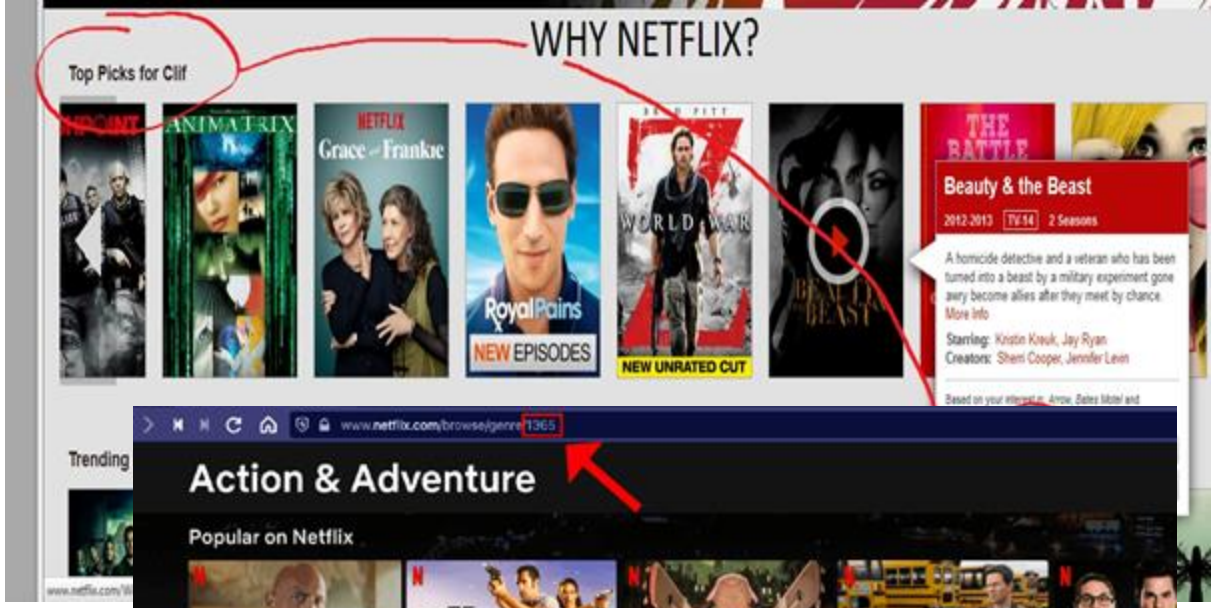
LECTURE: 01

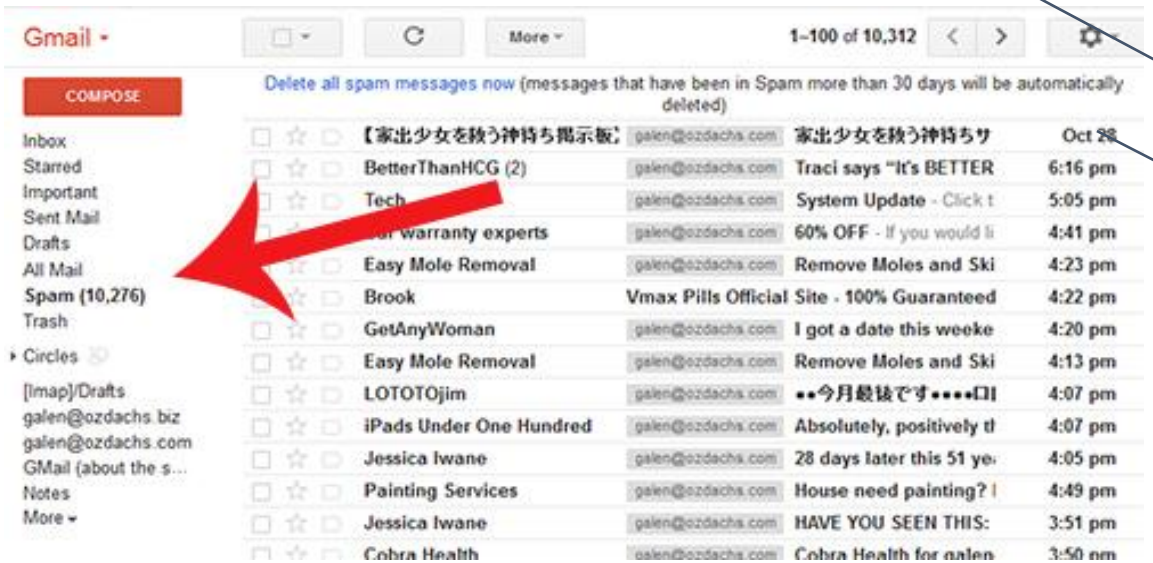
INTRODUCTION TO DATA SCIENCE

Today we will cover

1. What is Data Science?
2. What is Big Data and Its Key Features?
3. Overview of Data Science Life Cycle.
4. Career in Data Science.
5. Course Logistics and What to Expect.

What is data science?





Need availability of DATA!!!

DATA SCIENCE

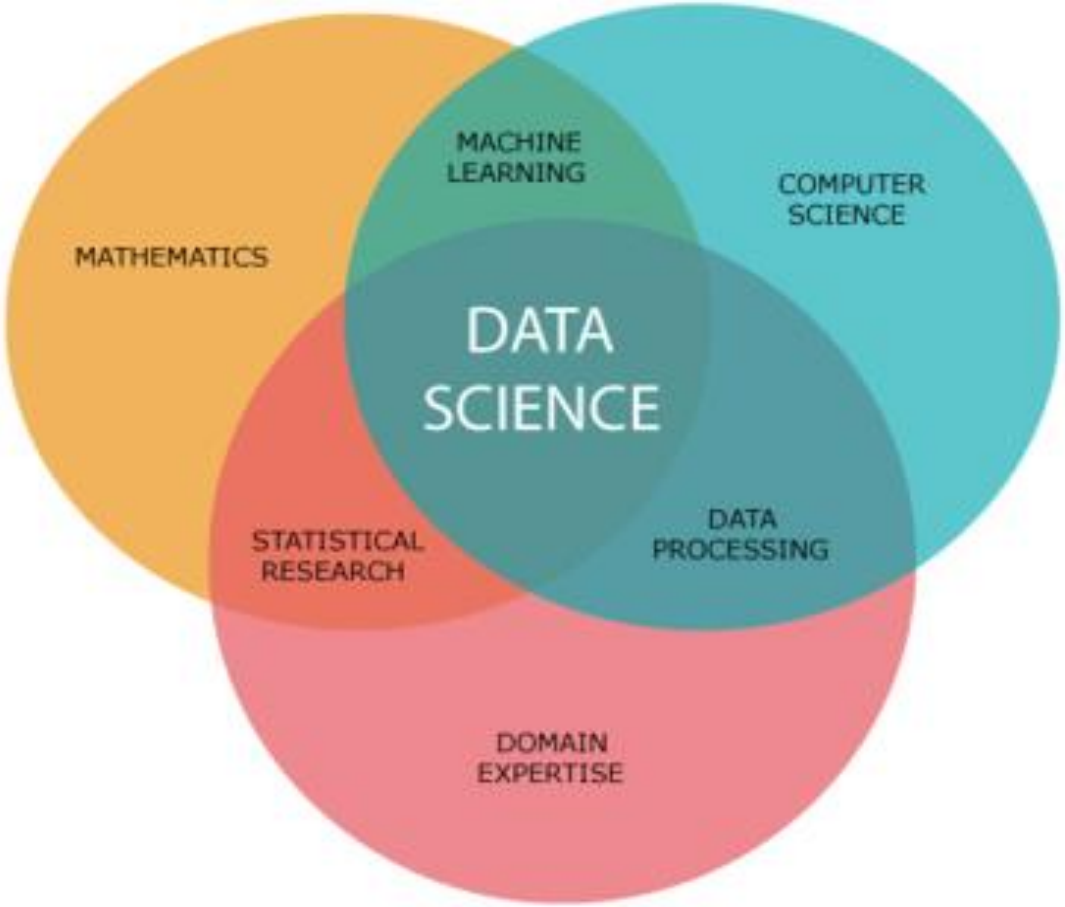
Data science is all about using data ***to solve problems***.

