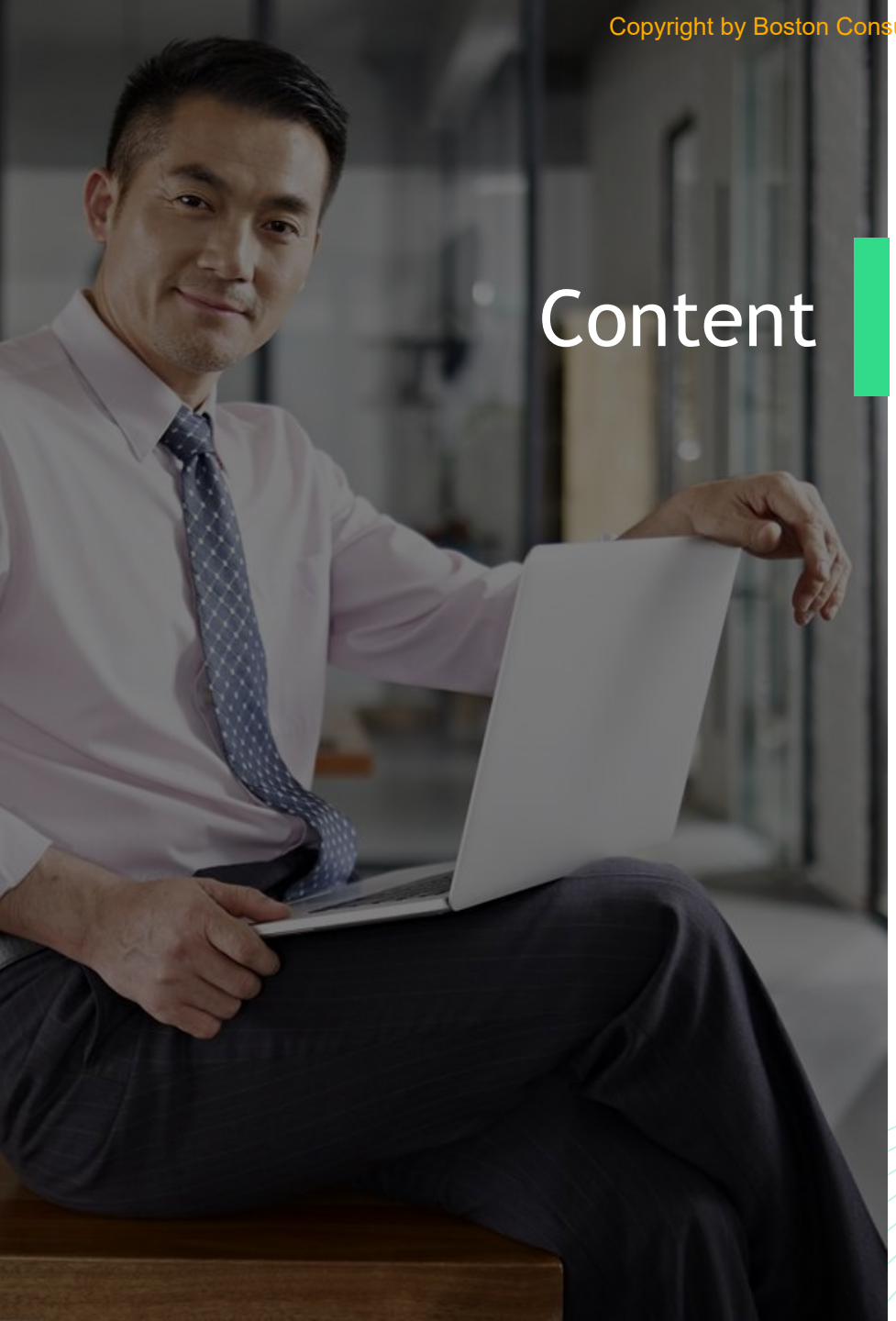
RISE BY DIGITALBCG ACADEMY

IU 4.5 – Statistical Analysis

Mini Project

2021

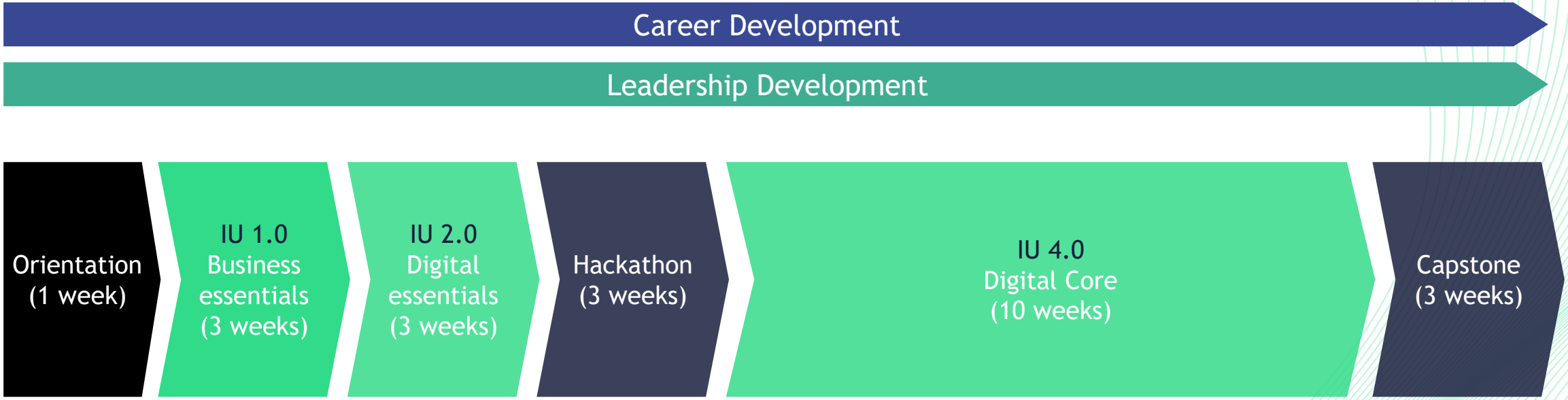


Content

180 min

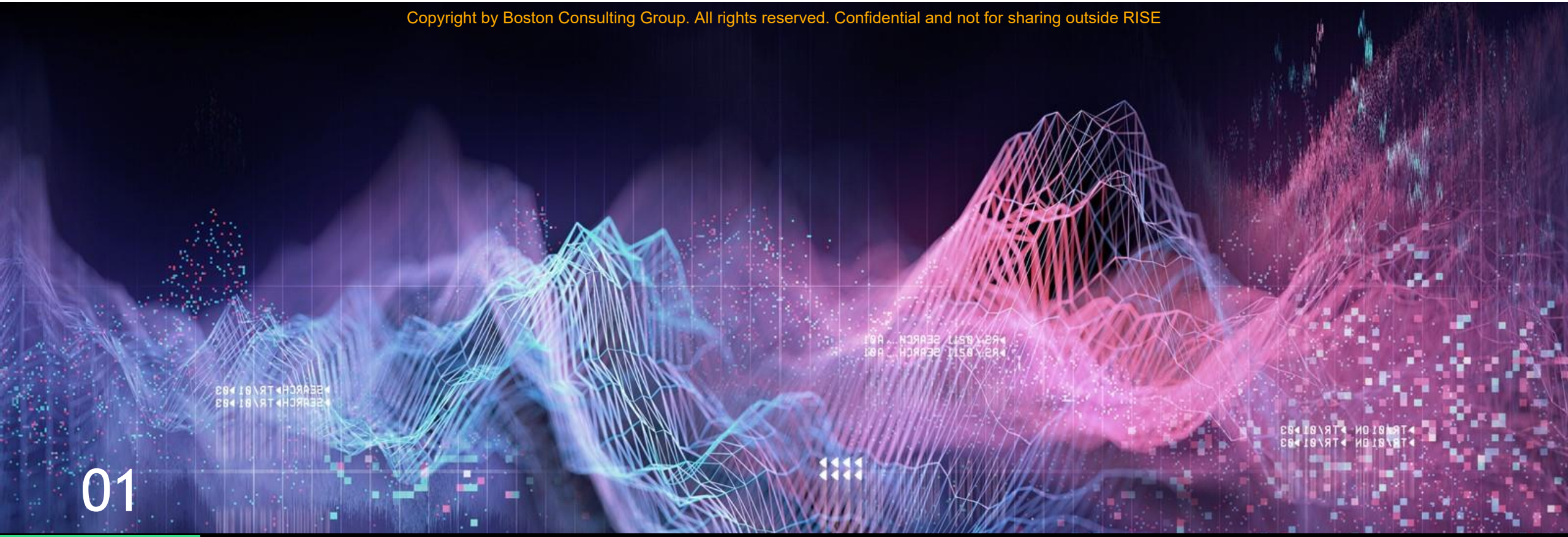
Where Are We in the Journey & Learning Objectives	5 min
Agenda for Today	30 min
• Mini Project: Introduction to problem statement	20 min
• Q&A	120 min
• Group activity to solve the project	
Wrap up	5 min

Where are we in the learning journey?



