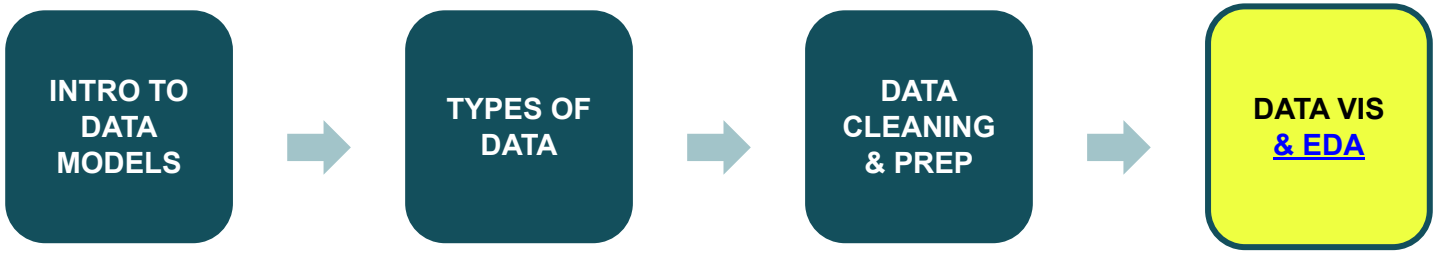
Exploratory Data Analysis

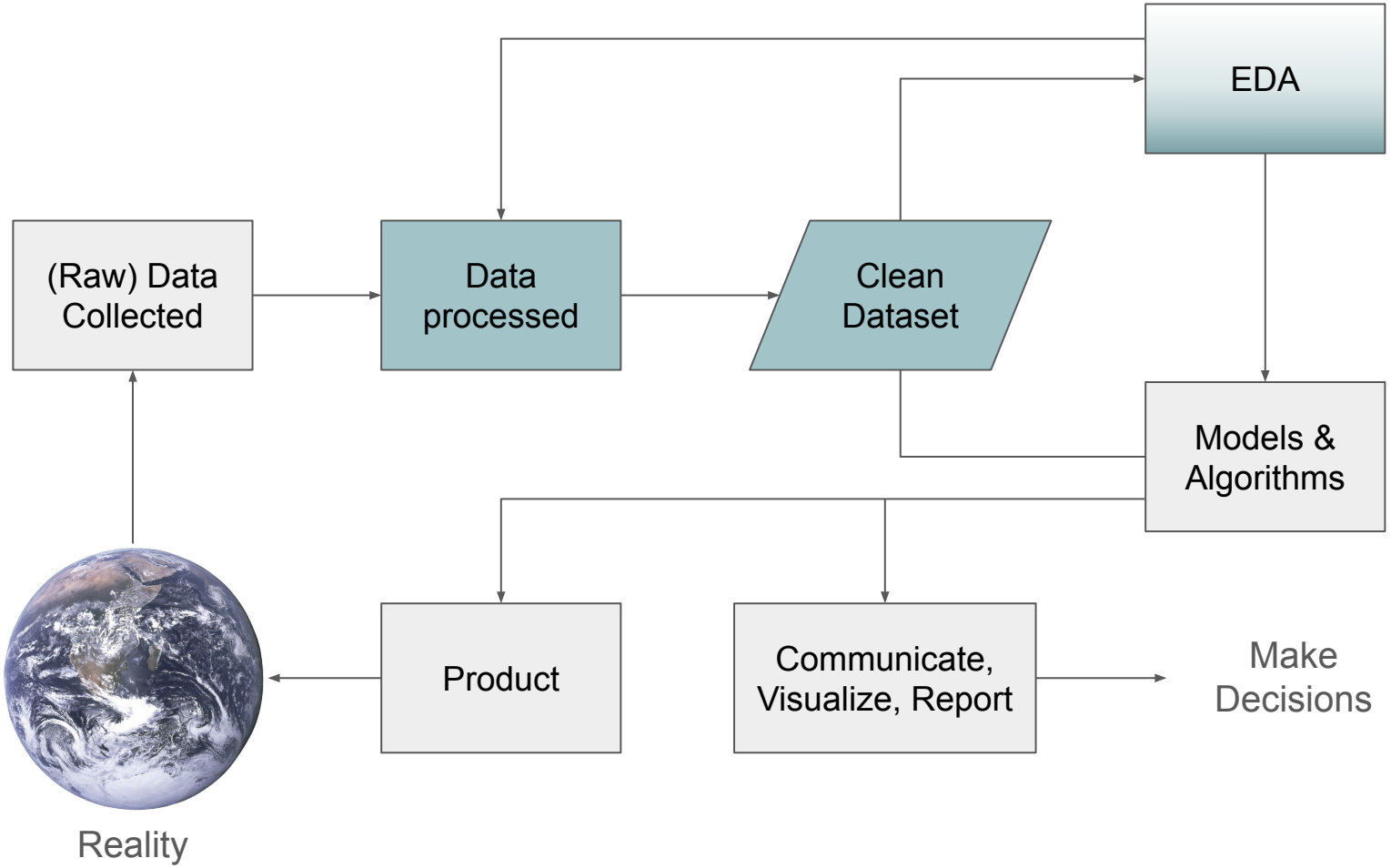
CPE 232: Data Models

Dr. Sansiri Tarnpradab

Department of Computer Engineering, KMUTT

Review


INTRO TO DATA MODELS
TYPES OF DATA
DATA CLEANING & PREP
DATA VIS & EDA


EDA
(Raw) Data Collected
Data processed
Clean Dataset
Models & Algorithms
Product
Communicate, Visualize, Report
Make Decisions
Reality

Data Exploration

- Initial step in data analysis
- To explore the data
- To see a bigger picture
- To uncover patterns and insights
- Usually via:
  - Visualization tools
  - Statistical techniques
- Aka Exploratory Data Analysis (EDA)

Ref: https://www.exploringdata.org/post/how-to-explore-data-dataexplorer-package/


KIRKLAND
Signature
ITM. / ART. 12173
Jelly Belly
THE ORIGINAL GOURMET JELLY BEANS
49 FLAVORS
ITM. / ART. 12173
NET WT 64 OZ (4 LB) (1.81 kg)