In simple terms: Using data *to draw an inference or predict an outcome*.
Such information can help us *to make better decisions*.

- **Decision making**
  - Which email is spam and which is not?
- **Product recommendation**
  - Which movie to watch?
- **Predicting the outcome**
  - Who will be the next President of the USA?
  - Many more ……

WHAT IS DATA SCIENCE?

Data science is an **interdisciplinary** field focused on **discovering patterns and describing relationships** using **data**.



Data science is the application of computational and statistical techniques to address or gain (managerial or scientific) insight into some problem in the real world.

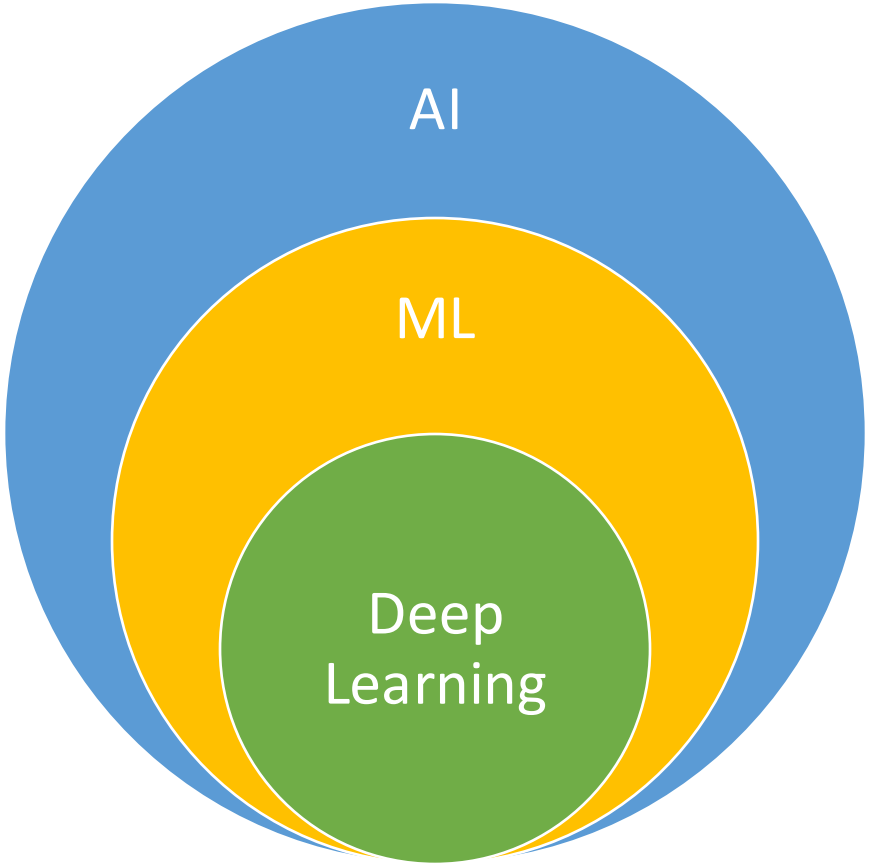
Zico Kolter
Machine Learning Prof, CMU
Chief Scientist of AI Research, Bosch

Definitions

***Artificial Intelligence*:** Any technique that enables computers to mimic human intelligence, e.g., using decision trees, rules, logic, ML

***Machine Learning*:** A subset of AI that uses statistical techniques that enable machines to use experience to improve at tasks

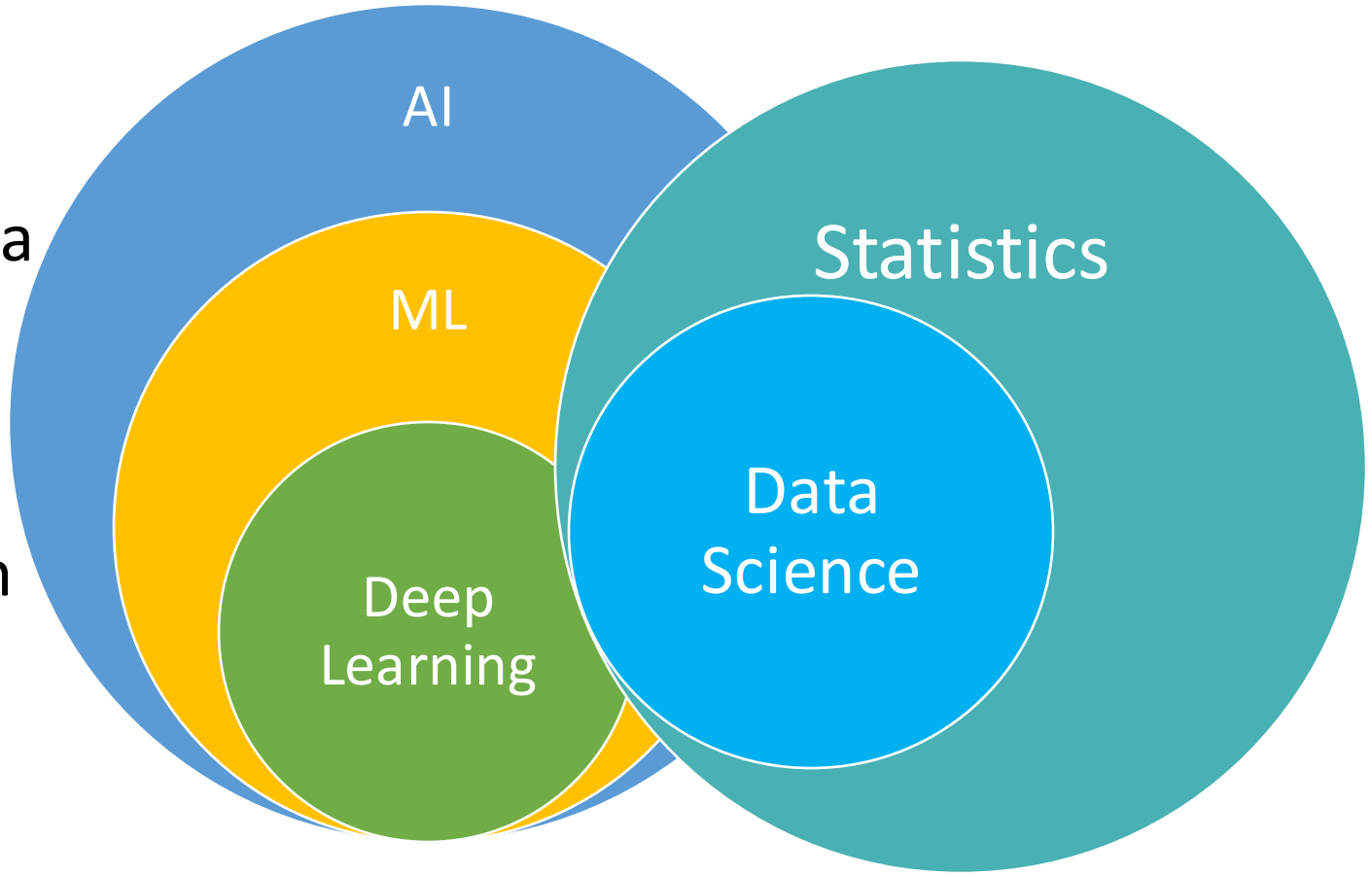
***Deep Learning*:** A subset of ML that uses vast amounts of data and multilayer neural nets to enable a machine to train itself to perform tasks



