

The background is a dark blue gradient. In the corners, there are white line-art illustrations of circuit boards or data paths, consisting of lines and small circles.

# DATA GATHERING AND CLEANING

## CHAPTER 5

# THERE ARE FIVE MAIN STEPS FOR DATA SCIENCE PROCESSING

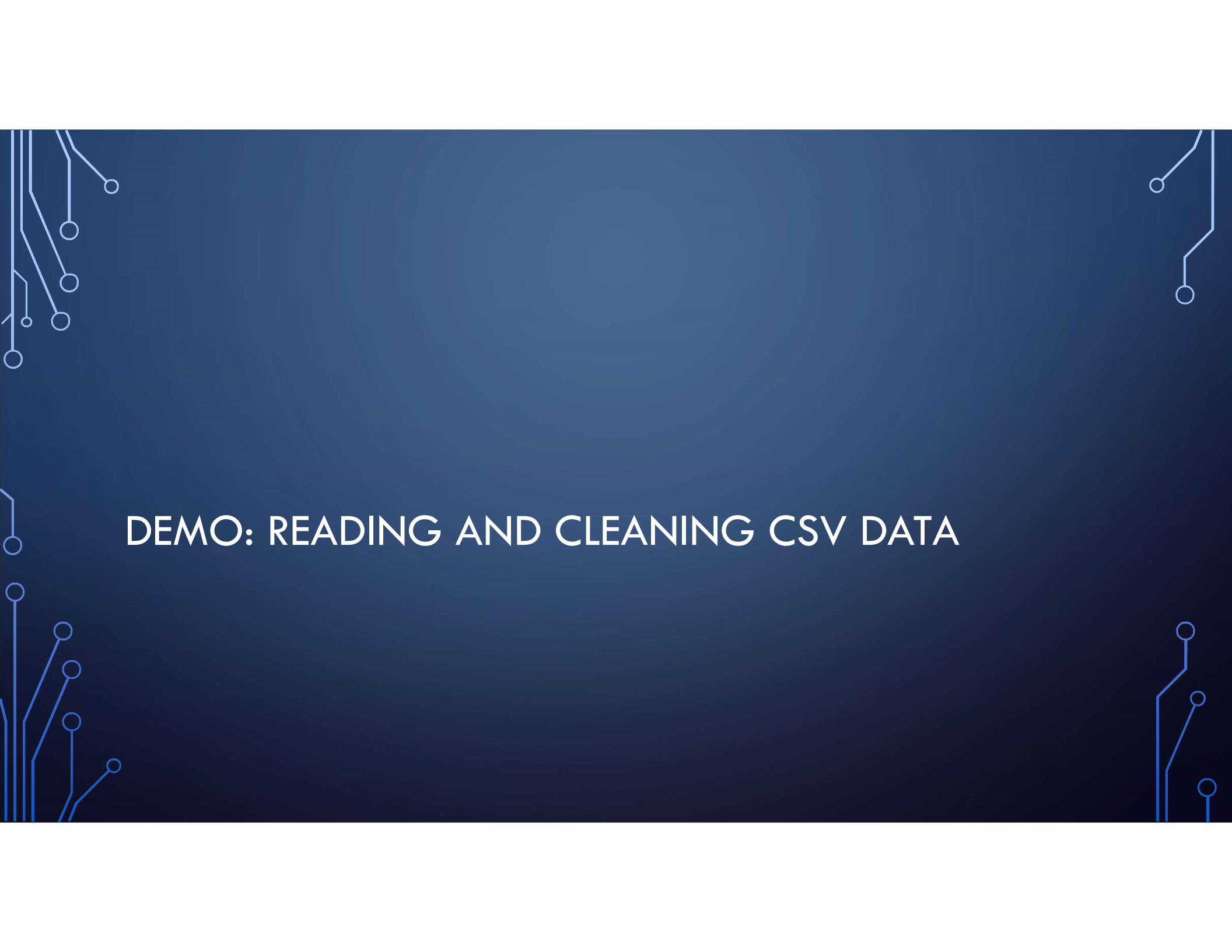
- **Data acquisition**
- **Data cleaning**
- **Exploratory analysis** is where you look at your cleaned data and make statistical processing fits for specific analysis purposes.
- An **analysis model** needs to be created. Advanced tools such as machine learning algorithms can be used in this step.
- **Data visualization** is where the results are plotted using various systems provided by Python to help in the decision-making process.

# PYTHON LIBRARIES

- **Pandas** is an open source Python library used to load, organize, manipulate, model, and analyze data by offering powerful data structures.
- **Numpy** is a Python package that stands for “numerical Python. It is a library consisting of multidimensional array objects and a collection of routines for manipulating arrays. It can be used to perform mathematical, logical, and linear algebra operations on arrays.
- **Matplotlib** is a Python library used to create 2D graphs and plots.

