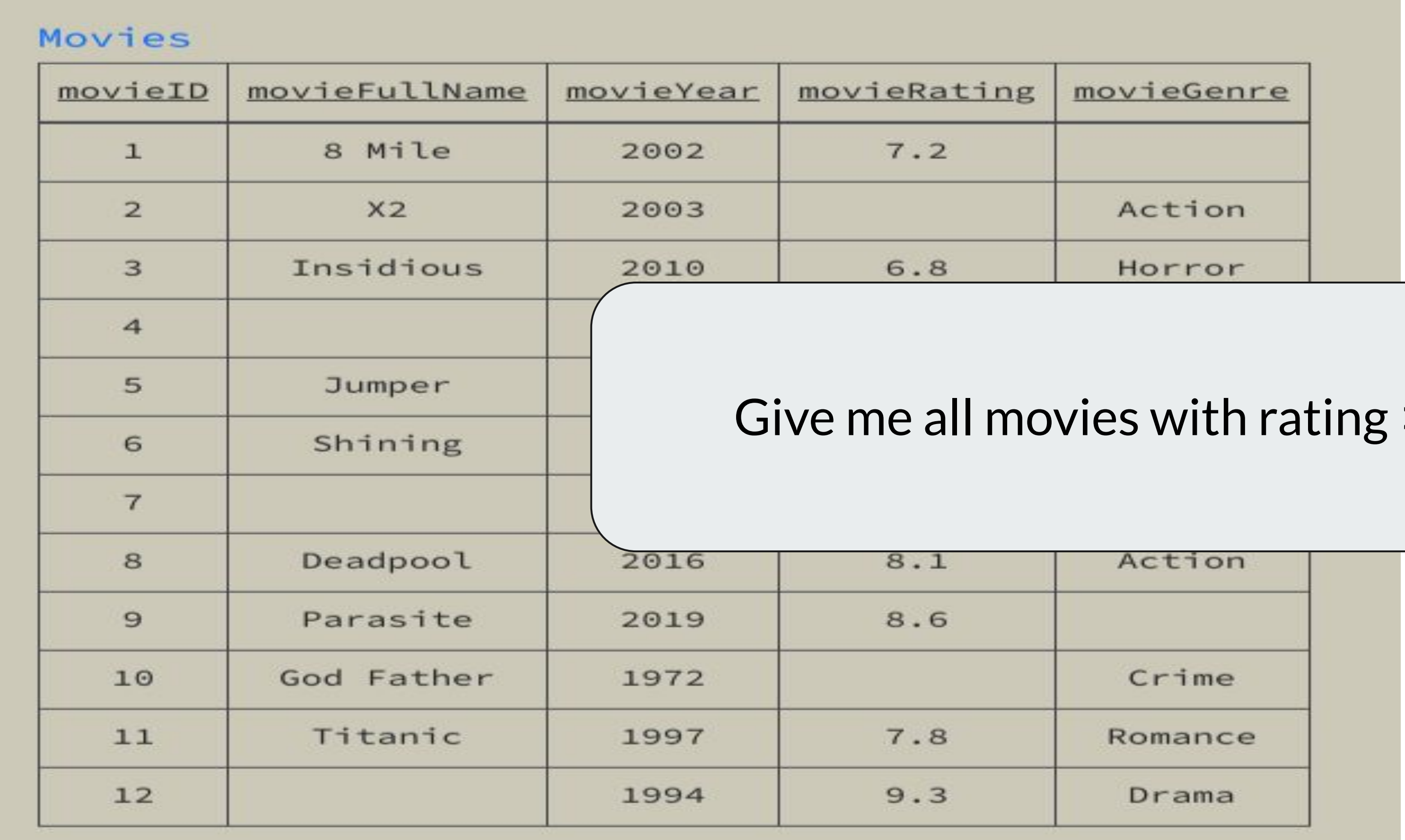
Acquire - The Data Hunt



Movies

movieID	movieFullName	movieYear	movieRating	movieGenre
1	8 Mile	2002	7.2	
2	X2	2003		Action
3	Insidious	2010	6.8	Horror
4		1971	5.5	Family
5	Jumper	2008		Action
6	Shining	1980	8.4	
7		2011	7.4	Romance
8	Deadpool	2016	8.1	Action
9	Parasite	2019	8.6	
10	God Father	1972		Crime
11	Titanic	1997	7.8	Romance
12		1994	9.3	Drama

