COMP 333 — Week 8 Exploratory Data Analysis

Context: The Data Analytics Process

The data analytics process is how the business community looks at data analytics.

Step 1: Business Understanding: What are the business goals and problems?

Step 2: Data Understanding: Explore and visualize the data.

Step 3: Data Preparation: Generate features

Step 4: Modeling: Create models

Step 5: Evaluating: Train models and evaluate effectiveness

Step 6: Deploying: Use this data-driven approach for the goal of the business on a regular basis.

This can be viewed as a highly iterative cycle:

- Define the Goal: What problem are you solving?
- Collect and Manage Data: What information do I need?
- Build the Model: Find patterns in the data that lead to solutions.
- Evaluate and Critique the Model: Does the model solve your problem?
- Present Results and Document: Establish that you can solve the problem, and how.
- Deploy Model: Deploy the model to solve the problem in the real world.

What is Data Analytics?

The aim of data analytics is to add value to your data

so it becomes **actionable** data

which means it helps you and your organisation to make decisions.

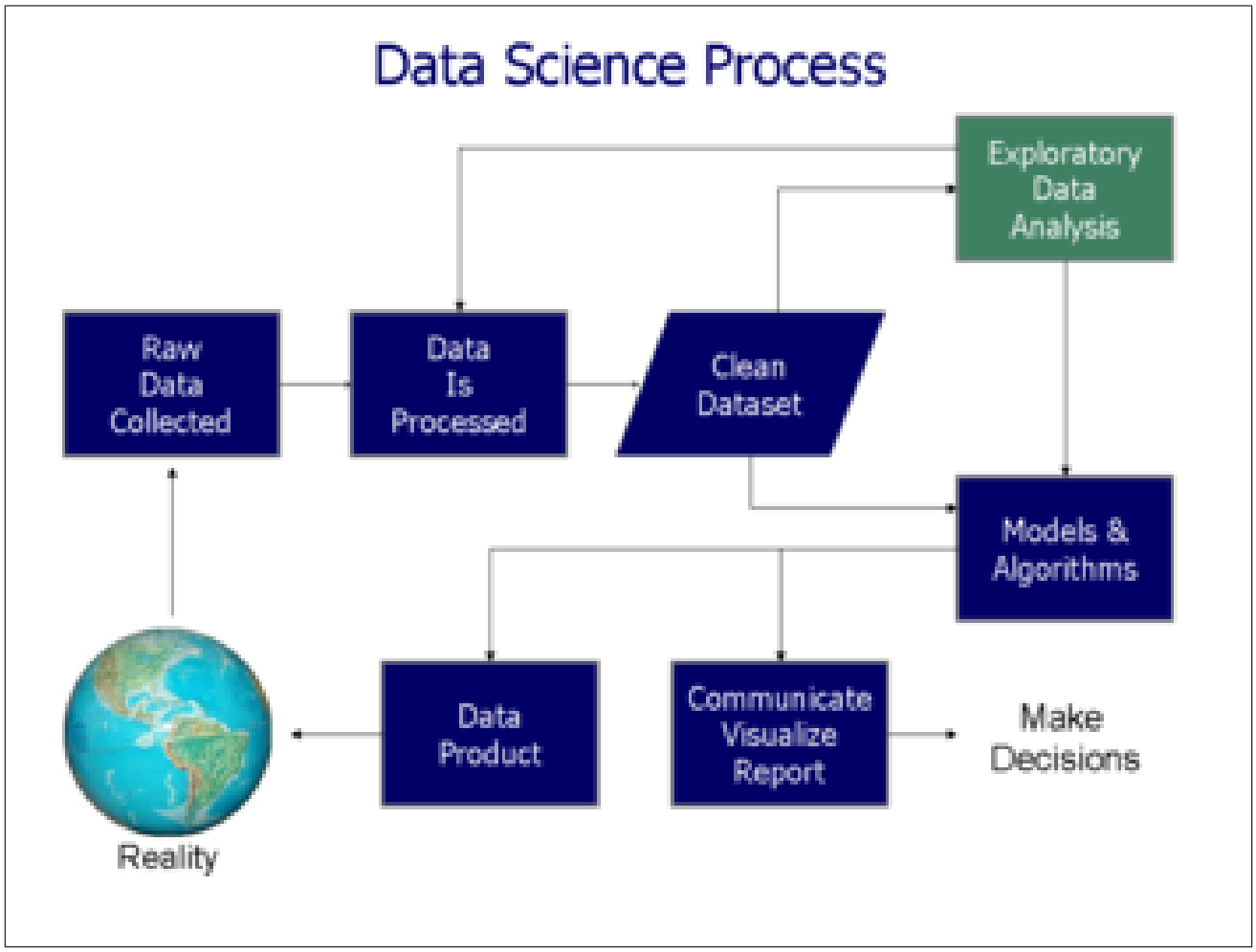
You will see it termed as *"monetization of data"* in the business world.