# CLEANING DATA

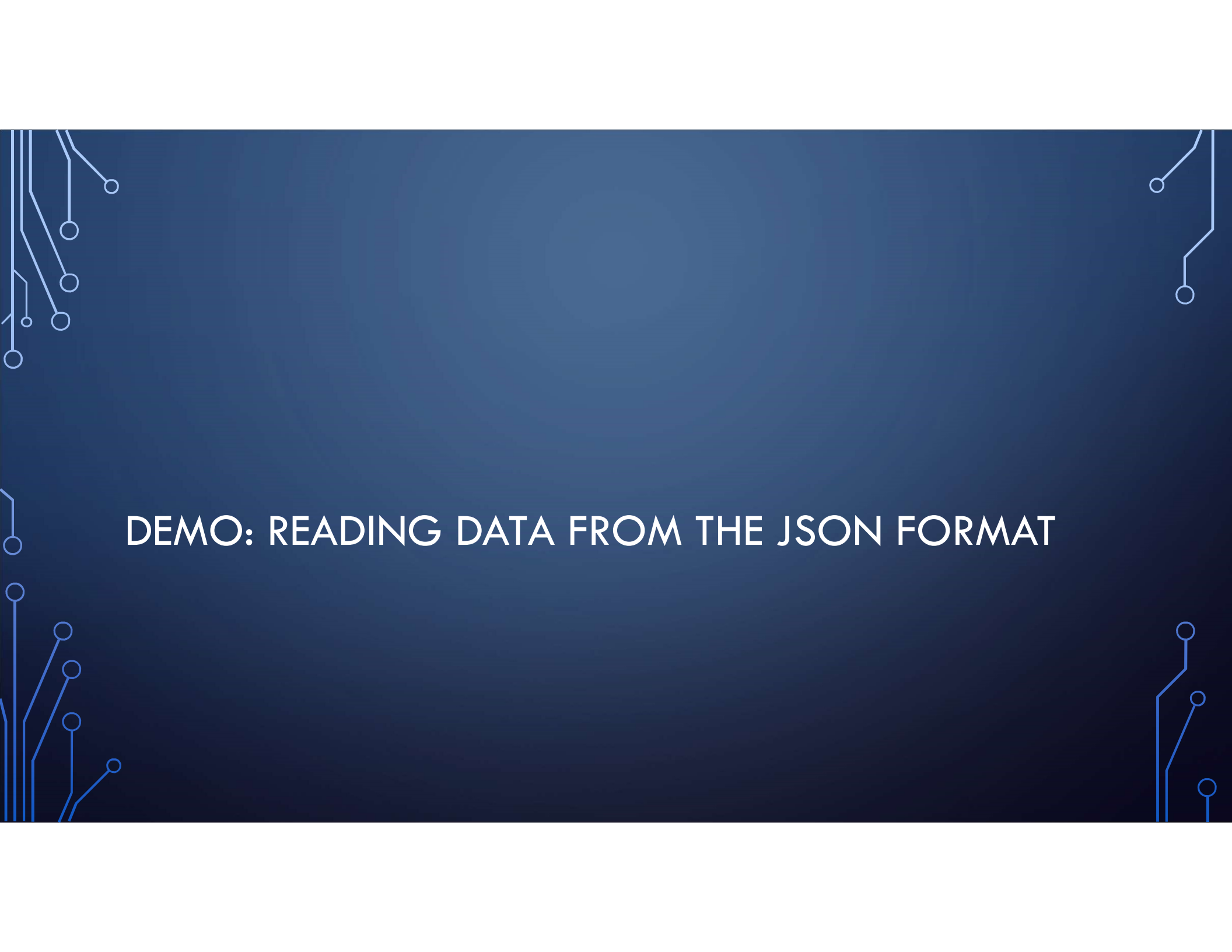
- Checking for Missing Values
- Handling the Missing Values

The background is a dark blue gradient. In the corners, there are white line-art illustrations of circuit boards or data paths, with lines and small circles representing components.

# DEMO: READING AND CLEANING CSV DATA

The background is a dark blue gradient. In the corners, there are white line-art illustrations of circuit boards or data paths, consisting of lines and small circles.

# DEMO: MERGING AND INTEGRATING DATA

The background is a dark blue gradient. In the corners, there are white line-art illustrations of circuit traces and nodes, resembling a printed circuit board (PCB) layout. These elements are positioned in the top-left, top-right, bottom-left, and bottom-right corners, framing the central text.