Acquire - The Data Hunt

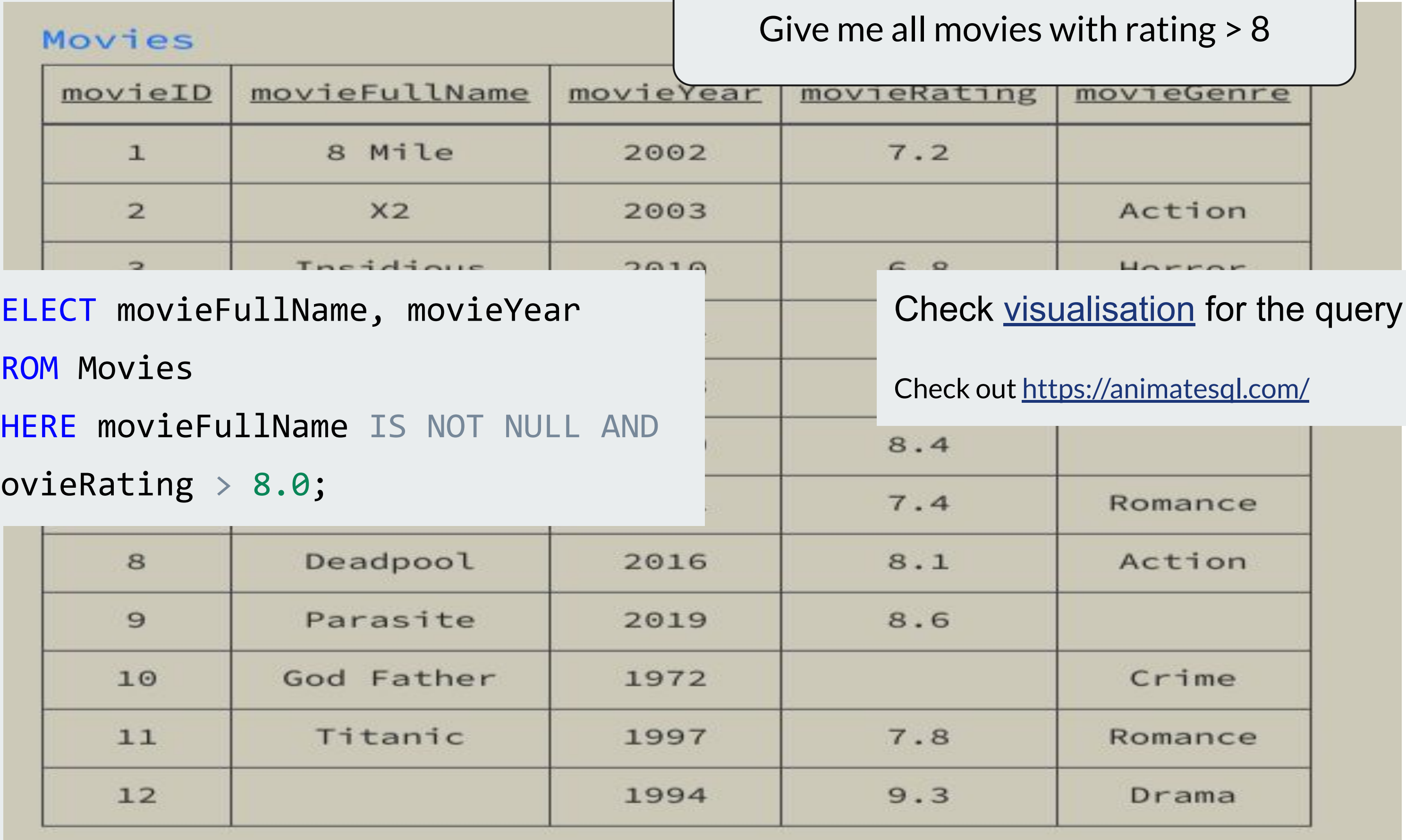


Movies

movieID	movieFullName	movieYear	movieRating	movieGenre
1	8 Mile	2002	7.2	
2	X2	2003		Action
3	Insidious	2010	6.8	Horror
4		[illegible]	[illegible]	[illegible]
5	Jumper	[illegible]	[illegible]	[illegible]
6	Shining	[illegible]	[illegible]	[illegible]
7		[illegible]	[illegible]	[illegible]
8	Deadpool	2016	8.1	Action
9	Parasite	2019	8.6	
10	God Father	1972		Crime