Importance of Context

Let's conduct EDA!

- ❏ Data
- ❏ Data Preparation
- ❏ Data Visualization

Putting them all together in order to…

- ➔ Identify obvious errors, outliers, etc
- ➔ Better understand patterns within the data
- ➔ Find interesting relations among variables

Some Steps

- ❏ Understand the dataset
- ❏ Handle missing data
- ❏ Get statistics of the data
- ❏ Detect outliers (if any)
- ❏ Explore data relationships
- ❏ Data transformations
- ❏ Draw insights

Note: The order above could vary.

Toy Example

Name	Math	Science	English
Alice	85	78	92
Bob	90	88	85
Charlie	78	72	80
David	65	70	75
Emma	95	92	89

- ❏ Dataset: Student scores on 3 subjects
- ❏ No missing data 👌

Toy Example: Get Statistics

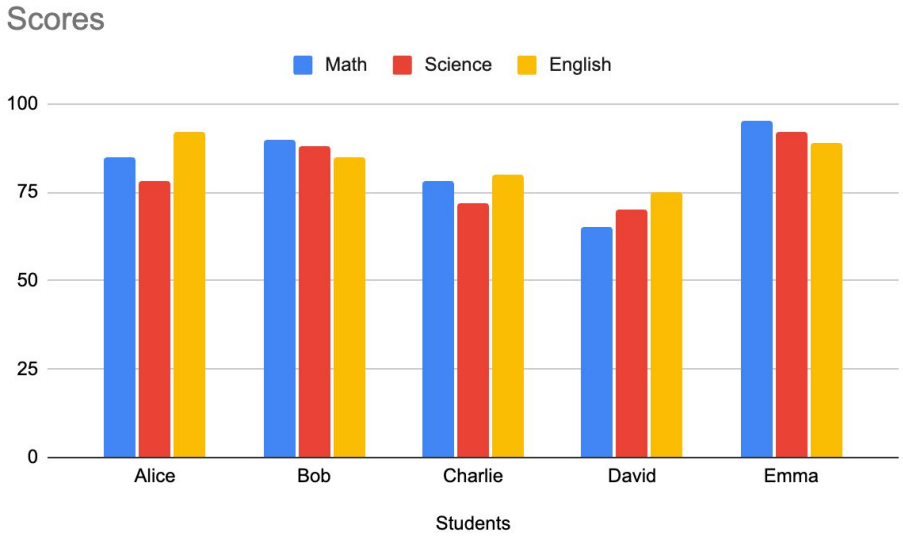
Name	Math	Science	English
Alice	85	78	92
Bob	90	88	85
Charlie	78	72	80
David	65	70	75
Emma	95	92	89
AVERAGE	**82.6**	**80**	**84.2**