Definitions

***Statistics*:** science of collecting and analyzing numerical data, either to describe properties of a dataset or to make inferences based on a subset of data

***Data Science*:** an applied branch of statistics that uses computer science techniques to manage, analyze, visualize and discover patterns in data



BIG DATA AND DATA SCIENCE

The rise of **data science** over the last 20 years is partially a result of **big data**.

Big data describes datasets with **large volume**, created and updated with **high velocity**, that have **variety in structure and format**.

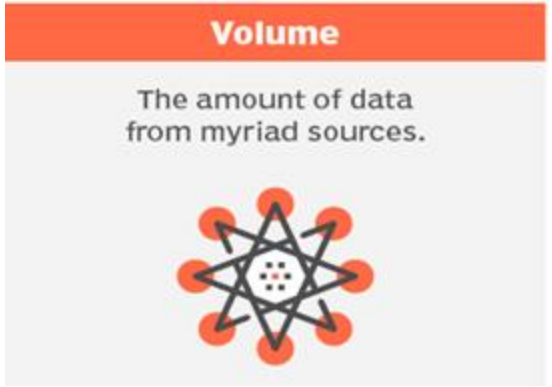
As more companies and organizations collect and use big data, the demand for people with data science skills grows.

The 3 V's of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3 V's: *volume*, *velocity* and *variety*.

3Vs of Big Data

BIG DATA AND DATA SCIENCE

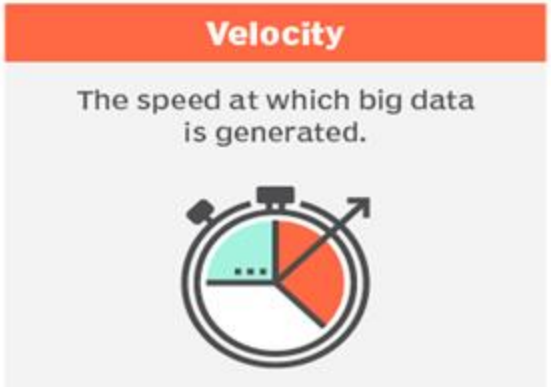


3 V's:

- **<u>Volume:</u>** vast amount of data generated or collected. the volume of global data has increased exponentially.

  Data scientists use specialized tools and software to work with big datasets, such as:
  - **Apache Spark**,
  - **Hadoop**,
  - **Cloud-based storage**, like *Amazon Web Services*, *Google Cloud*, or *Microsoft Azure* for storing and analyzing large amounts of data.

BIG DATA AND DATA SCIENCE

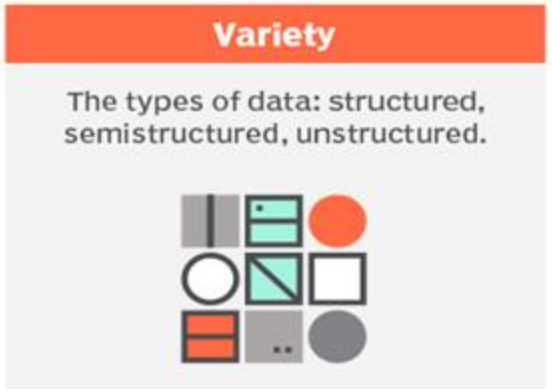


3 V's:

- **Velocity:** Represents the speed at which data is generated, processed, and analyzed, often in real-time or near to real-time.

  Think about everyday:
  - **900** million photos are uploaded on Facebook,
  - **500** million tweets are posted on Twitter,
  - **0.4** million hours of video are uploaded on Youtube and
  - **3.5** billion searches are performed in Google.
  - This is like a nuclear data explosion

BIG DATA AND DATA SCIENCE



3 V's:

- **<u>Variety:</u>** Big data comes in a variety of forms: tables, spreadsheets, images, videos, sound, text, etc.

  Data scientists deal with variety in data by *using techniques from*
  - statistics,
  - computer science,
  - machine learning, and
  - artificial intelligence.
  - The ability to deal with data variety helps set data science apart from other computational and analytical fields.

EXAMPLE: ANALYZE PATIENT'S MEDICAL RECORD, PREDICT DISEASE OUTBREAK

Applies Techniques and Models: Analyze patient data to identify risk factors, predict disease progression, and recommend personalized treatments.

Raw Data: Medical records, patient demographics, lab results.