11	Titanic	1997	7.8	Romance
12		1994	9.3	Drama

Give me all movies with rating > 8

Acquire - The Data Hunt



Movies

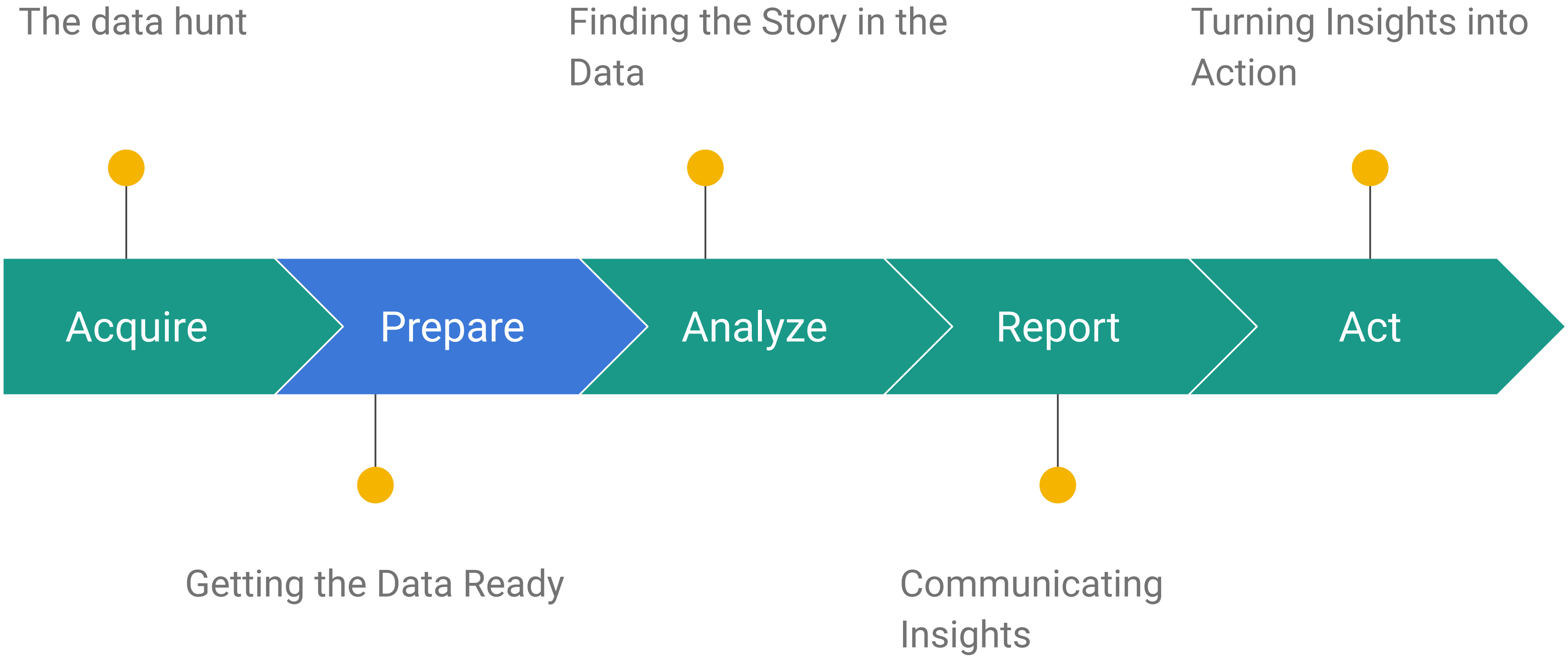
movieID	movieFullName	movieYear	movieRating	movieGenre
1	8 Mile	2002	7.2	
2	X2	2003		Action
3	Insidious	2010	6.8	Horror
			8.4	
			7.4	Romance
8	Deadpool	2016	8.1	Action
9	Parasite	2019	8.6	
10	God Father	1972		Crime
11	Titanic	1997	7.8	Romance
12		1994	9.3	Drama