Toy Example: Detect Outliers

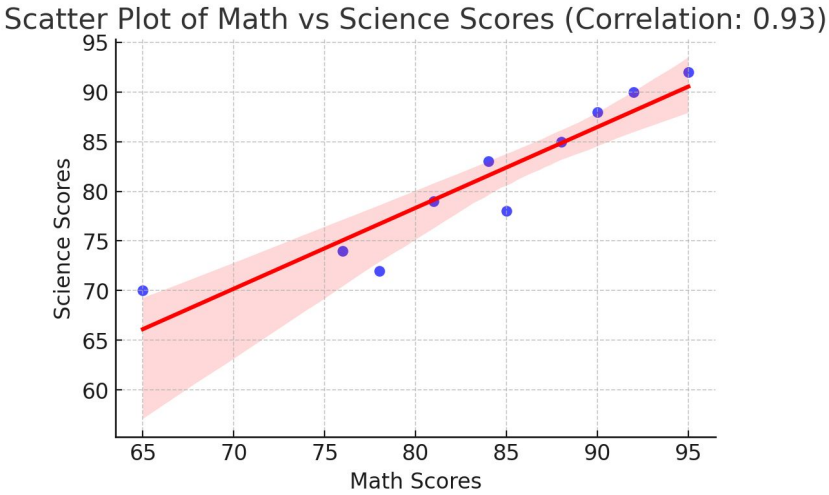
Name	Math	Science	English
Alice	85	78	92
Bob	90	88	85
Charlie	78	72	80
David	65	70	75
Emma	95	92	89
AVERAGE	**82.6**	**80**	**84.2**

- If a score is too low or too high compared to others, it can be flagged.

Toy Example: Explore Data Relationships



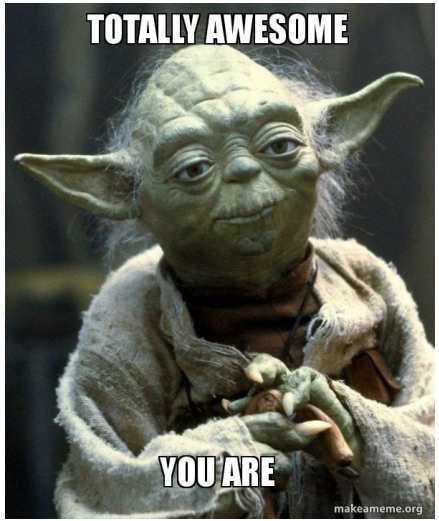
Use bar charts to compare subject scores.



- There's a strong positive correlation between math and science scores.
- Students who score higher in math also tend to perform well in science.

