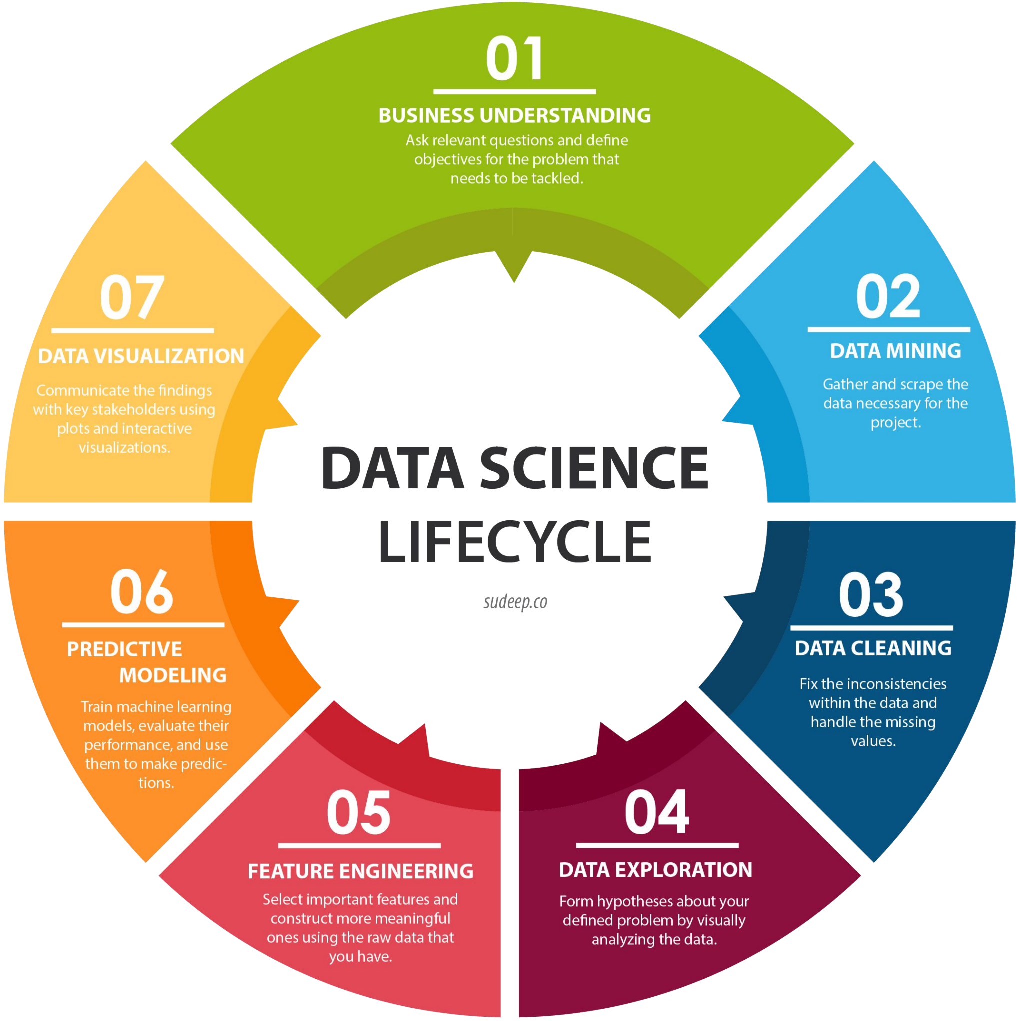
Exploratory Data Analysis and visualisation.

By
Ketul Gupta
Software Engineer.

The Data science life cycle.



DATA SCIENCE LIFECYCLE

sudeep.co

01 BUSINESS UNDERSTANDING

Ask relevant questions and define objectives for the problem that needs to be tackled.

02 DATA MINING

Gather and scrape the data necessary for the project.

03 DATA CLEANING

Fix the inconsistencies within the data and handle the missing values.

04 DATA EXPLORATION

Form hypotheses about your defined problem by visually analyzing the data.

05 FEATURE ENGINEERING

Select important features and construct more meaningful ones using the raw data that you have.

06 PREDICTIVE MODELING

Train machine learning models, evaluate their performance, and use them to make predictions.

07 DATA VISUALIZATION

Communicate the findings with key stakeholders using plots and interactive visualizations.

EDA=FIRST LOOK AT DATA!