Give me all movies with rating > 8

```sql
SELECT movieFullName, movieYear
FROM Movies
WHERE movieFullName IS NOT NULL AND
movieRating > 8.0;
```

Check visualisation for the query!

Check out https://animatesql.com/



The data hunt

Finding the Story in the Data

Turning Insights into Action

Acquire

Prepare

Analyze

Report

Act

Getting the Data Ready

Communicating Insights

Prepare
Getting the Data Ready

Issues in raw data

Order ID	Customer Name	Order Date	Price ($)	Country
12345	John Smith	2023-12-15	55.99	USA
98765	jane doe	15/12/2023	12.5	UK
12345	J. Smith	12/15/2023	55.99	US
45678	Sarah Johnson	2023-13-05	-20	Canada
33322	William Lee	null	89.99	Australia

Issues in raw data

Order ID	Customer Name	Order Date	Price ($)	Country
12345	John Smith	2023-12-15	55.99	USA
98765	jane doe	15/12/2023	12.5	UK
12345	J. Smith	12/15/2023	55.99	US
45678	Sarah Johnson	2023-13-05	-20	Canada
33322	William Lee	null	89.99	Australia

Prepare
Getting the Data Ready

Types of Issues:

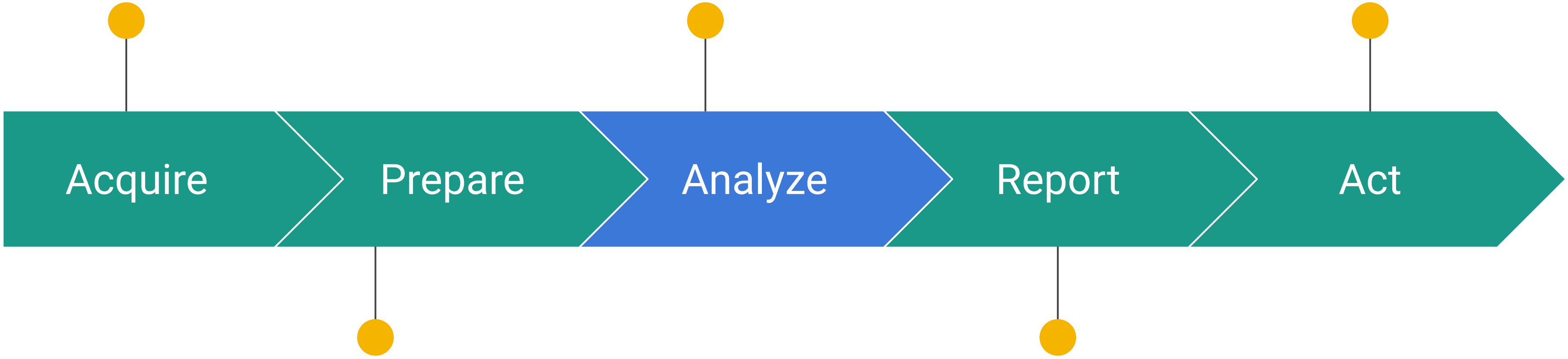
Inconsistent values: Different spellings, date formats.

Duplicate records: Identify and handle them.

