ASKING QUESTION

The purpose of computing is insight, not numbers.
– Richard W. Hamming

Prepared by Dr. Salimah Mokhtar

Learning Objectives:

1. To recap what data scientist do.
2. To explain data processing pipeline.
3. To discuss asking a question.
4. To categorize data analytics questions.
5. To identify 5 questions Data Science can answer with Machine Learning.

Recap – What Data Scientists Do

- **Ask questions** in order to answer or solve problems or find new solutions to problems with which businesses must deal.
- **Define data** necessary to support or help answer the question.
- **Work with existing data** to find, collect, store and explore data.
- **Determine** the needed **analysis type** for a particular situation or type of data.
- Use **algorithms** and other **tools** to parse, clean, check quality and utilize data.
- **Transform insights** learned into formats usable by non-data scientists (for use by others within the business), including the creation of graphs, charts, infographics and more.
- Create software to **automate data science tasks** based on specific business requirements and data sources.

Companies that struggle to get meaningful insights from their data are often not asking the right questions

GUKODO - STOCK.ADOBE.COM

Domain Area

- Data Science working in a **domain area** collaborates with the business / organization with a series of questions and answers that allow data scientist to deliver the analysis/model/data product that the business want.
- Questions are required to **fully understand** what the business wants and not find data scientist **making assumptions** about what others are thinking.

*Asking the right questions, like those you identified here is what separate Data Scientists that know '**why**' from folks that only know what (tools and technologies).*

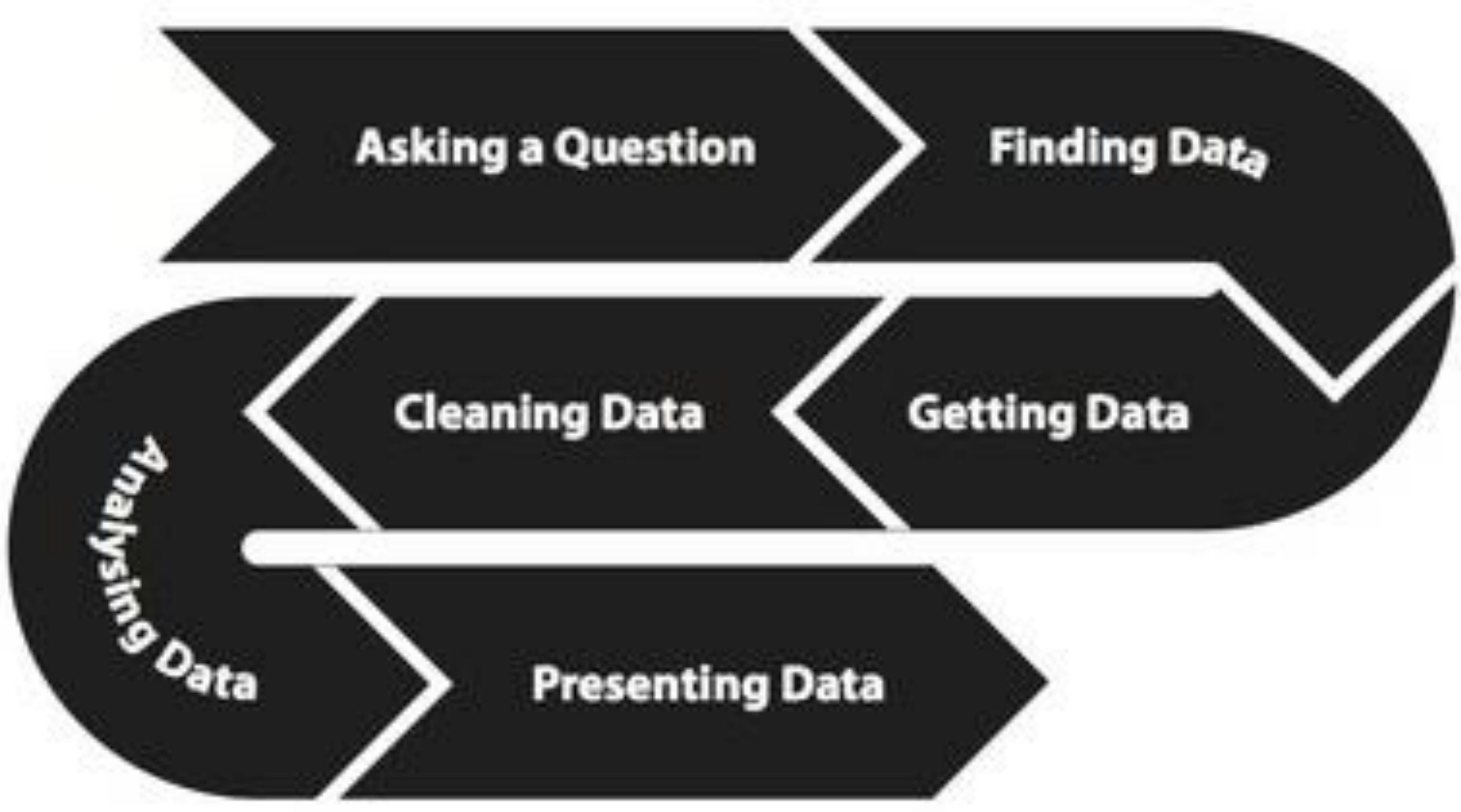
- Kayode Ayankoya (Data Scientist at Microsoft)

"Asking the right questions involves **domain knowledge** and **expertise**, coupled with a keen ability to see the **problem**, see the available **data**, and match up the two."