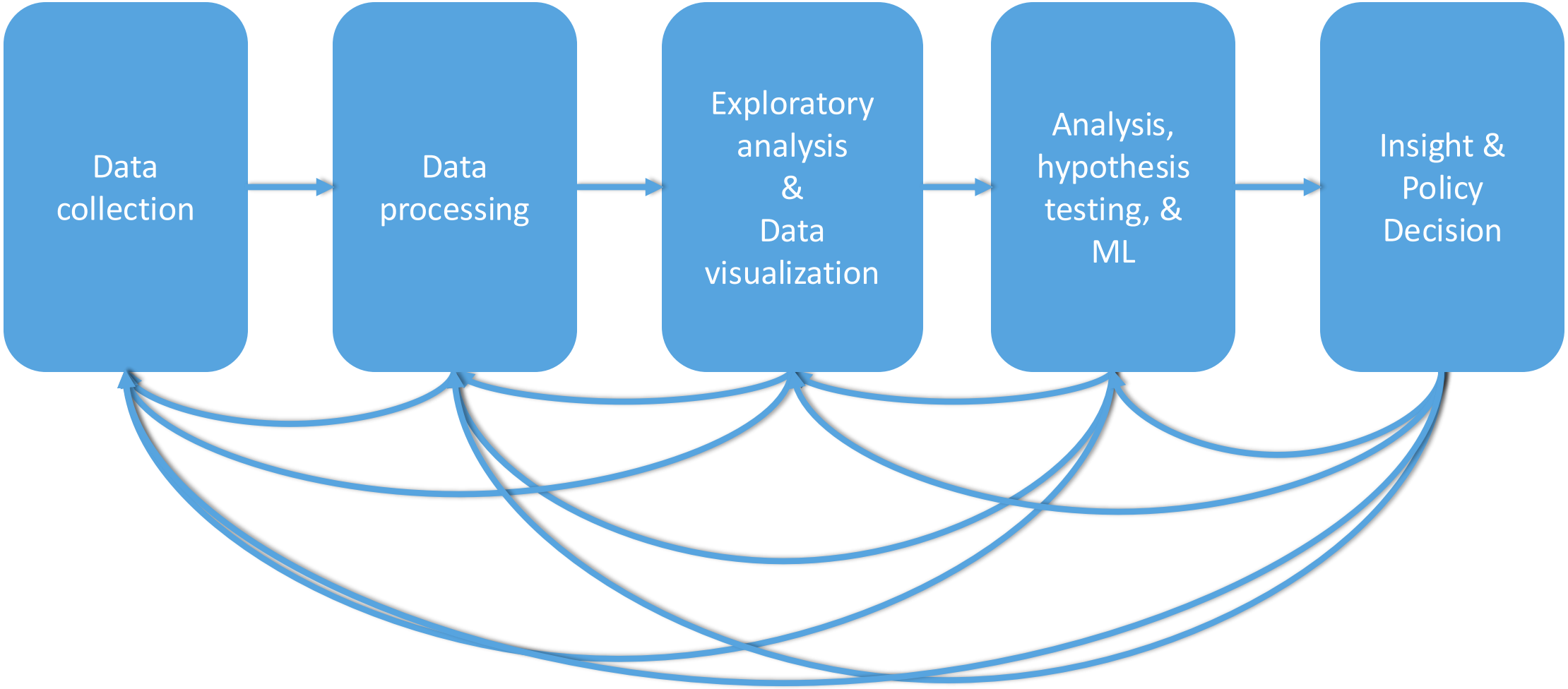
Insights and Predictions: Identifies patterns in patient data to

- predict disease outbreaks,
- optimize treatment plans,
- provide insights for medical research.

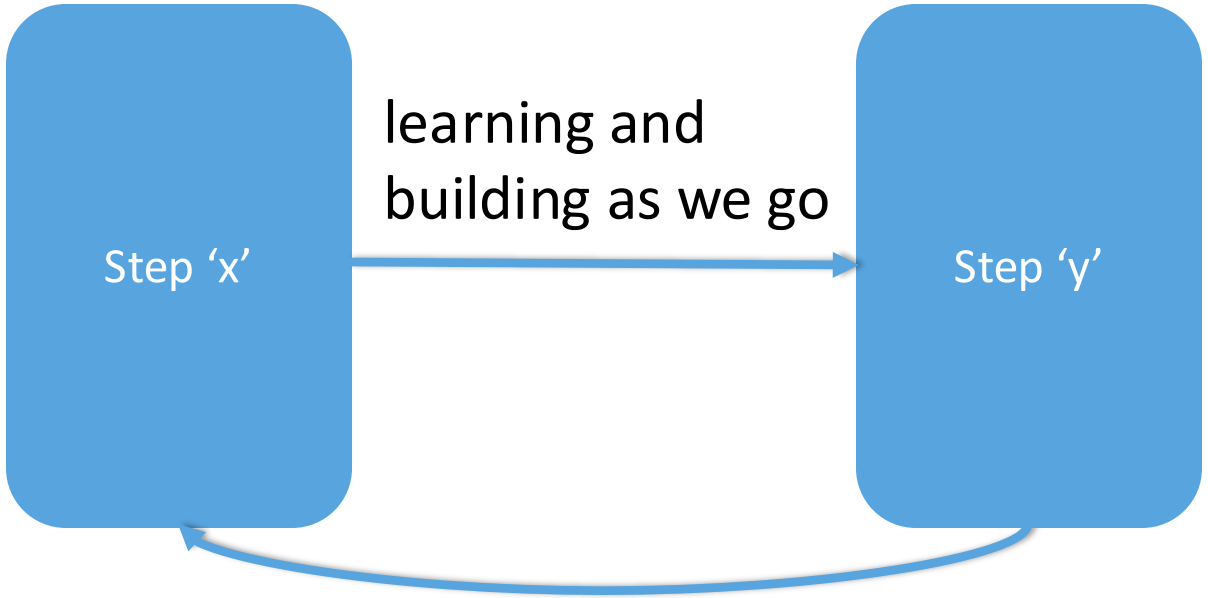
SOME MORE EXAMPLES

- Given the results of a **drug trial**, determine if the drug is effective.
- Given a dataset of **movies with ratings**, predict what movies someone will like.
- Given a set of **labeled images**, identify what is in a given picture.
- Given a dataset **describing people** who have and haven’t paid off their loans, predict if a new person will repay a loan.

THE DATA LIFECYCLE



REMEMBER: DATA SCIENCE IS NOT A STRICTLY ONE-WAY LINEAR PROCESS; IT'S DYNAMIC, ITERATIVE, AND ADAPTIVE



If need to revisit previous steps due to new insights or challenges that arise during later stage. **Example:** Later realize you need to collect more data. This allows for *constant refinement* and *improvement* as new information emerges and insights evolve.

BEFORE THAT: DEFINE PROBLEM STATEMENT

What problem are you going to solve?

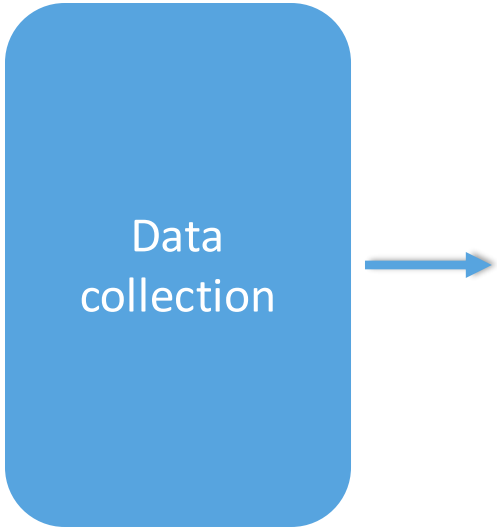
- Why do we need a well-defined problem statement?

Example: “I want to increase the profit” - is it a well defined problem statement?

How much to increase the average profit/ revenue such as 20% or 30% ?
What is the average time frame to increase the revenue?