Missing values: Deletion, Imputation.

Invalid data: Out-of-range values, errors.

Outliers: Investigate if they're true errors or meaningful extremes.



Acquire — The data hunt

Prepare — Getting the Data Ready

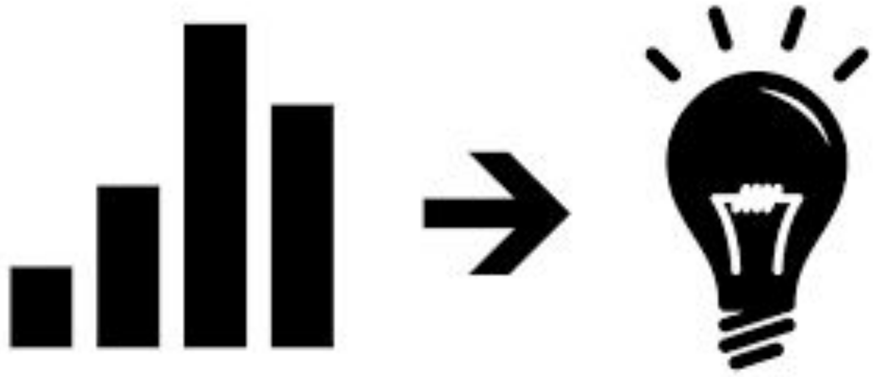
Analyze — Finding the Story in the Data

Report — Communicating Insights

Act — Turning Insights into Action

Analyze

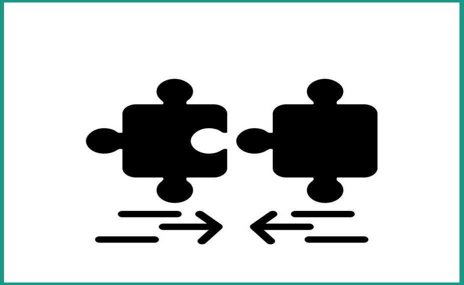
Finding the Story in the Data



Why do you want to analyse?

What is likely to happen in the future?

What are the natural divisions within my data?



What items or events tend to occur together?