Learning objectives

- Generate statistical summary of data sets and perform exploratory data analysis
- Test of significance - statistical hypothesis testing (A/B-test) between two groups of data



01

Mini Project: Introduction to the problem statement

Introduction – Cookie Cats

Play and complete levels

Cookie Cats is a popular mobile puzzle game where players complete a task and level up

Encounter gates

While players completes levels, they encounter gates after completing certain number of levels

Gates serves the purpose of giving players enforced break and drive revenue from in-game purchases

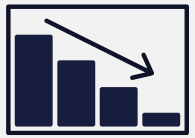
Wait at the gate

Gates force players to wait for sometime before they can play further or make in game purchases

Business Problem

Even though Cookie Cats overall popularity is growing over time with players subscribing to try out the game, but,

Revenue from in-game purchases has been declining over time



Total number of active players are also declining with players uninstalling the game after playing for few days

Project objectives

Business Objective: Increase YoY revenue from game purchases by increasing retention rate of gamers

Hypothesis

- Company CEO believes that players are churning because the first gate encountered at level 30 is too early which forces players to wait before they can proceed further in the game
- In order to increase player retention rate, developers ran AB-test by moving the first gate from level 30 to level 40 for some players
  - i.e., group A would encounter the gate at level 30
  - And group B would encounter the gate at level 40

Analytics Objective

Test CEO's hypothesis to analyze if moving the first gate from level 30 to 40 increases retention rate and number of game rounds played

What is provided?

Data provided for 90,189 players who installed the game while the AB-test was running

Variables	Variable description
User ID	a unique number that identifies each player
version	whether the player was put in the control group A (gate_30 - a gate at level 30) or the test group B (gate_40 - a gate at level 40)
sum_gamerounds	the number of game rounds played by the player during the first week after installation
retention_1	did the player come back and play 1 day after installing?
retention_7	did the player come back and play 7 days after installing?

Group activity | Assignment

Exercise
Work plan

1. Detect and resolve problems in the data (Missing value, Outliers, Unexpected value, etc.) *[20 -40 mins] [Marks: 10]*

2. Plot summary statistics and identify trends to answer basis business questions *[30-45 mins] [Marks: 15]*
   i. What is the overall 7-day retention rate of the game?
   ii. How many players never played the game after installing?
   iii. Does the number of users decrease as the level progresses highlighting the difficulty of the game

3. Generate cross tab for two player groups to understand the difference in the 1-day and 7-days retention rate & total number of game rounds played *[20-30 mins] [Marks: 20]*

4. Perform two-sample test for groups A and B to test statistical significance amongst the groups in the sum of game rounds played **i.e., if groups A and B are statistically different** *[90 -120 mins] [Marks: 35]*
   i. Check the assumptions of two sample test
      a. Normal distribution - Apply Shapiro test
      b. Homogeneity of variance – Apply Levene's Test
   ii. Apply the relevant two sample significance test method based on the results from test for normality and homogeneity

5. Based on significance testing results, if groups A and B are statistically different, which level has more advantage in terms of player retention? *[20-40 mins] [Marks: 20]*

6. Bonus question: Use bootstrap resampling to plot retention rate distribution for both groups to visualize effect of different version of the game on retention *[20 - 40 mins]*

Estimated time
to complete: 4 -5 hours

What are we giving you?

Mini Project 2 - IU 4.5 Statistical Analysis

Packages and setup

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import os

from scipy.stats import shapiro
import scipy.stats as stats

#parameter settings
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

Starter code provided in Jupyter notebook

Q4. Perform two-sample test for groups A and B to test statistical significance amongst the groups in the sum of game rounds played i.e., if groups A and B are statistically different

Initial data processing

In [5]:
```
#Define A/B groups for hypothesis testing
#user_df["version"] = np.where(user_df.version == "gate_30", "A", "B")
group_A=pd.DataFrame(user_df[user_df.version=="A"]['sum_gamerounds'])
group_B=pd.DataFrame(user_df[user_df.version=="B"]['sum_gamerounds'])
```

Q4.1 Shapiro test of Normality

In [6]:
```
#---------------------- Shapiro Test ----------------------
# NULL Hypothesis H0: Distribution is normal
# ALTERNATE Hypothesis H1: Distribution is not normal

#test for group_A
```

Recap: Statistical analysis topics

- Slicing and dicing data frames to generate summary statistics
  - Filtering data frames based on column(s)
  - splitting data into multiple data frame(s)
  - shape of data frame, counting rows and length
  - Adding/deleting column(s) to a data frame
  - Aggregations and merging data frames
  - Grouping, pivot table, cross tabs
  - Scatter plots, histogram, box plots, etc.