Data-Driven: Creating a Data Culture

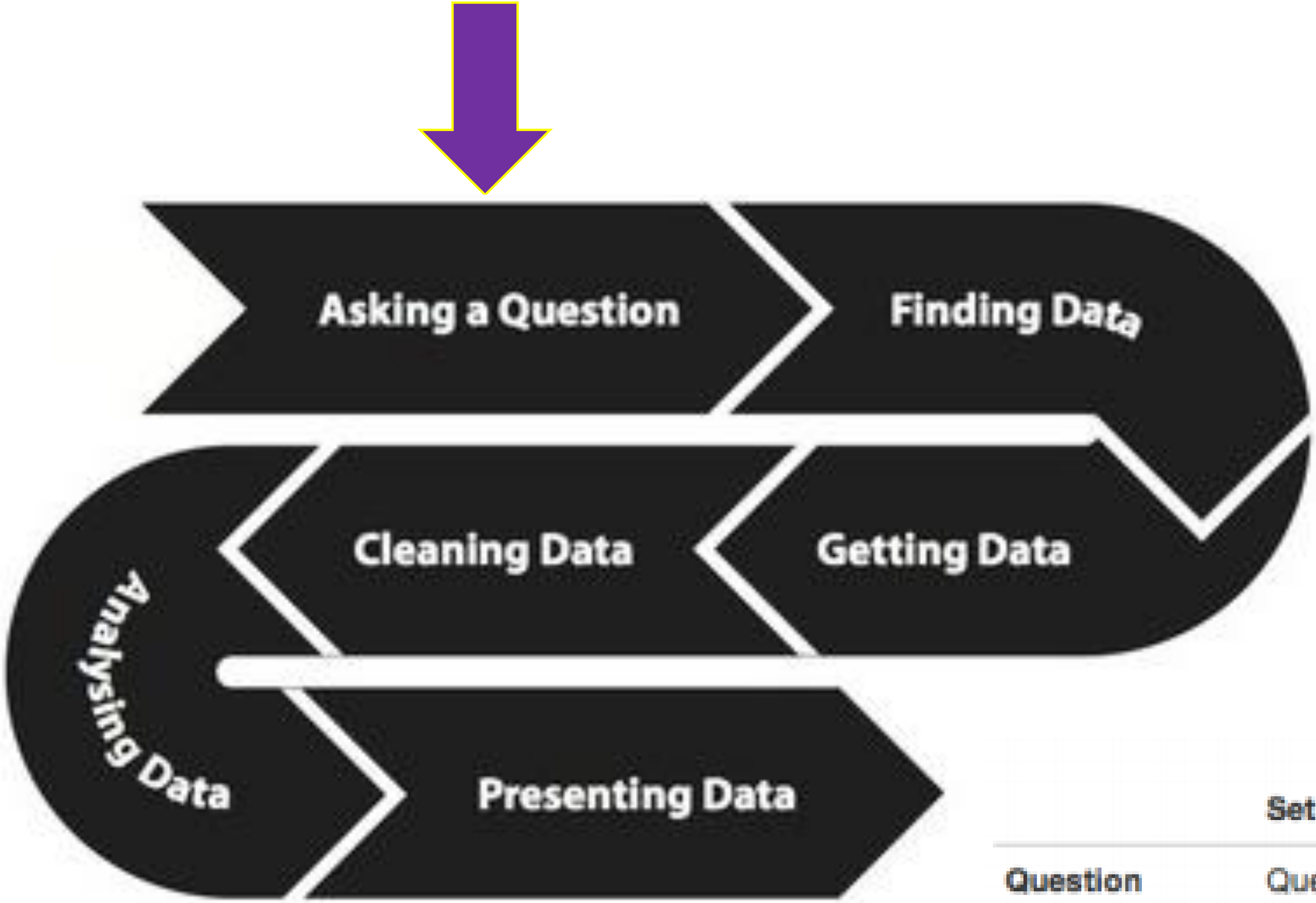
The Data Processing Pipeline

Data pipeline, is a set of data processing elements connected in series, where the output of one element is the input of the next one.



Source: https://community.openspending.org/resources/handbook/

First Stage - Stating the Question



1. Setting Expectations
2. Collect information (data), comparing the data to your expectations, and if the expectations don’t match,
3. Revising your expectations or fixing the data so your data and your expectations match.

	Set Expectations	Collect Information	Revise Expectations
Question	Question is of interest to audience	Literature Search/Experts	Sharpen question

"If I had an hour to solve a problem and my life depended on the solution, I would spend the first 55 minutes determining the proper question to ask, for once I know the proper question, I could solve the problem in less than five minutes" — **Albert Einstein**.

"A problem well-stated is a problem half-solved."

-- Charles Kettering,
head of research at GM

Asking questions is easy!

But asking the **RIGHT** questions is somewhat not straight forward.

The right question can provide **direction** and **purpose**.

Important step:

Defining a **data science problem** (or problem statement) is one of the most important steps in data science pipeline.

https://towardsdatascience.com/how-to-ask-the-right-questions-as-a-data-scientist-913621907411

Ultimately, final goals → to formulate better questions and well-defined problem statements → to solve using data science approach and generate business insights and drive actionable plans.

- ○ don’t know exactly what they want.
- ○ might have a vague idea of what the problem is and what they need as a solution.

Business people

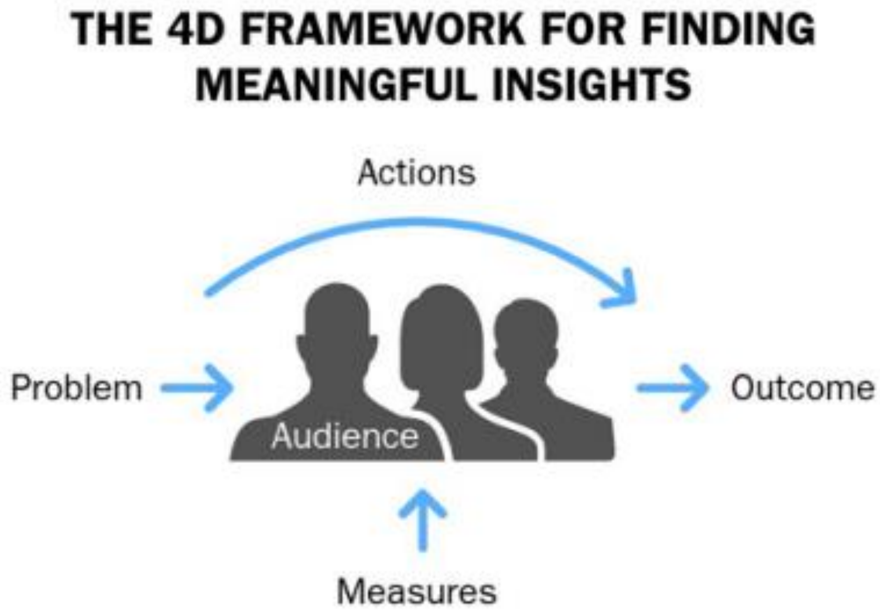
Data scientists

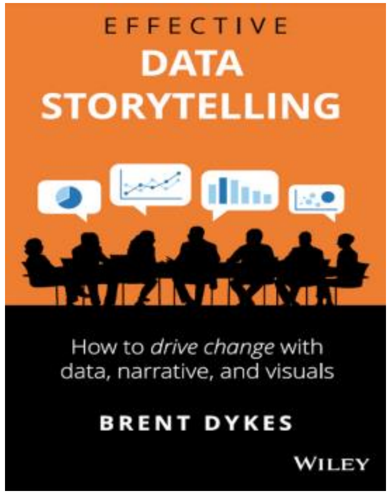
What Questions to Ask?