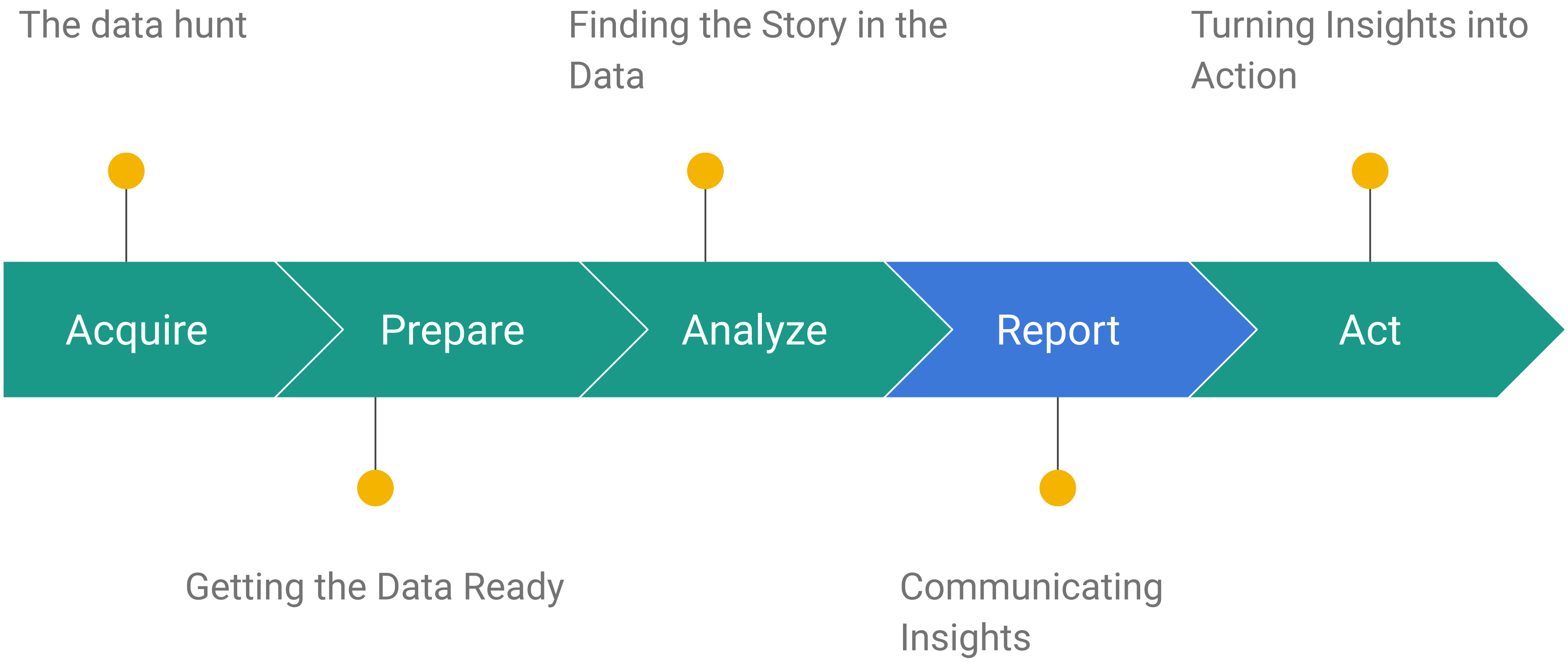
Common Analysis Types

- **Regression**: Predicting future values (e.g., Sales forecasting, stock price prediction).

- **Clustering**: Grouping similar data points together (e.g., Customer segmentation)

- **Association Rule Mining**: Finding patterns of co-occurrence (e.g., "Customers who bought X also bought Y").

- **Classification**: Predicting categories (Email: spam/not spam).