A problem well defined is a problem half-solved. — Charles Kettering

THE DATA LIFECYCLE



1. COLLECT DATA

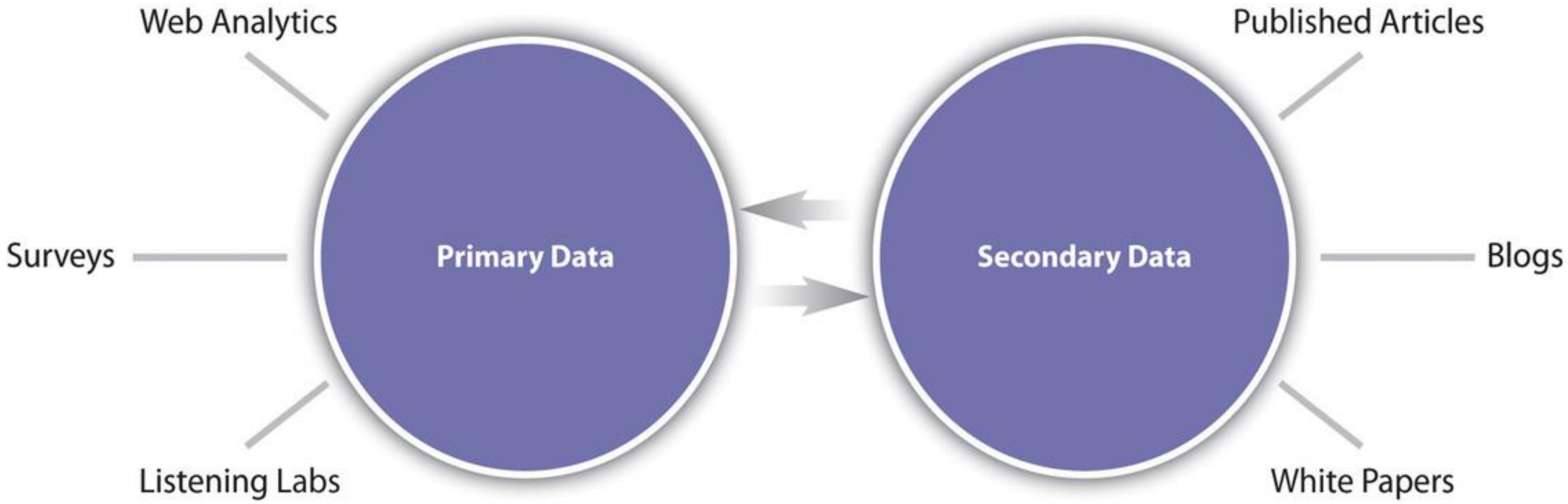
Data collection is a systematic approach to gather relevant information from a variety of sources.

- Gathered from *external sources*
- Gathered from *existing company databases*
- Gathered by *tools created by you*

DATA COLLECTIONS METHODS:

Two types of data collections methods:

1. Primary Data Collection
2. Secondary Data Collections



DATA COLLECTIONS METHODS: PRIMARY

Situation: Some unique problem and no related research is done on the subject.

Solution: Collect new data → Primary data collection.

Example: Average time that employees spend during lunch break across companies.

- ***Problem*** → No public data available of these.
- ***Solution:*** Collect the data through various methods.
- ***Different Methods****: Surveys*, *Interviews of employees* and by *Monitoring the time spent* by employees in cafeteria.
- This method is time consuming

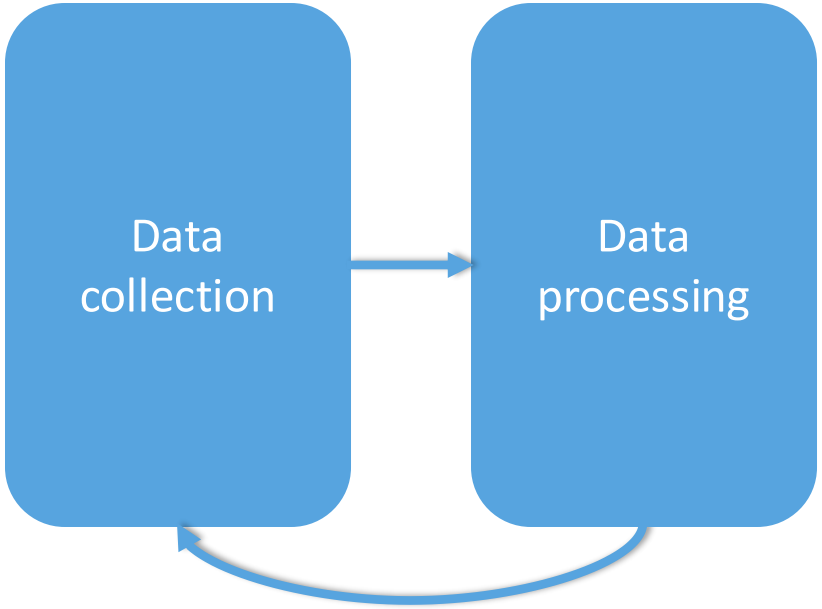
DATA COLLECTIONS METHODS: SECONDARY

Situation: Some problem and the data is readily available or collected by someone else.

Solution: Use the data → Secondary data collection.

- ***Different Methods****:* The *internet*, *news articles*, *government register/poll*, *magazines* and so on.
- This method is less time consuming than the primary method.

THE DATA LIFECYCLE



2. DATA PROCESSING

Clean or scrub data to ensure the data quality

- **Important:** Do sanity check on data (*a preliminary validation process to confirm data is reasonable, valid, and usable for analysis*)