Need to understand the problem & translate business needs into technical model requirements.

- ✓ What business problem are you trying to solve?
- ✓ What’s the pain points in your current situation?
- ✓ What’s your use case? Can you give me an example?
- ✓ What’s the end product you wish to deliver?
- ✓ If we want to start small, are there any specific use cases you want to take first?

Framework: Help Formulate the Right Questions




EFFECTIVE DATA STORYTELLING
How to *drive change* with data, narrative, and visuals
BRENT DYKES
WILEY

Available on GDrive

Problem: A key challenge or issue your audience wants to address. Often, it is something they would like to make more efficient or effective than it currently is.

Outcome: A strategic goal or desired end result your audience wants to achieve. If a problem represents something occurring in your *current* state, an outcome signifies a preferred *future* state. The more explicit the outcome (a specific target), the more helpful it will be to your analysis.

Actions: The key activities and strategic initiatives your audience is putting into place to fix a problem or achieve a desired outcome. These actions attempt to close the gap between where an organization currently is and its desired future state.

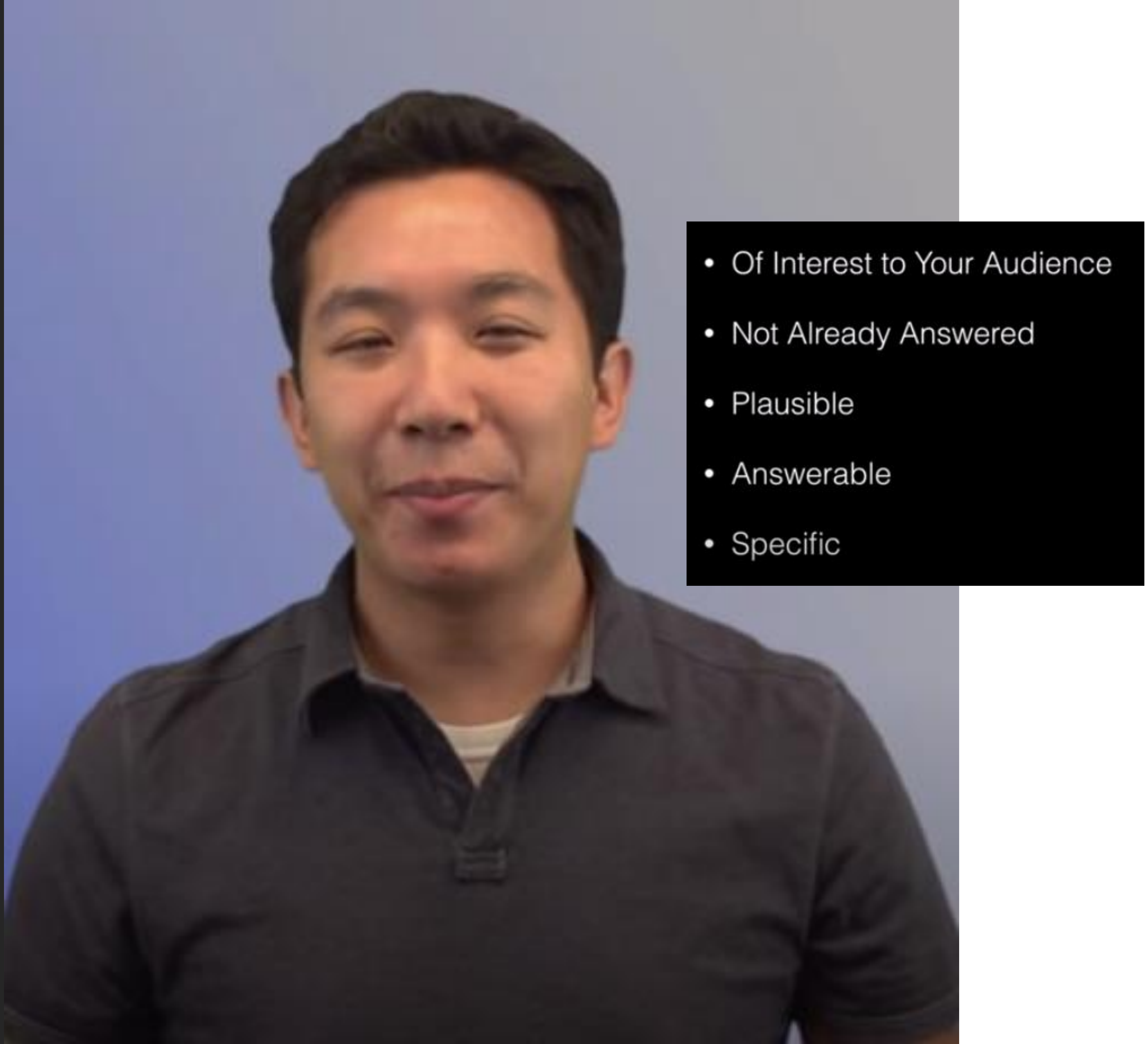
Measures: The key metrics and other data used to highlight the problem, monitor the effectiveness of the initiatives, and define the achievement of the desired outcome.

Characteristic of a Good Question

By Dr. Roger Peng

https://www.youtube.com/watch?v=cJYgCI4inU0

6:55 minutes



- Of Interest to Your Audience
- Not Already Answered
- Plausible
- Answerable
- Specific

Good Question

Let’s say you want to find out how rich people are across the world and how that differs in different geographic regions.

#1: Specific Terminology

Are the terms you are asking specific to that domain?