Acquire

The data hunt

Prepare

Getting the Data Ready

Analyze

Finding the Story in the Data

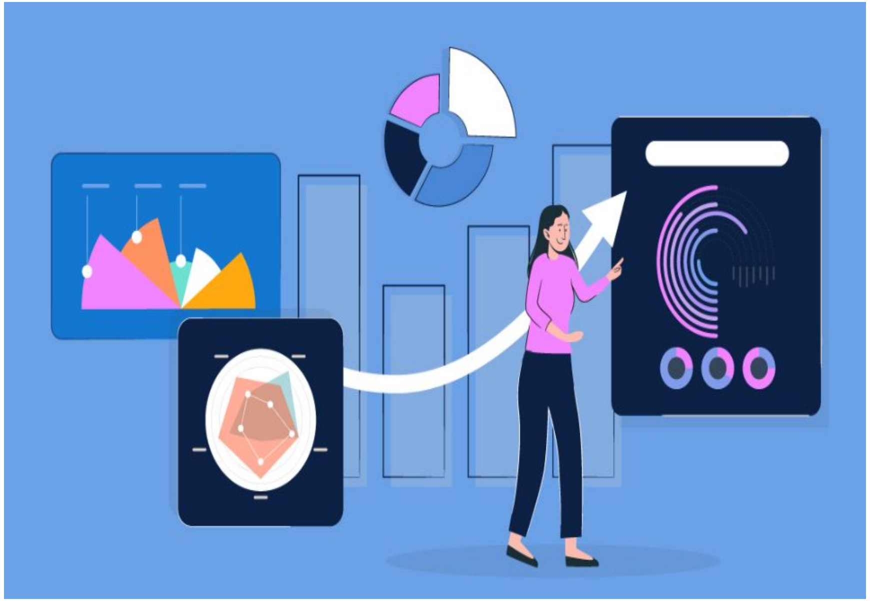
Report

Communicating Insights

Act

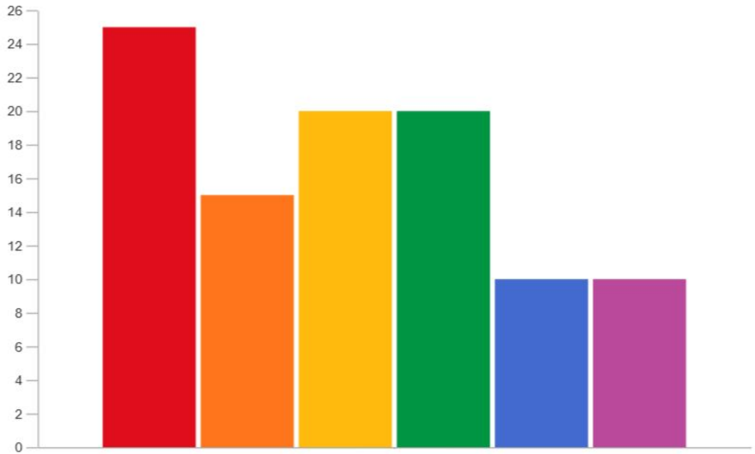
Turning Insights into Action

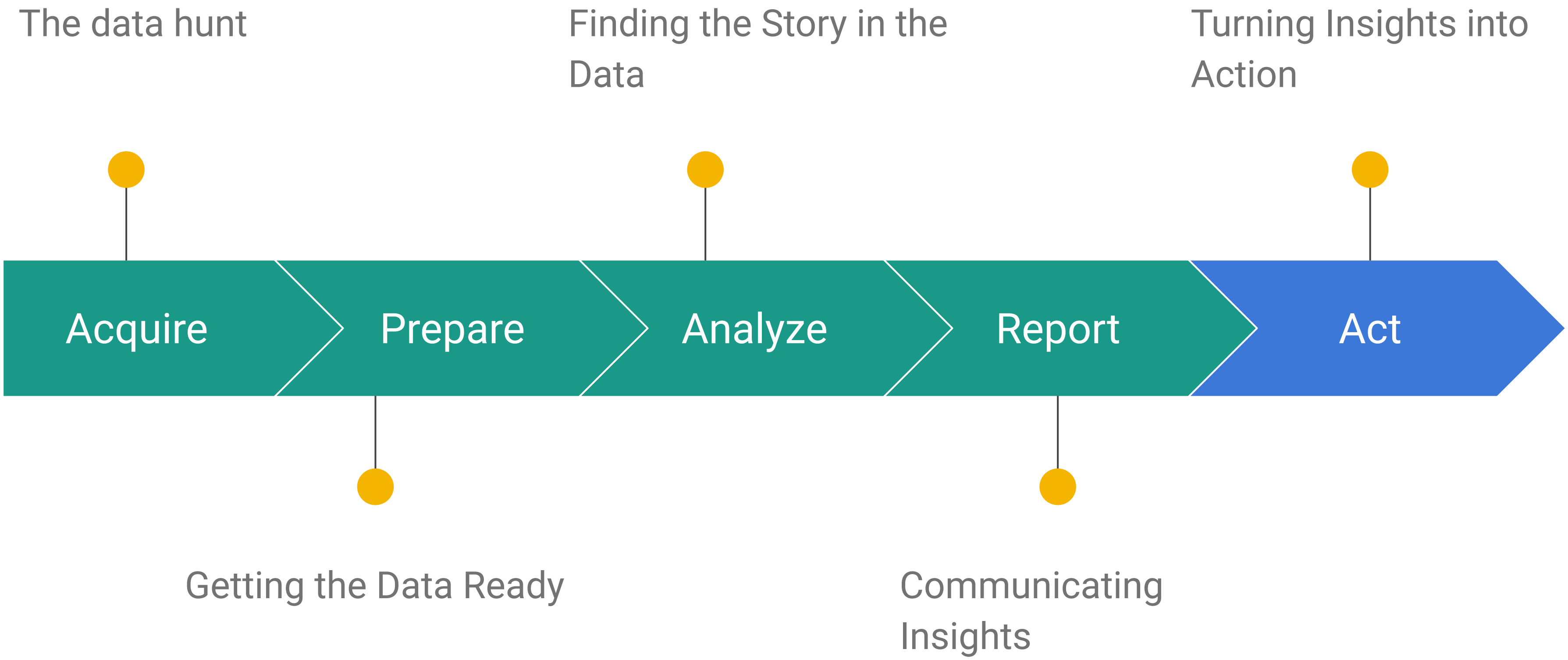
Report
Communicating Insights



Report - Communicating Insights

- Data alone doesn't create change.
- Presentation is important!