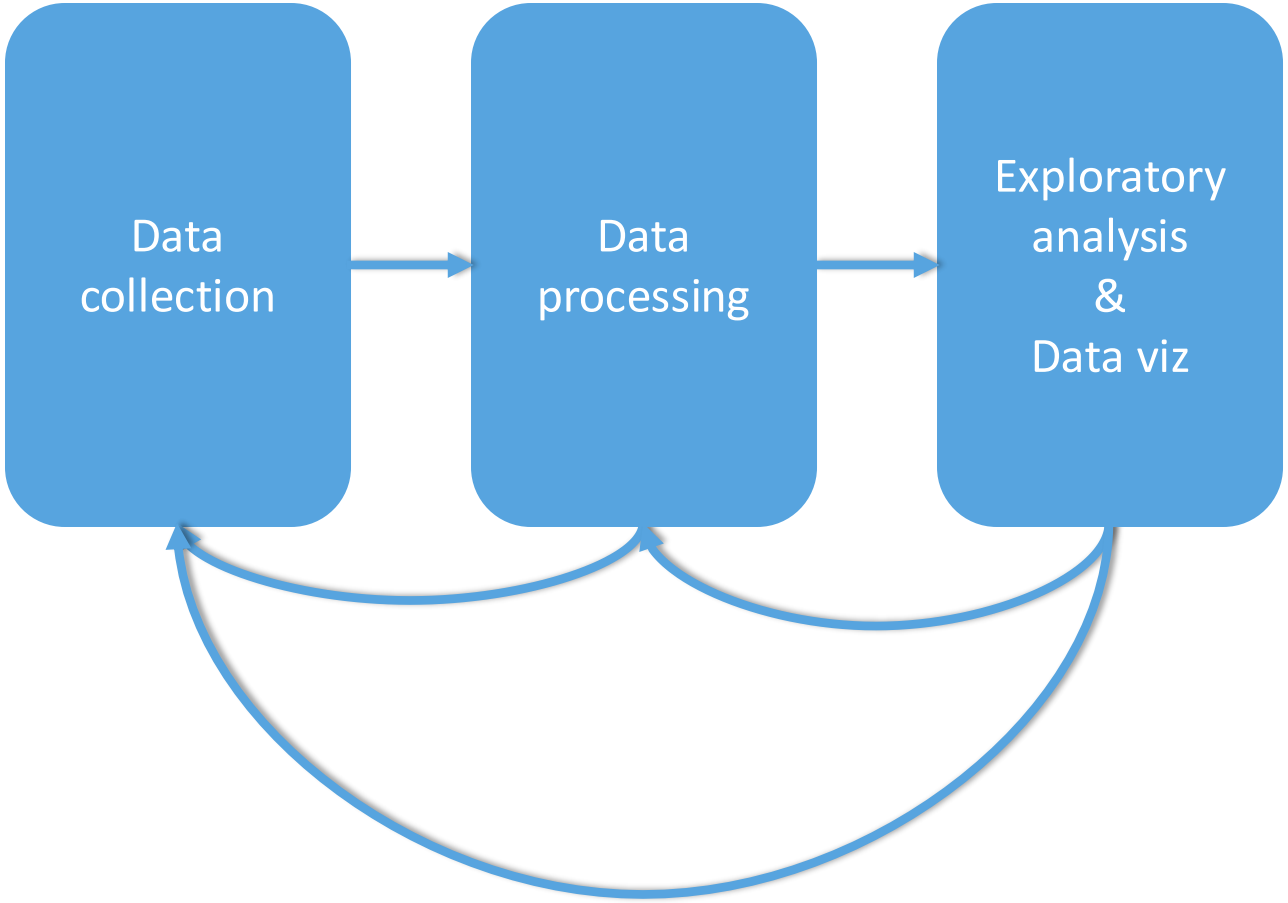
- **Why ?** Bad quality may lead to unexpected results or misleading information.
  - Deal with duplicates
  - Formatting
  - Weird outliers
  - Mistakes

2. DATA PROCESSING: EXAMPLE

You Collect data about students' test scores.

- But → Some left SCORE blank, and others wrote "N/A” (Missing Values).
- Process Data
  - Replace "N/A" entries with a neutral value like 0 and
  - Fill in the missing scores with the correct value

THE DATA LIFECYCLE



3. EXPLORE DATA: FIGURE OUT WHAT YOU HAVE

You'll be sitting on like a terabyte of raw data

- What is there?
- Are there any interesting correlations?
- Do you have everything you need?

3. EXPLORE DATA CONT.

- Extract useful insights from the data, *understanding patterns*, and setting the stage for effective *model building* and *decision-making*
- Important to *analyse the data* and build familiarity with the data
- Skipping this step may lead to inaccurate models as well as insignificant variables in your models.

"**It is important to understand *what you CAN DO* before you learn to measure how WELL you seem to have DONE it**." – *As quoted by John Tukey, developer of Exploratory Data Analysis*

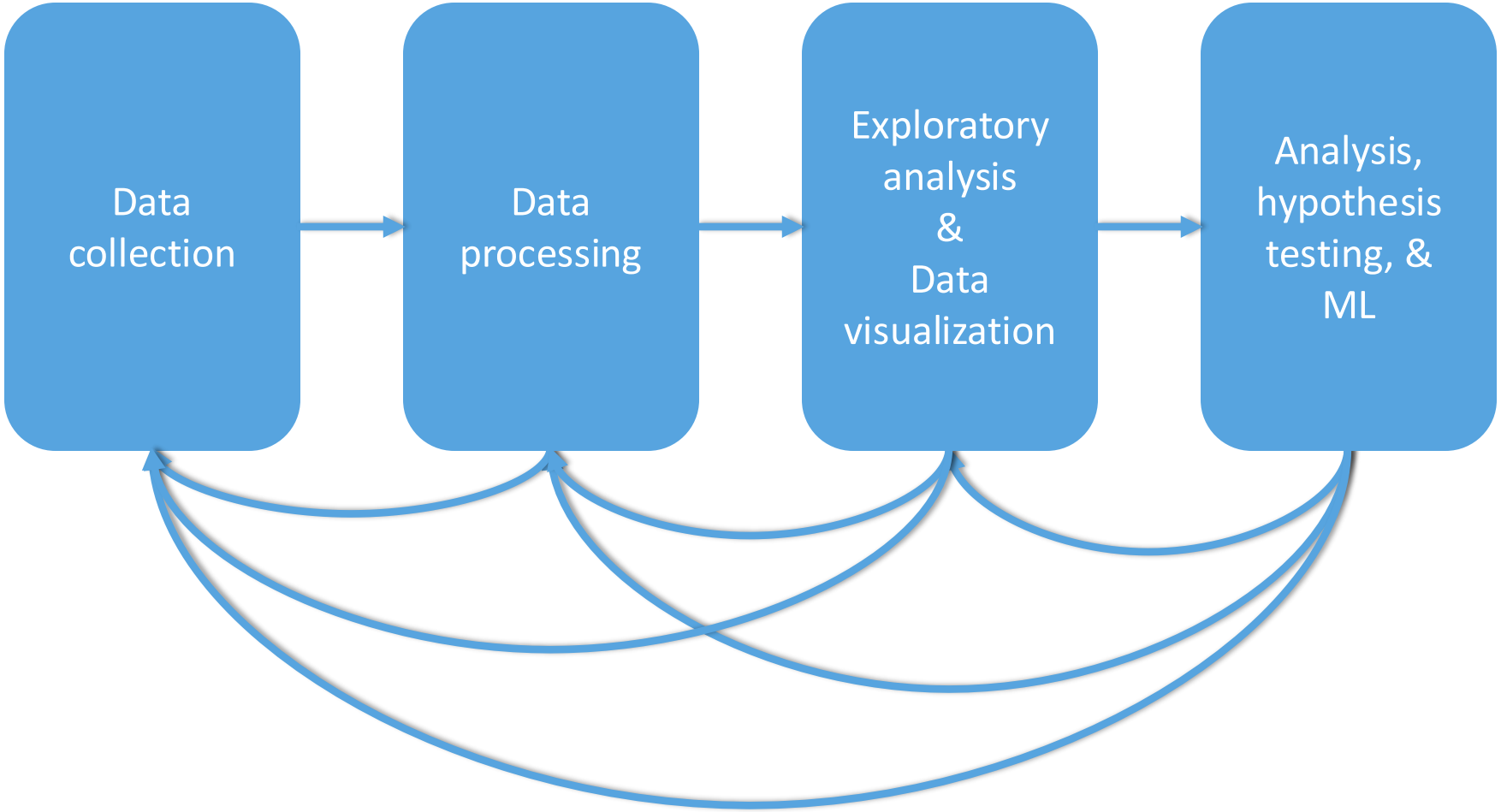
3. EXPLORE DATA: EXAMPLE

You want to understand more about the scores and how they relate to other factors like study hours.