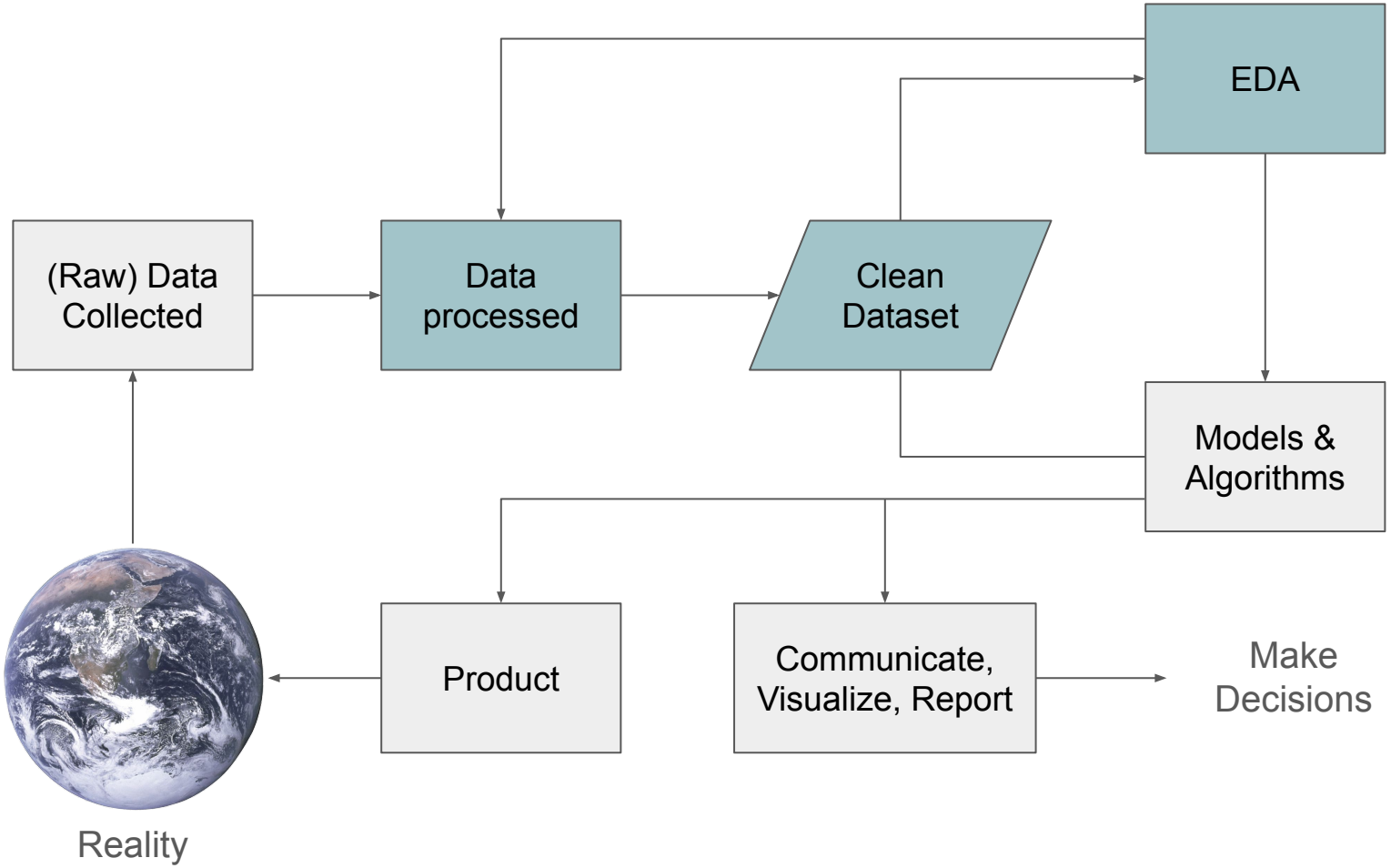
Toy Example: Draw Insights

- ➔ Students with high math scores tend to have high science scores.
- ➔ David has the lowest scores in all subjects and may need extra support.



Ref: https://makeameme.org/meme/totally-awesome-you

TOTALLY AWESOME, YOU ARE.


EDA
(Raw) Data Collected
Data processed
Clean Dataset
Models & Algorithms
Product
Communicate, Visualize, Report
Make Decisions
Reality

In Summary

- Intro to EDA
- Purpose of EDA
- Some steps
- Toy Example

Ref: https://makeameme.org/meme/when-class-is-21b23e18e8