Acquire

The data hunt

Prepare

Getting the Data Ready

Analyze

Finding the Story in the Data

Report

Communicating Insights

Act

Turning Insights into Action

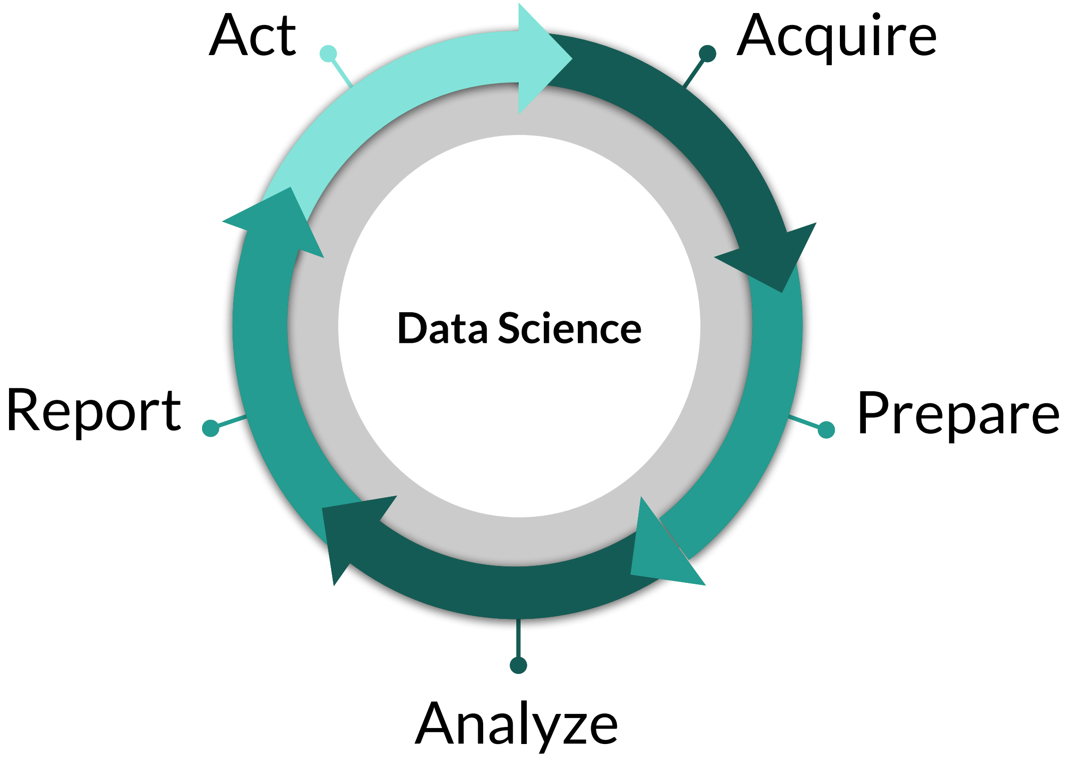
Act

Turning Insights into Action

Act - Turning Insights into Action

- Goal : Use data-driven insights to inform decisions that improve our business, research, or outcomes.
- Decision-making isn't the end - Feed new data back into the process.

Takeaway - Data science process is a continuous cycle!



Act
Acquire
Data Science
Report
Prepare
Analyze

The data science techniques we've discussed work well with datasets that fit on a single machine.

But what happens when the data explodes?

The data scien
discussed work
on a single ma

But what happ
explodes?

Volume

Variety

Speed of information

Big Data!