# DEMO: READING DATA FROM THE JSON FORMAT

The background is a dark blue gradient. In the corners, there are white line-art illustrations of circuit boards or neural networks, with lines and small circles representing nodes and connections.

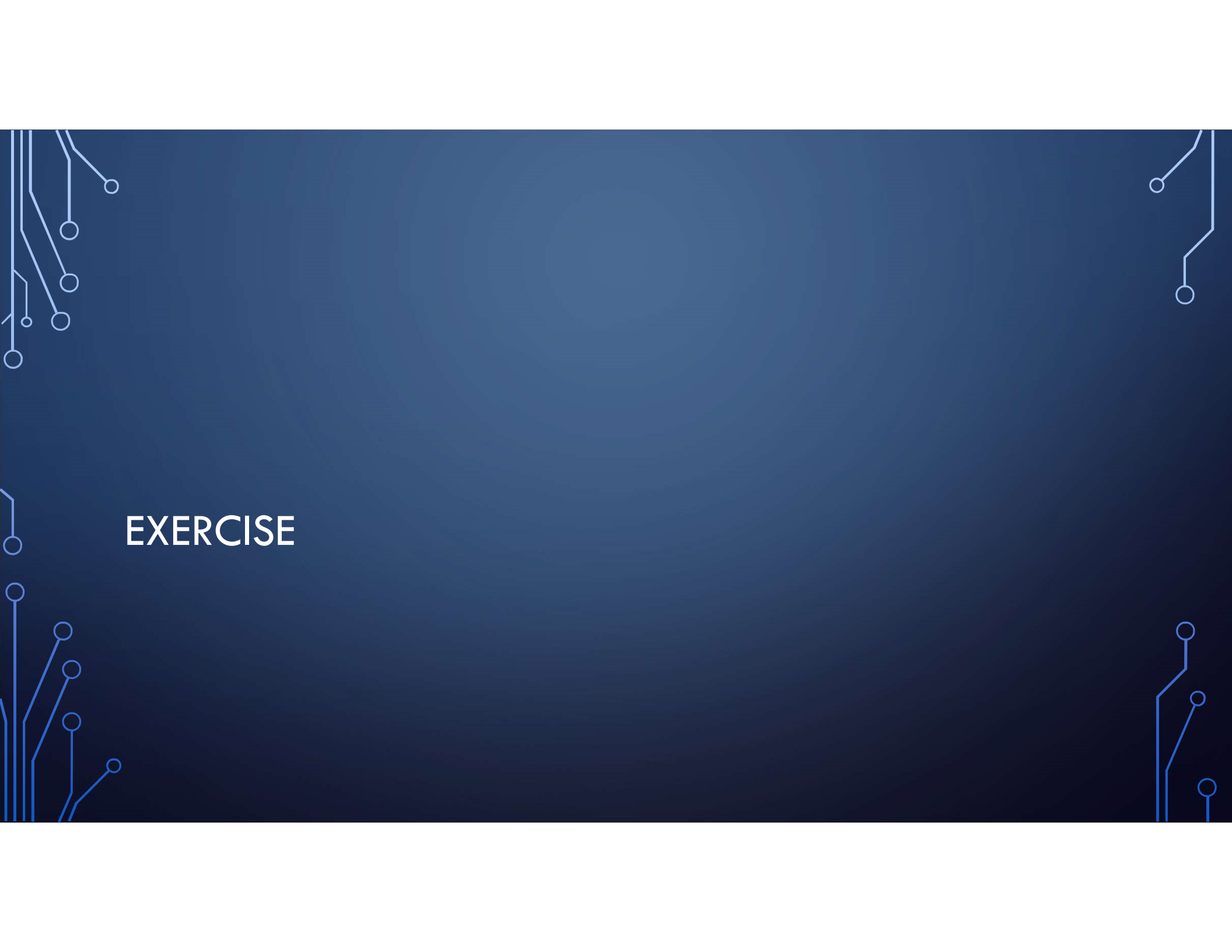
# DEMO: READING DATA FROM THE HTML FORMAT

The background is a dark blue gradient. In the corners, there are white line-art illustrations of circuit boards or neural networks, with lines and small circles representing nodes and connections.

# DEMO: READING DATA FROM THE XML FORMAT

# SUMMARY

- How to apply cleaning techniques to handle missing values
- How to read CSV-formatted data offline and directly from the cloud
- How to merge and integrate data from different sources
- How to read and extract data from JSON, HTML, and XML formats

The background is a dark blue gradient with a subtle, abstract pattern of light blue lines and circles, resembling a circuit board or a network diagram. The pattern is more dense on the left and right sides, with lines and circles extending from the edges towards the center.

# EXERCISE