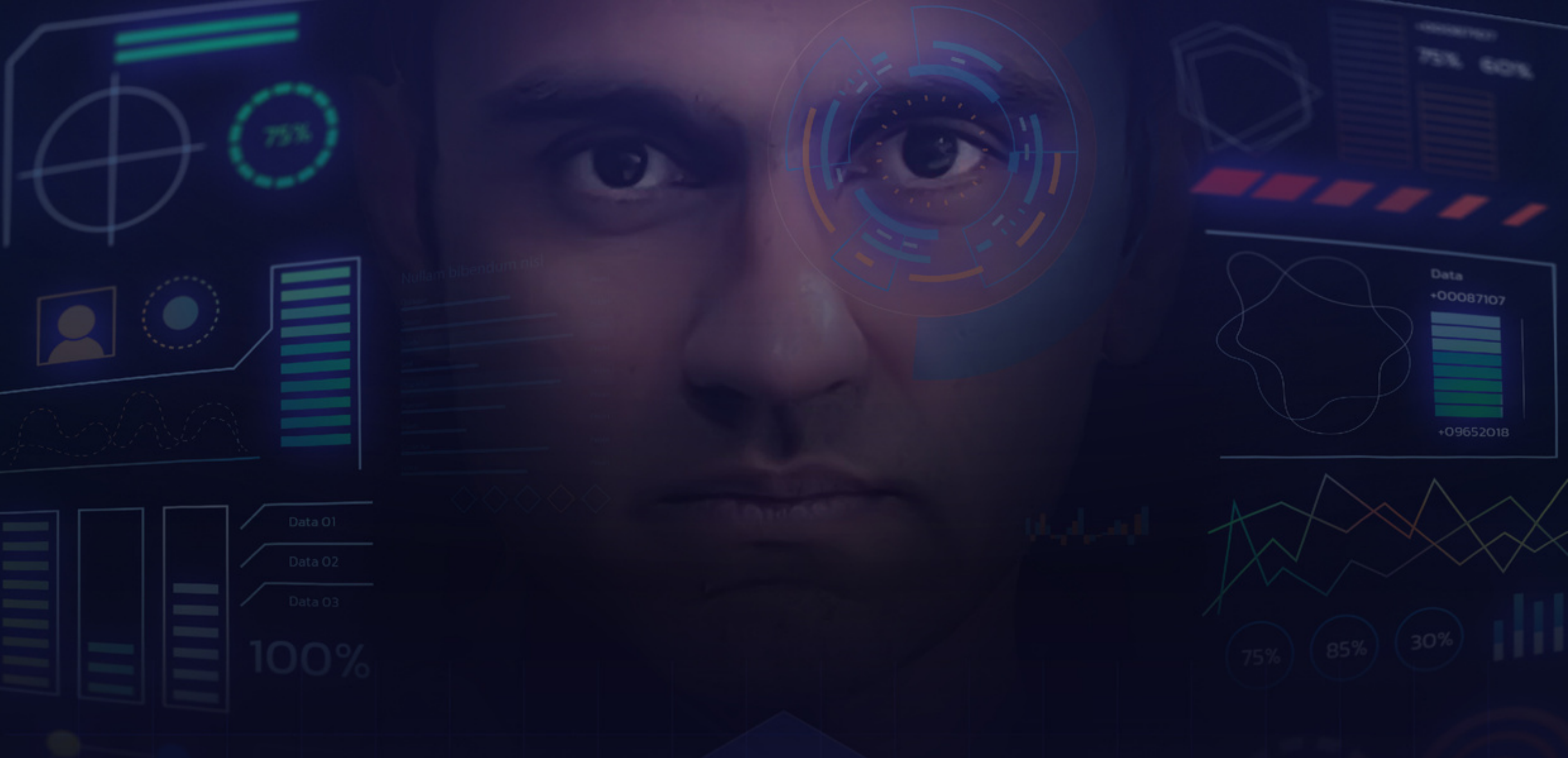

CODE
BASICS

Data & AI

TAKEAWAYS

codebasics.io

The Role of Data in AI

1 Data is the integral part of any AI application – without good data the AI model is as good as nothing.

2 Data is majorly classified into two types – Structured data and unstructured data.

3 With respect to data collection, data is classified again into two types namely manual and automated data.

4 Data ins transported, transformed and stored before usage just like the fuel before consumption.

Data Infrastructure in a Company

1 Data infrastructure is generally created by 4 personas in any company: Requestor, Creator, Regulator and Consumer.

2 Data collection, transformation, storage and distribution are the 4 stages involved in a typical company building data infrastructure for AI applications.

3 Data Engineering team (creator) is generally involved in all stages of generating data infrastructure.

4 Data governance team (regulator) is generally involved in the data transformation and data distribution stage.

Data Collection: Overview

1 Companies collect data by two major means – one is organisational data which is collected from their own software, the other one is external data which is bought from external vendors.

2 APIs and Web scraping are two popular methods of getting data from a website.

Data Storage and Transformation: Overview

1 ETL process is mostly used in companies where they process the data in batches as the insights required are not real time whereas ELT is used in incoming data is instantaneous and the insights required are real time (Netflix, Amazon etc.)

2 APIs and Web scraping are two popular methods of getting data from a website.

3 Volume, Velocity and Variety are the 3 factors to be considered for selecting the right database.

4 Apache Spark enables transformation of huge data through distributed computing.

5 Master data management is typically performed by the data governance team.

Data Distribution: Privacy, Ethics & Governance

1 Data privacy, compliance and governance is important to protect the misuse of data which has the potential to create a big negative impact.

2 In 2024, EU union released EU AI act which regulates the usage of AI and imposes heavy fines for non-compliance.

3 Masking and anonymisation is a method to protect the identity of individuals by hiding their PII (Personally Identifiable Information).

4 Mandatory Compliance Training is one of the ways to ensure data compliance in the organization.