A good first start or “pre-question” would ask “How do economists measure the wealth of individuals?” or “what indicators measure individual wealth?” These might lead you to terms like GDP per capita, GNP per capita, etc.

#2: Specific Scope

You need to have a scope, or you’ll have to ask a series of follow up questions. For example, are you interested in particular countries, regions or continents? Are you concerned about their current wealth or you want data from xyz years ago?

#3: Specific Source

Source refers to a broad sense of whom you’re asking this question. Is it reasonable?

https://towardsdatascience.com/how-to-ask-good-questions-be41b517c1b

https://towardsdatascience.com/how-to-ask-questions-data-science-can-solve-e073d6a06236

- Traditional **hypothesis-driven** science was based on asking specific questions of the world and then generating the specific data needed to confirm or deny it.
- Today **data-driven** science focuses on generating data on a previously unheard of scale or resolution, in the belief that new discoveries will come as soon as one is able to look at it.

Both ways of thinking will be important to us:

- ✓Given a **problem**, what available **data** will help us answer it?
- ✓Given a **data set**, what interesting **problems** can we apply it to?

For a client, a data science project usually begins in one of two ways.
One, the client says, "We have this business situation, and we need help resolving it."
Or two, a client says, "We have this data set, and we think we could use it to benefit our business."
Think of it as clients who want to know *an answer* and clients who want to know *what's possible*.

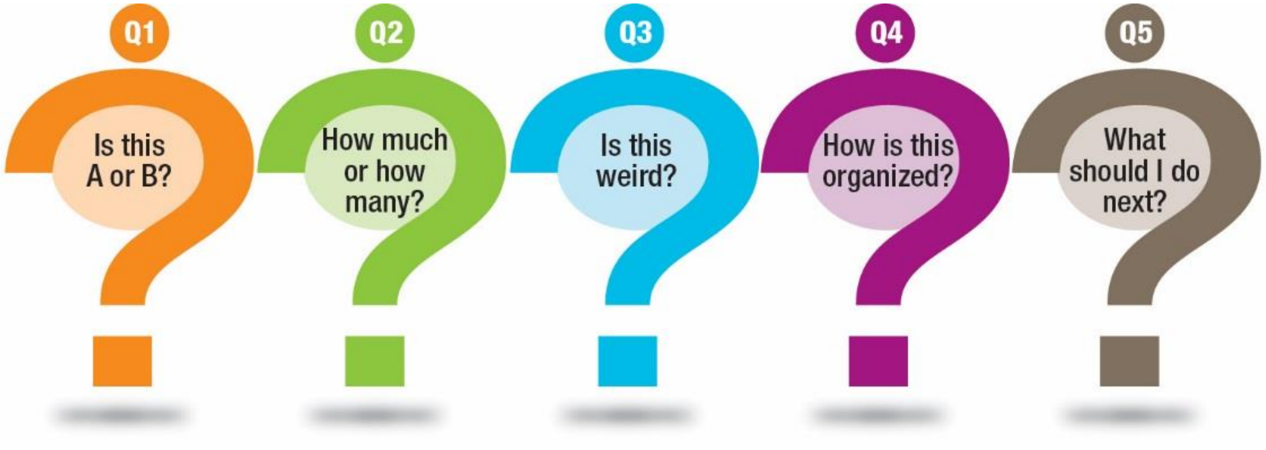
Data Scientists Always Ask Questions

Becoming a data scientist requires learning to **ask questions about data**.

Data scientist job is going to be to **turn numbers into insight**.

Data scientist ask **questions** like:

- *What things might you be able to learn from a given data set?*
- *What do you/your people really want to know about the world?*
- *What will it mean to you once you find out?*