The main steps of the data analytics are

- descriptive data analysis
- data wrangling
- exploratory data analysis

These steps fit into an overall data analytics process
where you combine an understanding of the data and the business
to come up with data-driven input into the decision-making of the organization.

The wikipedia view of data science shows where EDA fits in the process



Exploratory Data Analysis (EDA)

EDA grew out of the statistics community.

EDA is the heart of data analytics.

EDA involves data wrangling and descriptive data analysis.

EDA develops a data-driven solution to your problem
by exploring the data to find which features lead to a solution.

The steps of EDA

Step 1: Data wrangling: collect, load, enrich data

Step 2: Descriptive data analysis: check data types, check distributions

Step 3: Feature engineering

Step 4: Modeling

Step 5: Story-Telling

A checklist for EDA:

Q1. What question(s) are you trying to solve (or prove wrong)?

Q2. What kind of data do you have and how do you treat different types?

Q3. What's missing from the data and how do you deal with it?

Q4. Where are the outliers and why should you care about them?

Q5. How can you add, change or remove features to get more out of your data?

Statistics View of EDA

Statistics moved from the structured *confirmatory data analysis*
behind the scientific method
to a freer data-driven approach, called Exploratory Data Analysis
through the work of John Tukey:
John Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.

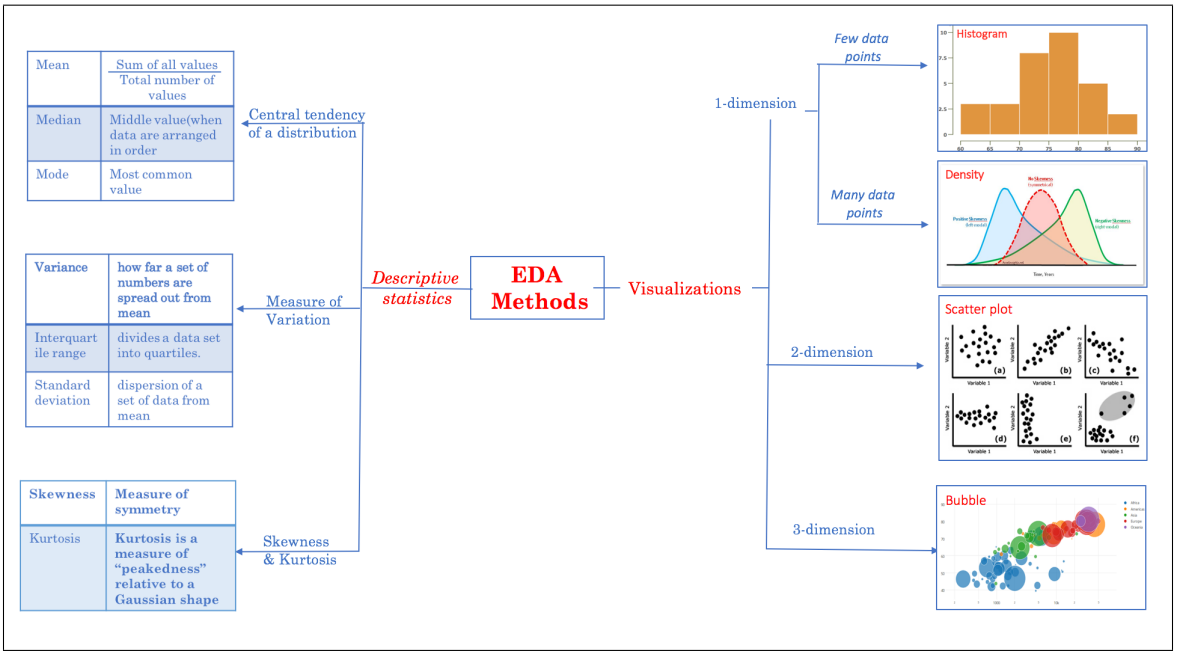
Definition NIST Engineering Statistics Handbook
EDA is an approach to data analysis
that postpones the usual assumptions about what kind of model the data follow
with the more direct approach of
allowing the data itself
to reveal its underlying structure and model.