- Exploratory Analysis: Find Connection between *study hours and scores* by calculating the average score, find the range of scores, and notice that some students scored exceptionally well.

- Data Visualization: Visually show the relationship between study hours and average scores for different groups of students in order to reveal that students who study more tend to have higher average scores.

THE DATA LIFECYCLE



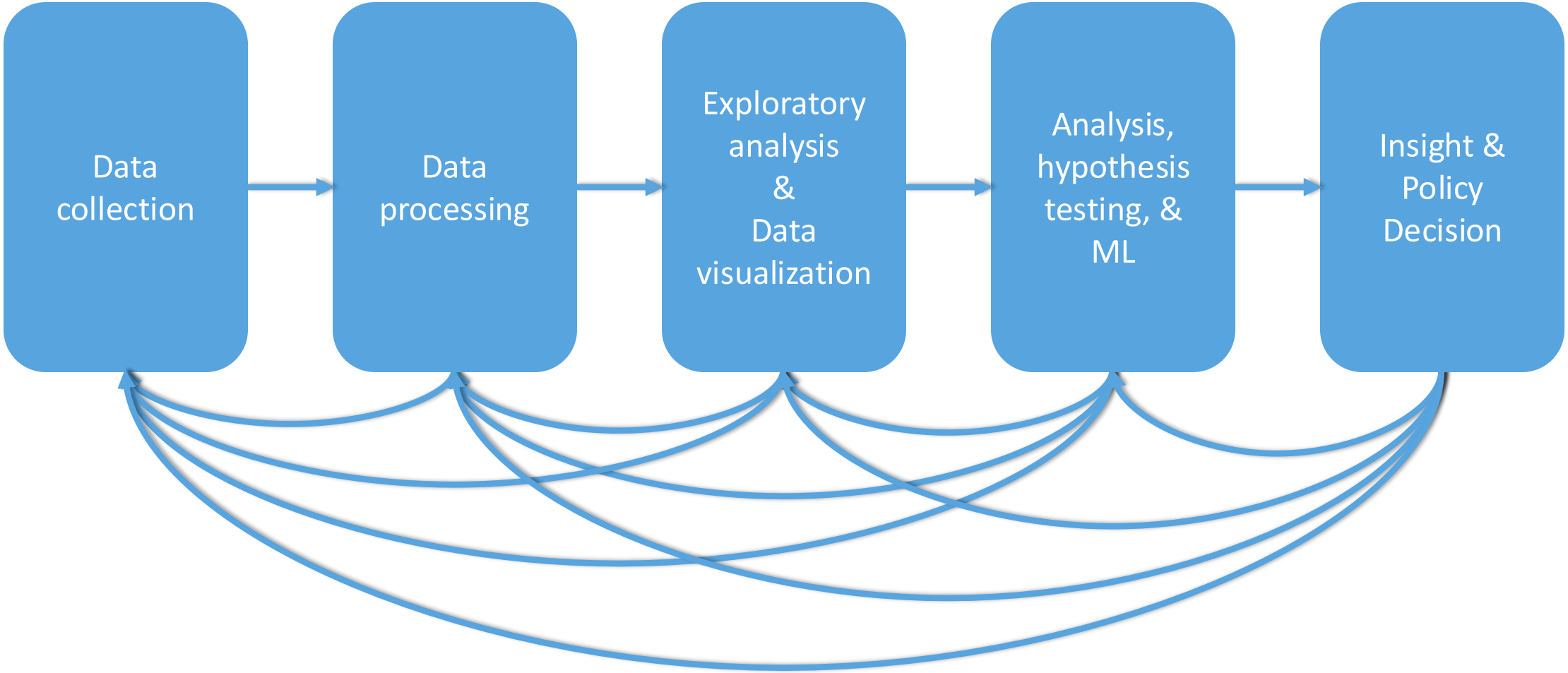
4. BUILD A MODEL

Steps to the solution.

Example: build a machine learning model that **predicts a student's test score based on their study hours**.

- Train the model → using the historical data,
- Once it's trained, **you can input a student's study hours to get a predicted test score**.

THE DATA LIFECYCLE



5. INTERPRETATION **Translate these findings into actionable insights.**

- Deriving Insights and make policy decisions if needed
- Present the results from your analysis to the stakeholders.
- Convince People: Explain the specific conclusion and critical findings, probably in understandable manner.

The last step is getting a bunch of non-technical people to understand what your magical model is doing and why it's right and they should listen to you.

5. INTERPRETATION

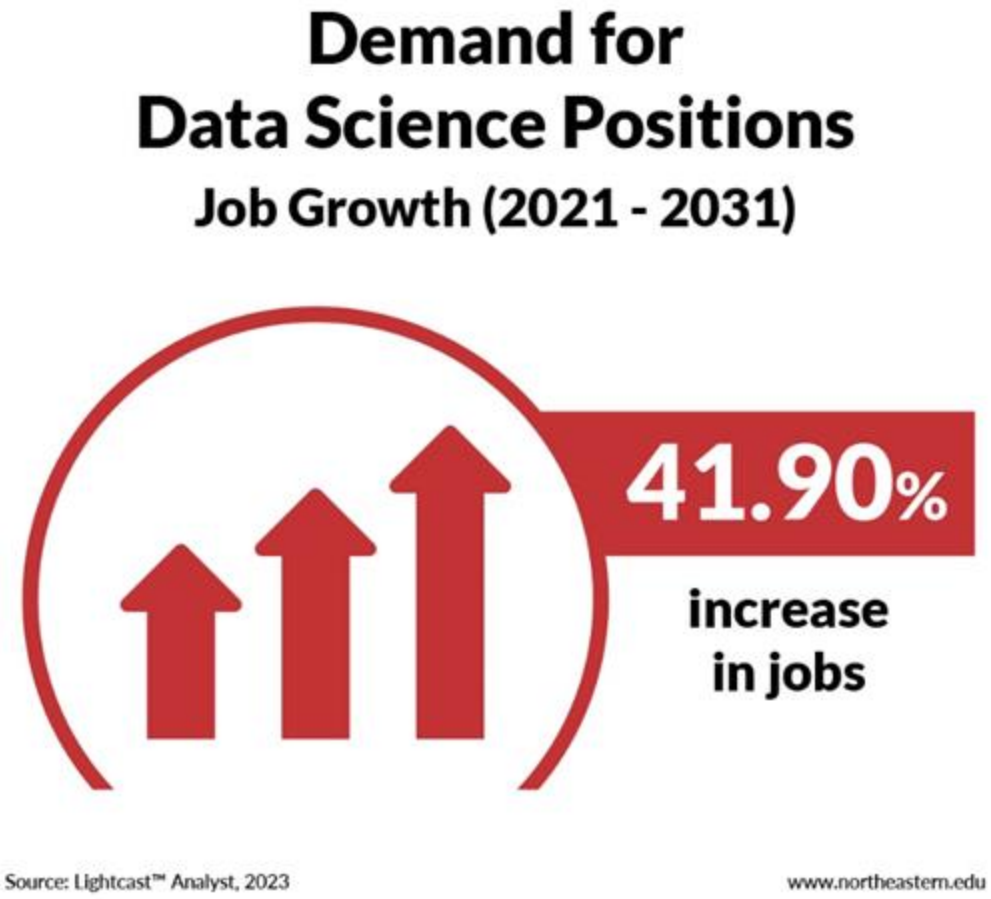
Example:

- Let's say your analysis confirmed that there's a strong positive correlation between study hours and test scores.
- This insight is crucial for both students and educators to understand the importance of consistent study habits.