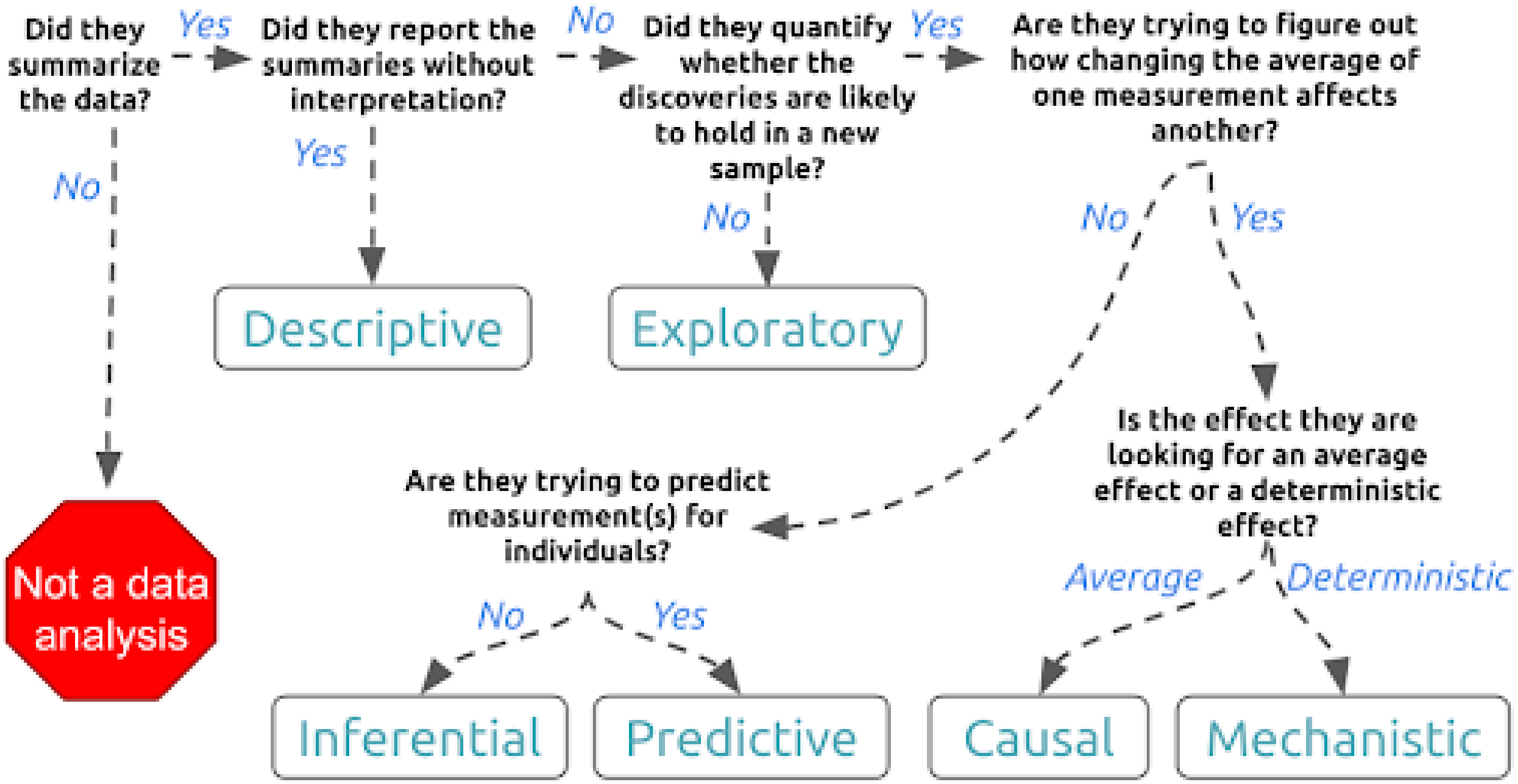


The Data Analytic Question

- **Data** can be used to **answer** many **questions**, but remember **NOT** all of them.
- Before performing a data analysis the key is to **define the type of question being asked**. Some questions are easier to answer with data and some are harder.
- There is a broad categorization of the types of data analysis questions, ranked by how easy it is to answer the question with data.

Source: **The Elements of Data Analytic Style** http://worldpece.org/sites/default/files/datastyle.pdf

Data analysis flowchart



Did they summarize the data?

Did they report the summaries without interpretation?

Did they quantify whether the discoveries are likely to hold in a new sample?

Are they trying to figure out how changing the average of one measurement affects another?

Is the effect they are looking for an average effect or a deterministic effect?

Are they trying to predict measurement(s) for individuals?

http://science.sciencemag.org/content/347/6228/1314

Descriptive

- A **descriptive data analysis** seeks to summarize the measurements in a **single number** without interpretation.
- **Examples** include determining the proportion of males, the mean number of servings of fresh fruits and vegetables per day, or the frequency of viral illnesses in a set of data collected from a group of individuals.
- Another **example** is the United States Census. The Census collects data on the residence type, location, age, sex, and race of all people in the United States at a fixed time.
- The **Census** is descriptive because the goal is to summarize the measurements in this fixed data set into population counts and describe how many people live in different parts of the United States.
- There is **no interpretation** of the result itself as the result is a **fact**.

Exploratory

- An **exploratory data analysis (EDA)** builds on a descriptive analysis by searching for **discoveries**, **trends**, **correlations**, or **relationships** between the measurements of multiple variables to generate ideas / hypotheses or discover new insights.

Examples:

  - Where are the accidents that led to the highest number of injuries or deaths?
  - What is the relationship between the driver’s age and the number of accidents?

- An exploratory analysis like this one seeks to make discoveries, but rarely can confirm or deny those discoveries.

Inferential

- An **inferential data analysis** goes beyond an exploratory analysis by quantifying whether an observed pattern will likely hold beyond the dataset in hand.
- Inferential data analyses are the **most common statistical analysis** in the formal scientific literature.
- An **example** is a study of whether air pollution correlates with life expectancy at the state level in the United States.
- The **goal** is to identify the strength of the relationship in both the specific data set and to determine whether that relationship will hold in future data.

Predictive

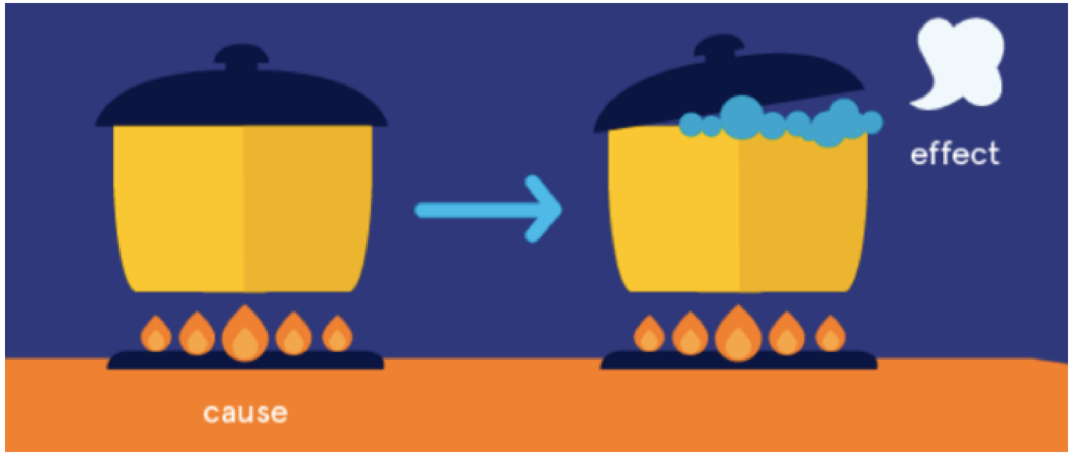
- While an inferential data analysis quantifies the relationships among measurements at population-scale, a **predictive data analysis** uses a subset of measurements (the features) to predict another measurement (the outcome) on a single person or unit.
- An **example** is when organizations like FiveThirtyEight.com use polling data to predict how people will vote on election day.
- In some cases, the set of measurements used to predict the outcome will be intuitive. There is an obvious reason why polling data may be useful for predicting voting behavior.
- But predictive data analyses only show that you **can predict one measurement from another**, they don’t necessarily explain why that choice of prediction works.