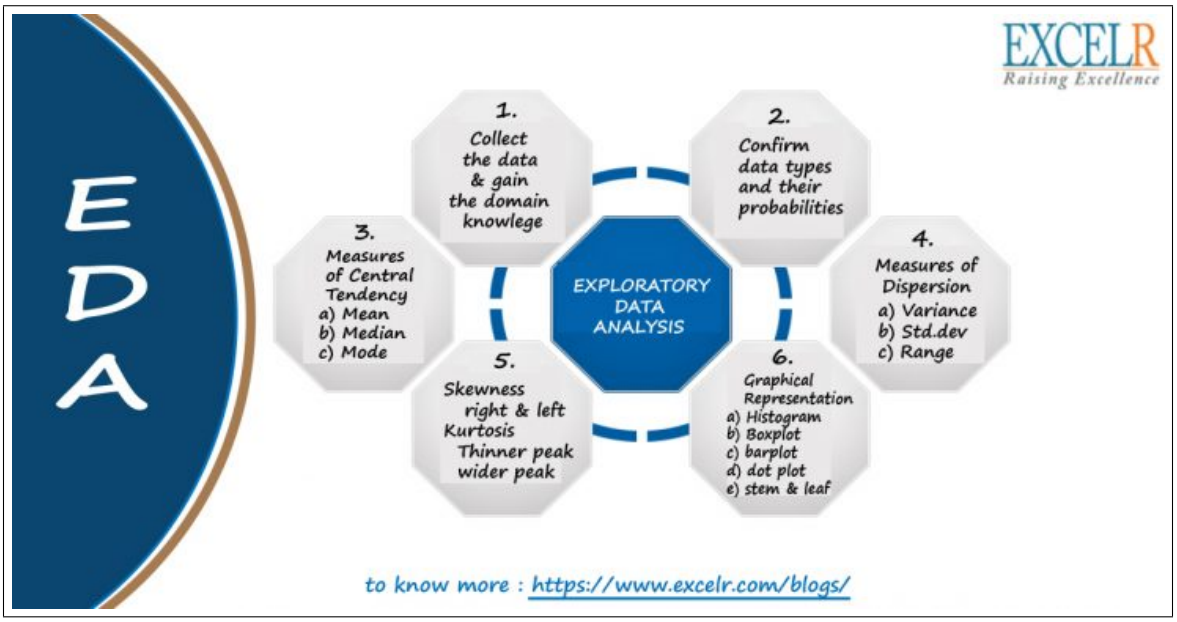
Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis
that employs a variety of techniques (mostly graphical) to

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious models; and
7. determine optimal factor settings.

The EDA approach is not a set of techniques,
but an attitude/philosophy
about how a data analysis should be carried out.

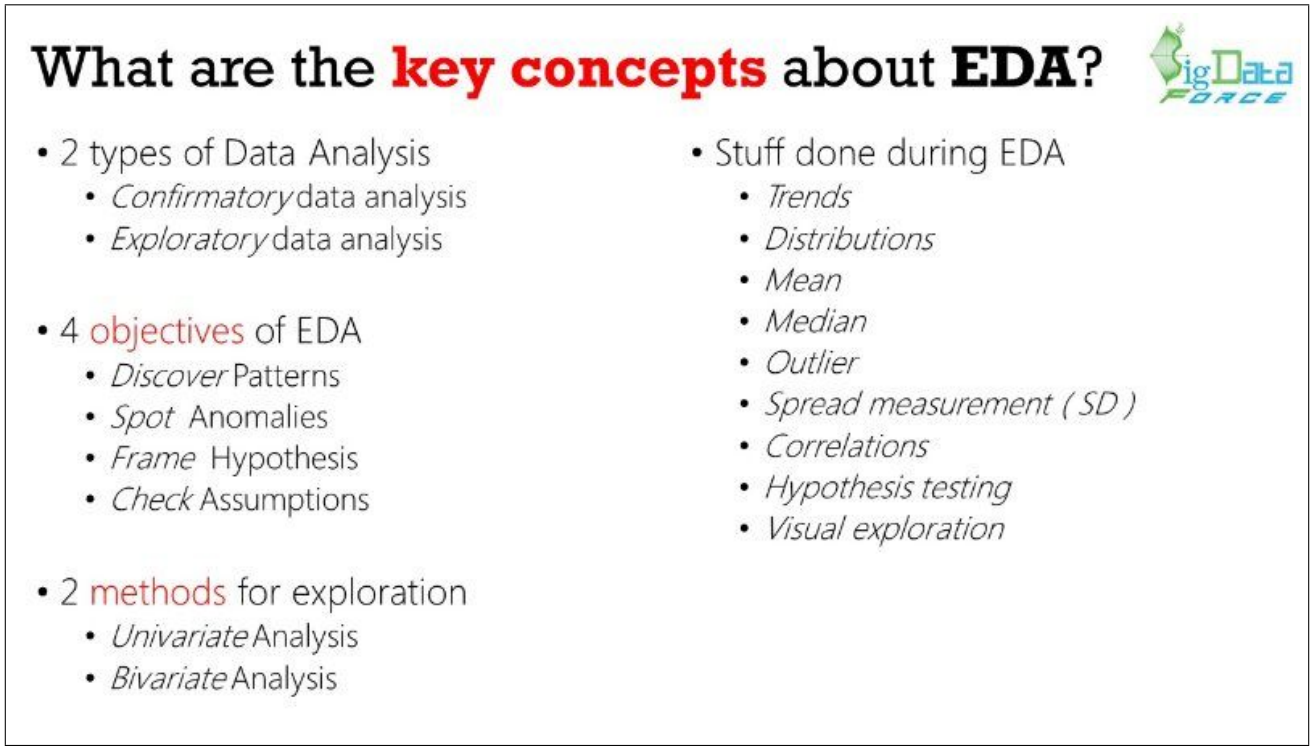
In many ways, the statistics view is sometimes little more than DDA:





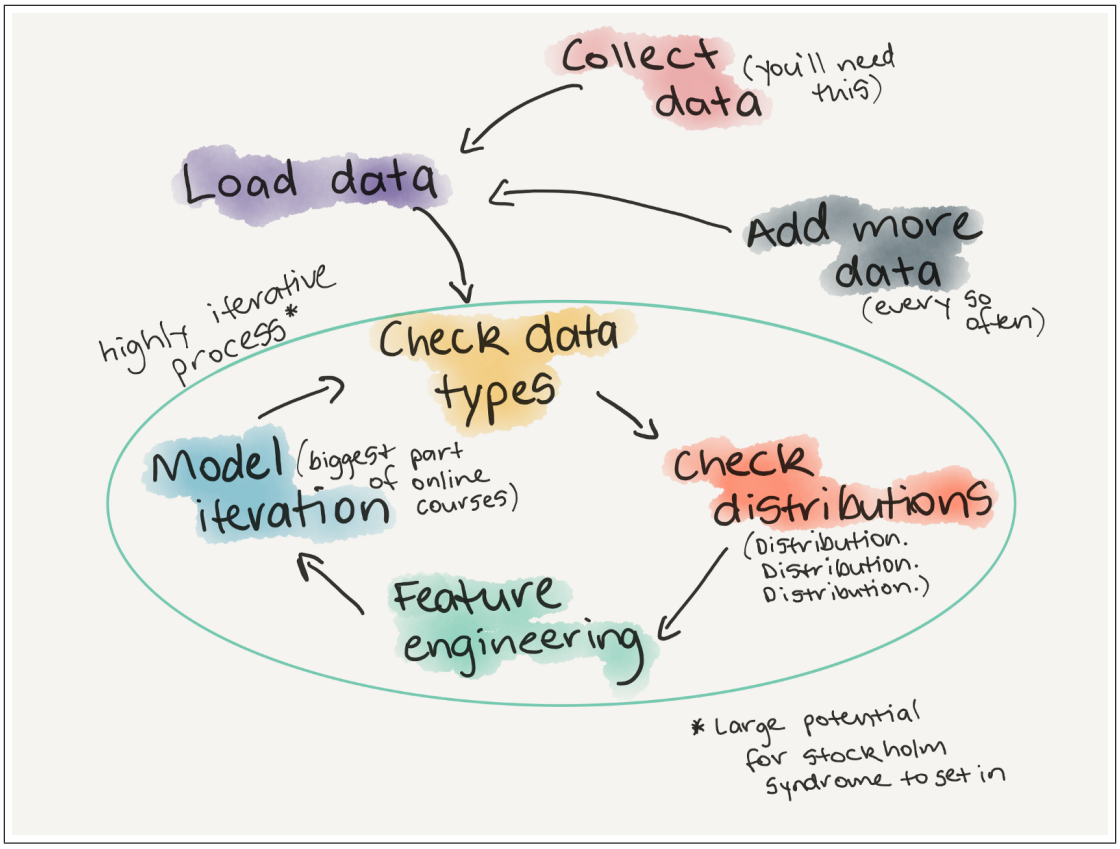
Or DDA slightly extended to include parts of model building

such as frame an hypothesis, check assumptions, and test hypothesis



Modern View of EDA

Today's view of EDA has been built around machine learning model construction:



Daniel Bourke, A Gentle Introduction to Exploratory Data Analysis, `https://towardsdatascience.com/a-gentle-introduction-to-exploratory-data-analysis-f11d843b8184`

Data Wrangling is evident is

Collect Data

Load Data

Enrich Data

DDA is evident in

Check Data Types

Check Distributions

What is novel in the modern view are the steps of

Feature Engineering

Modeling

Advanced Descriptive Methods

The EDA step of model construction involves Feature Engineering and both steps require a great deal of insight into your data.

Beyond the descriptions we have discussed under DDA, it is common to see descriptive techniques for

- Fitting curves and distributions
- Dimension reduction to simplify high-dimensional data
- Clustering of observations

The most common technique for *dimension reduction* is PCA (Principal Component Analysis).