- **Policy Decisions:** You can lead to policy decisions that encourage regular study time or improved teaching methods.
- **Convincing people to do what you want:**
  - Often, you must share findings with non-technical groups like marketing or executives.
  - Your goal is clear communication, allowing stakeholders to create actionable plans based on the results.

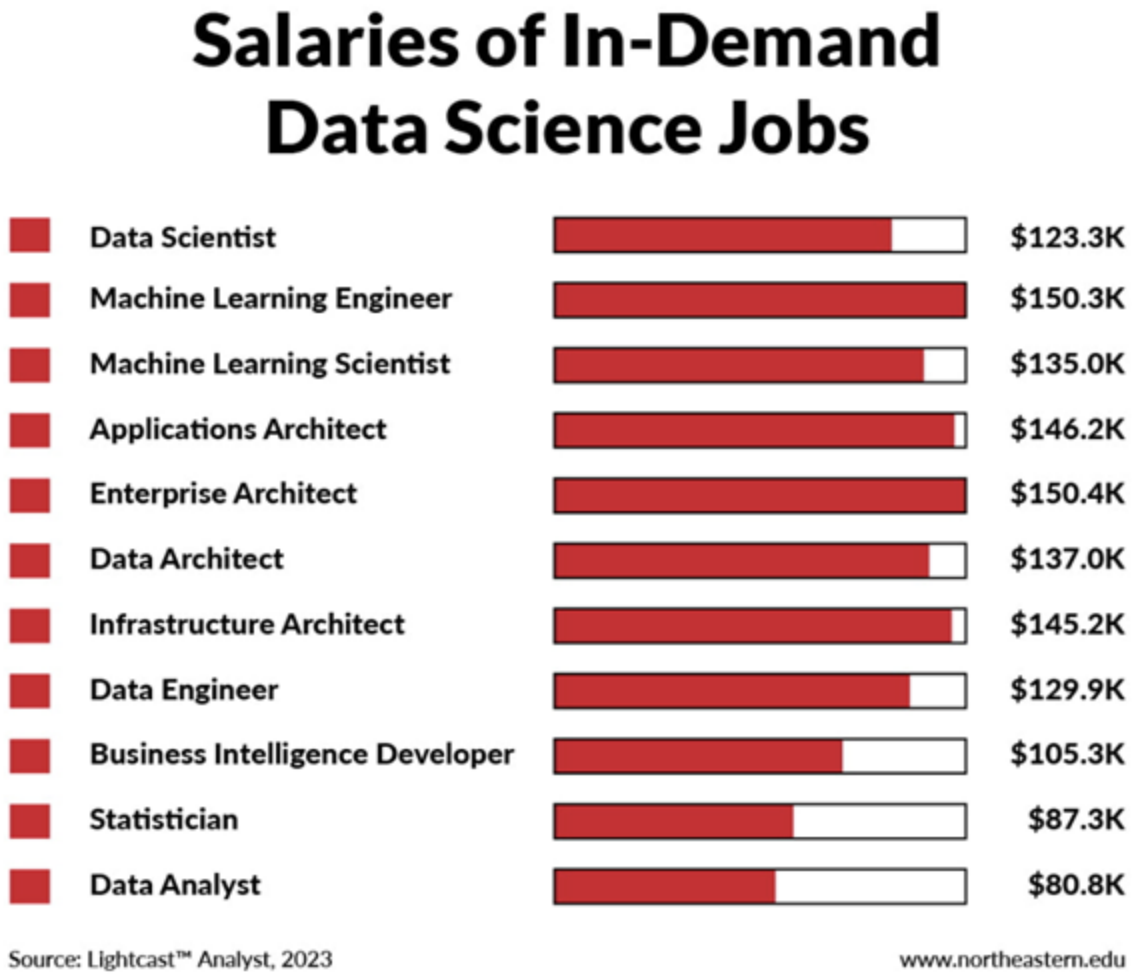
Career in Data Science

- **High Demand**
- **Job Growth**
- **High Salary Potential:**
- **Versatility:** Applies to various industries (e.g., healthcare, finance, marketing, tech)



Career in Data Science

- **High Demand**
- **Job Growth**
- **High Salary Potential:**
- **Versatility:** Applies to various industries (e.g., healthcare, finance, marketing, tech)



CAREER IN DATA SCIENCE

Data Analyst	Data Scientist	Machine Learning Engineer
Focus: • Primarily deals with analyzing and interpreting data to provide actionable insights, answering specific business questions • Emphasizes data visualization, statistical analysis, and creating reports.	**Focus:** • Involves both analysis and interpretation of complex data. • Applies statistical models and machine learning algorithms to extract insights. • Emphasizes pattern recognition and predictive **Probable Skills:** modeling.	**Focus:** • Specializes in deploying and operationalizing machine learning models. • Focuses on the technical implementation, scaling, and optimization of models for production.
Probable Skills: Proficient in statistical analysis, data cleaning, and visualization tools. Strong Excel skills and familiarity with databases. May have knowledge of basic programming for data manipulation.	**Probable Skills:** Strong statistical and mathematical background. Proficient in programming languages (e.g., Python, R). Expertise in machine learning and data modeling.	**Probable Skills:** Strong programming and software engineering skills. Expertise in machine learning frameworks and libraries. Knowledge of model deployment and optimization.

Career in Data Science

Required Skills

- **Technical:**
  - Programming (Python, R, SQL)
  - Data manipulation & cleaning (Pandas, NumPy)
  - Machine Learning (Scikit-learn, TensorFlow, PyTorch)
  - Data visualization (Matplotlib, Tableau, Power BI)
- **Soft Skills:**
  - Critical thinking and problem-solving
  - Communication (to explain findings to non-technical stakeholders)
  - Curiosity and learning agility

"The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades …"

Hal Varian
Chief Economist at Google

THIS COURSE

You'll learn to take data:

- Process it
- Visualize it
- Understand it
- Communicate it
- Extract value from it

SOME TECHNOLOGIES WE MIGHT USE (MOSTLY)

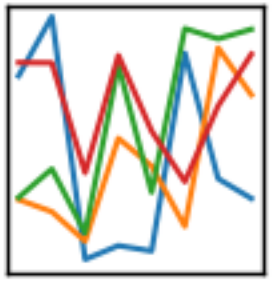
python™

jupyter

colab

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Overleaf