Causal

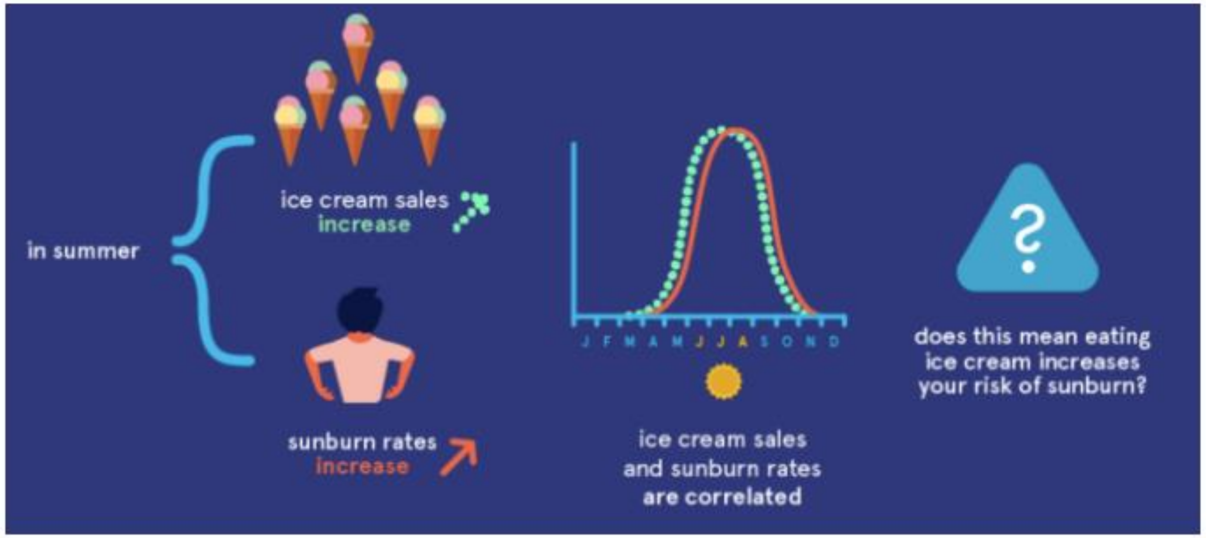
- A **causal data analysis** seeks to find out what happens (**effect**) to one measurement if you make another measurement change (**cause**).
- An **example** is a randomized clinical trial to identify whether fecal transplants reduces infections due to Clostridium difficile. In this study, patients were randomized to receive a fecal transplant plus standard care or simply standard care.
- In the resulting data, the researchers identified a relationship between transplants and infection outcomes.
- The researchers were able to determine that fecal transplants caused a reduction in infection outcomes.
- A causal data analysis identifies both the **magnitude and direction of relationships between variables**.

The Essence of Causation

> *It tries to answer the question: does one variable impact the other?*



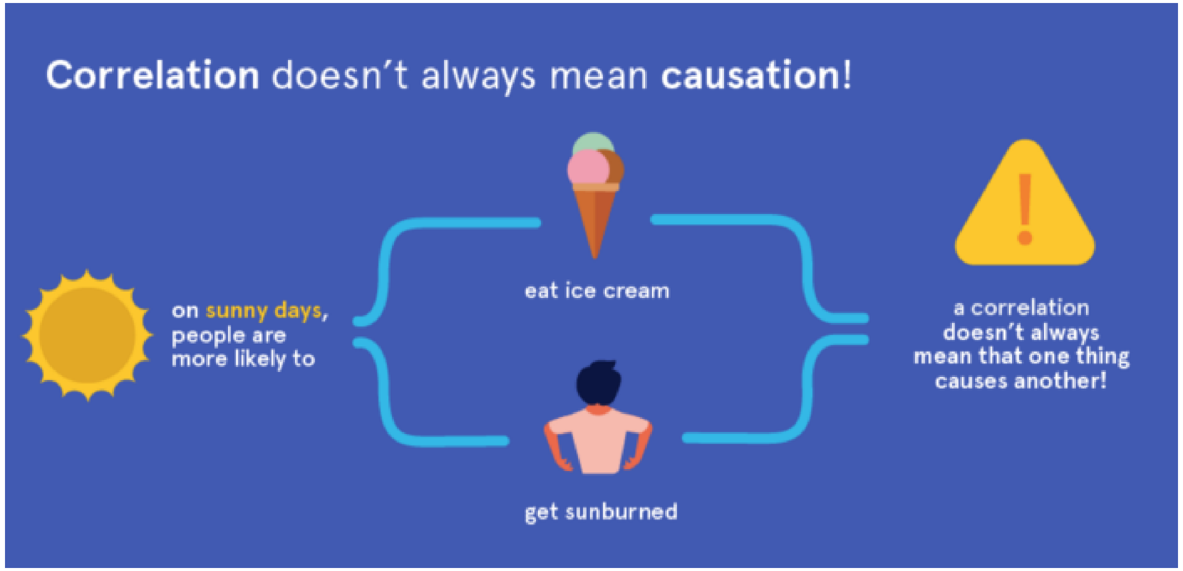
Causation: When one thing causes another to happen. Source:eufic.org



in summer
ice cream sales increase
sunburn rates increase
J F M A M J J A S O N D
ice cream sales and sunburn rates are correlated
?
does this mean eating ice cream increases your risk of sunburn?

Correlation: When two or more things appear to be real. Source:eufic.org

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables.
- simply a relationship between things.



Correlation doesn't always mean causation!
on sunny days, people are more likely to
eat ice cream
get sunburned
a correlation doesn't always mean that one thing causes another!

The more confident you become at identifying true correlations and causation within your dataset, the smarter you will be in data science domain.

Reference: https://medium.com/analytics-vidhya/correlation-causation-977f71bb1e36

Mechanistic

- Causal data analyses seek to identify average effects between often noisy variables.
- For **example**, decades of data show a clear causal relationship between smoking and cancer. If you smoke, it is a sure thing that your risk of cancer will increase. But it is not a sure thing that you will get cancer.
- The causal effect is real, but it is an effect on your average risk. A **mechanistic data analysis** seeks to demonstrate that changing one measurement always and exclusively leads to a specific, deterministic behavior in another.
- The **goal** is to not only understand that there is an effect, but how that effect operates.
- An **example** of a mechanistic analysis is analyzing data on how wing design changes air flow over a wing, leading to decreased drag. Outside of engineering, mechanistic data analysis is extremely challenging and rarely undertaken.

Q: What is the most important thing in Data Science?

- ❑ The data
- ❑ Hacking skills
- ❑ Working with large data sets
- ❑ The question you are trying to answer

Answer : ????