Introduction

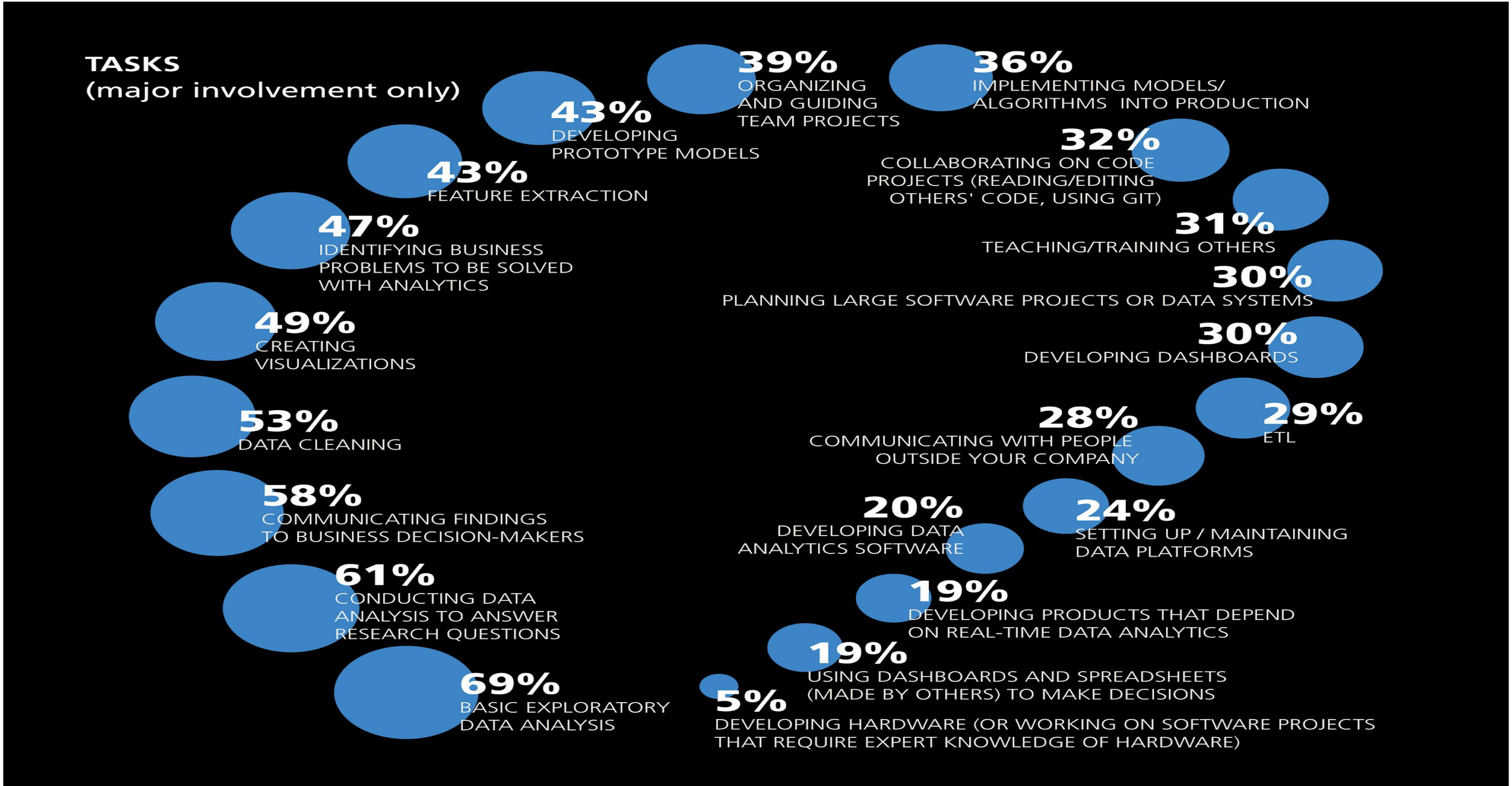
- Exploratory Data Analysis (EDA) and Visualization are important (necessary?) steps in any analysis task.
- get to know your data!
  - distributions (symmetric, normal, skewed)
  - data quality problems
  - outliers
  - correlations and inter-relationships
  - subsets of interest
  - suggest functional relationships
- Sometimes EDA or viz might be the goal!

Why EDA?

- Goal: get a general sense of the data
  means, medians, quantiles, histograms, boxplots
- You should always look at every variable - you will learn something!
- data-driven (model-free)
- ***Think interactive and visual***
  - Humans are the best pattern recognizers
  - You can use more than 2 dimensions!
    x,y,z, space, color, time….
- especially useful in early stages of data mining
  detect outliers (e.g. assess data quality)
- test assumptions (e.g. normal distributions or skewed?)
- identify useful raw data & transforms (e.g. log(x))

Bottom line:
It's always well worth looking at your DATA!

What data scientists spend the most time doing?



TASKS
(major involvement only)

69% BASIC EXPLORATORY DATA ANALYSIS

61% CONDUCTING DATA ANALYSIS TO ANSWER RESEARCH QUESTIONS

58% COMMUNICATING FINDINGS TO BUSINESS DECISION-MAKERS

53% DATA CLEANING

49% CREATING VISUALIZATIONS

47% IDENTIFYING BUSINESS PROBLEMS TO BE SOLVED WITH ANALYTICS

43% FEATURE EXTRACTION

43% DEVELOPING PROTOTYPE MODELS

39% ORGANIZING AND GUIDING TEAM PROJECTS

36% IMPLEMENTING MODELS/ ALGORITHMS INTO PRODUCTION

32% COLLABORATING ON CODE PROJECTS (READING/EDITING OTHERS' CODE, USING GIT)

31% TEACHING/TRAINING OTHERS

30% PLANNING LARGE SOFTWARE PROJECTS OR DATA SYSTEMS

30% DEVELOPING DASHBOARDS

29% ETL

28% COMMUNICATING WITH PEOPLE OUTSIDE YOUR COMPANY

24% SETTING UP / MAINTAINING DATA PLATFORMS

20% DEVELOPING DATA ANALYTICS SOFTWARE

19% DEVELOPING PRODUCTS THAT DEPEND ON REAL-TIME DATA ANALYTICS

19% USING DASHBOARDS AND SPREADSHEETS (MADE BY OTHERS) TO MAKE DECISIONS

5% DEVELOPING HARDWARE (OR WORKING ON SOFTWARE PROJECTS THAT REQUIRE EXPERT KNOWLEDGE OF HARDWARE)

Methods in EDA

Exploratory Data Analysis is majorly performed using the following methods:

- Univariate visualization – provides summary statistics for each field in the raw data set.
- Bivariate visualization – is performed to find the relationship between each variable in the dataset and the target variable of interest.
- Multivariate visualization – is performed to understand interactions between different fields in the dataset.
- Dimensionality reduction – helps to understand the fields in the data that account for the most variance between observations and allow for the processing of a reduced volume of data.

Examples.

Visualization

Various visualisations

- For numerical data:
- For categorical data:
- For correlation between various attributes:
- Dimensionality reduction:
- Univariate analysis
- Bivariate analysis
- Multivariate analysis

Thank you!