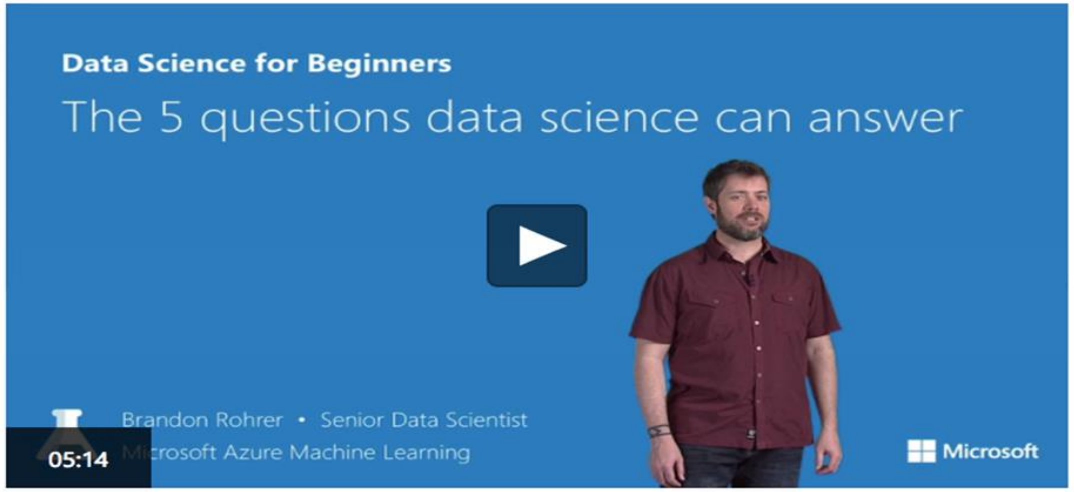
Examples of Questions

- **Hypothesis testing** tells us if our data applies to new situations. *“Do cat pictures drive more traffic than dog pictures?”*
- **Scenario analysis** analyzes many possible future outcomes under various conditions. We create many possible scenarios and then predict what will happen. *“What would happen if we raise the price of our product?”*
- **Optimization** generally asks simple, yet hard to answer, maximization and minimization questions. *“What supply routes minimize the cost of delivering packages?”*
- **Reinforcement learning** observes data and optimizes an outcome in real-time. *“When should I click to survive in the game Flappy Bird?”*

5 Questions Data Science (Machine Learning) Can Answer

Primary goal: To answer 5 types of questions using data – but to achieve and benefit from this, they also need to:

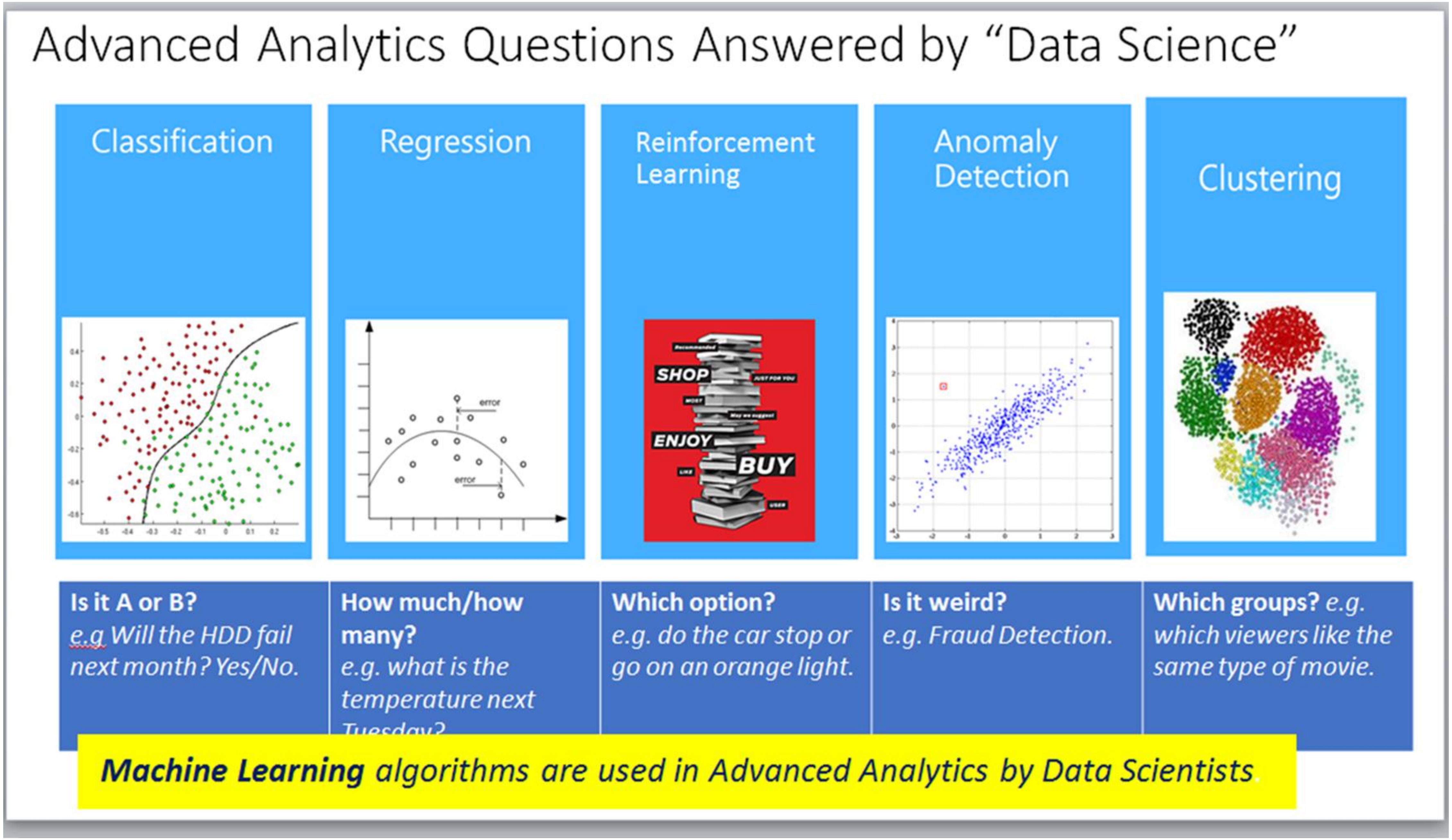
- Prepare data to answer the 5 questions (Data Engineering, Feature Engineering).
- To extract value on the answers from these 5 question (operationalizing data):
  - Predicting and forecasting.
  - AI/Intelligence/Cognitive applications.


Data Science for Beginners
The 5 questions data science can answer
Brandon Rohrer • Senior Data Scientist
Microsoft Azure Machine Learning
05:14
Microsoft

1. Is this A or B?
2. Is this weird?
3. How much/how many?
4. How is it organized?
5. What should I do next?

✓ CLICK HERE https://azure.microsoft.com/en-us/resources/videos/data-science-for-beginners-series-the-5-questions-data-science-answers/

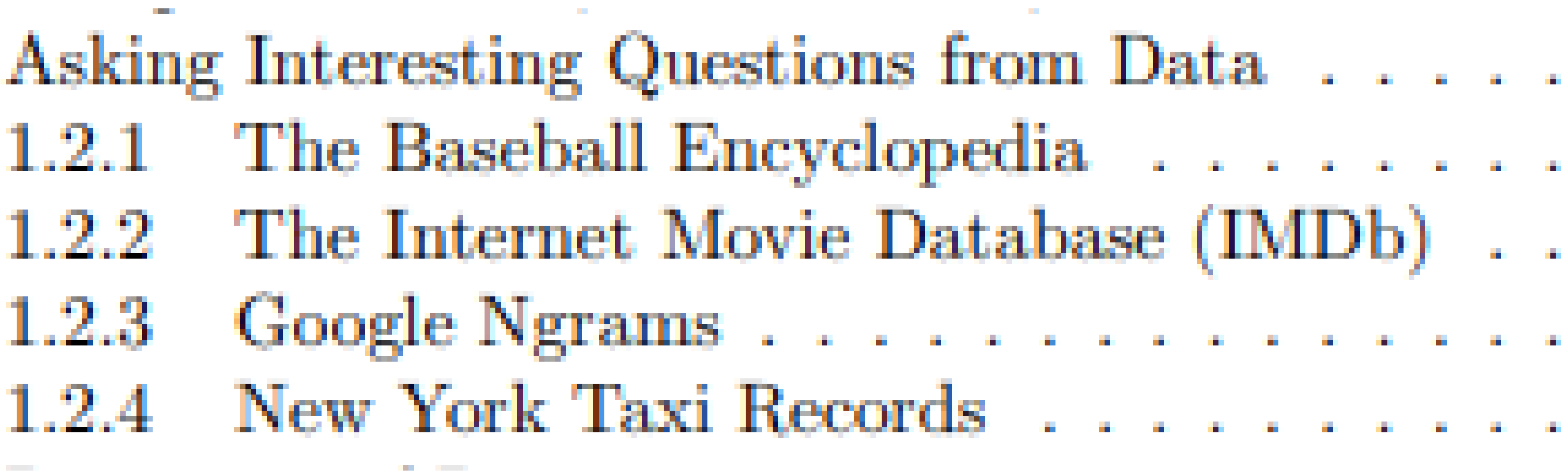
Brandon Rohrer https://www.linkedin.com/in/brohrer/



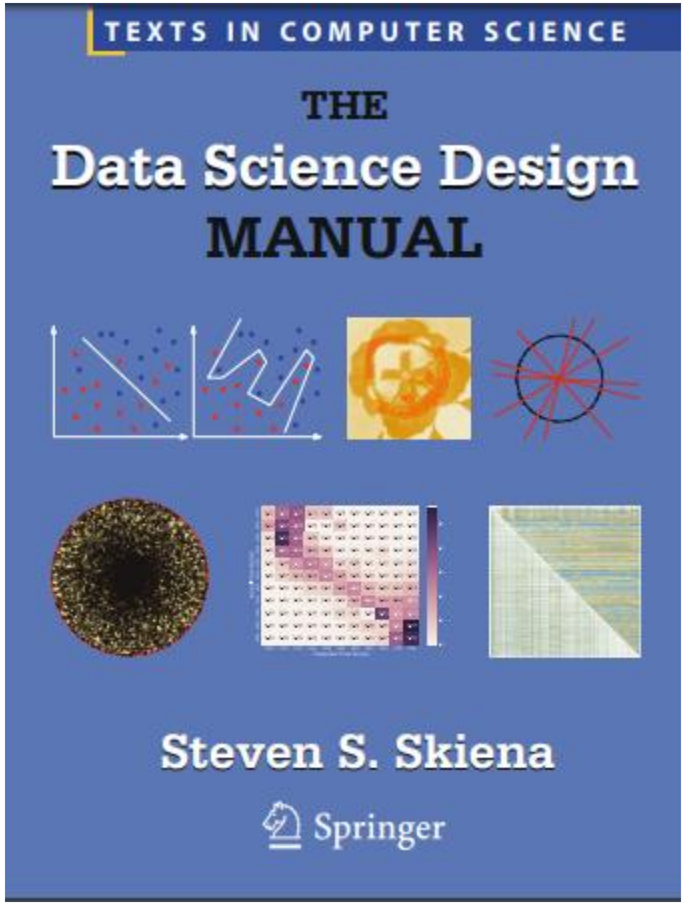
Advanced Analytics Questions Answered by "Data Science"

Classification	Regression	Reinforcement Learning	Anomaly Detection	Clustering
Is it A or B? *e.g Will the HDD fail next month? Yes/No.*	**How much/how many?** *e.g. what is the temperature next Tuesday?*	**Which option?** *e.g. do the car stop or go on an orange light.*	**Is it weird?** *e.g. Fraud Detection.*	**Which groups?** *e.g. which viewers like the same type of movie.*

***Machine Learning** algorithms are used in Advanced Analytics by Data Scientists*



Available on GDrive



TEXTS IN COMPUTER SCIENCE

THE Data Science Design MANUAL

Steven S. Skiena

Springer

For Group Project

Discuss the **type of questions** that you will ask.

- ➢Descriptive questions
- ➢Exploratory questions
- ➢Inferential questions
- ➢Predictive questions (optional)