- Hypothesis testing
  - Null and alternate hypothesis
  - P-value and significant testing based on confidence intervals

- One sample t-test
  - One-tailed and two-tailed

- Two sample t-test
  - Assumptions – normality test and homogeneity of variances
  - Types of two sample t-test
    - z-test
    - t-test with equal variances
    - t-test with unequal variances

Guidelines on submitting mini project solution

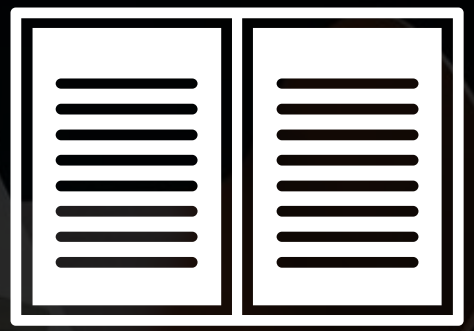
1) Work during the live class with your respective groups to brainstorm ideas and solve analytics objectives based on the work plan provided; Divide tasks within the group where possible

2) Submission guidelines – All team members must submit their solution via LMS using the Jupyter notebook template provided; One member from the group to submit the solution via Microsoft Teams as well, specifying the group number

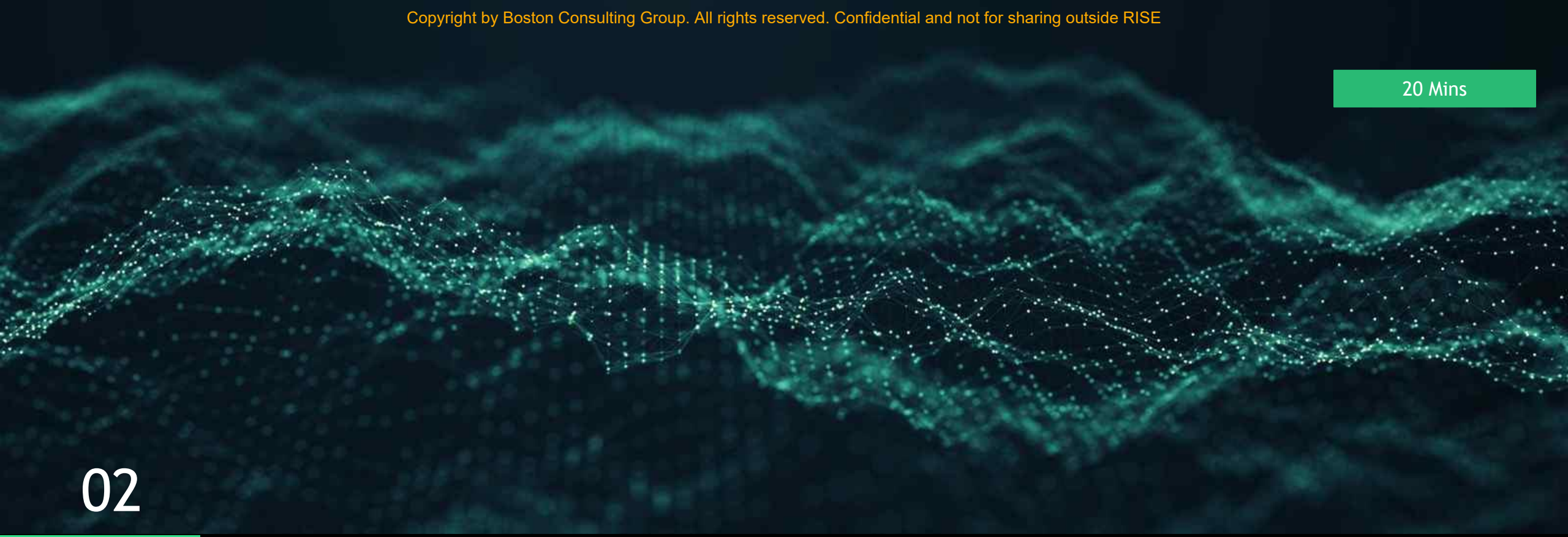
3) Final solution submission must be done during the final work and submission session

4) Grading will be done across - (i) Submission: 40% (ii) Concepts applied: 30% (based on marks per question) (iii) Insights/recommendation: 30% (based on marks per question)

5) Digital badges will be provided for top 2 teams with highest grades



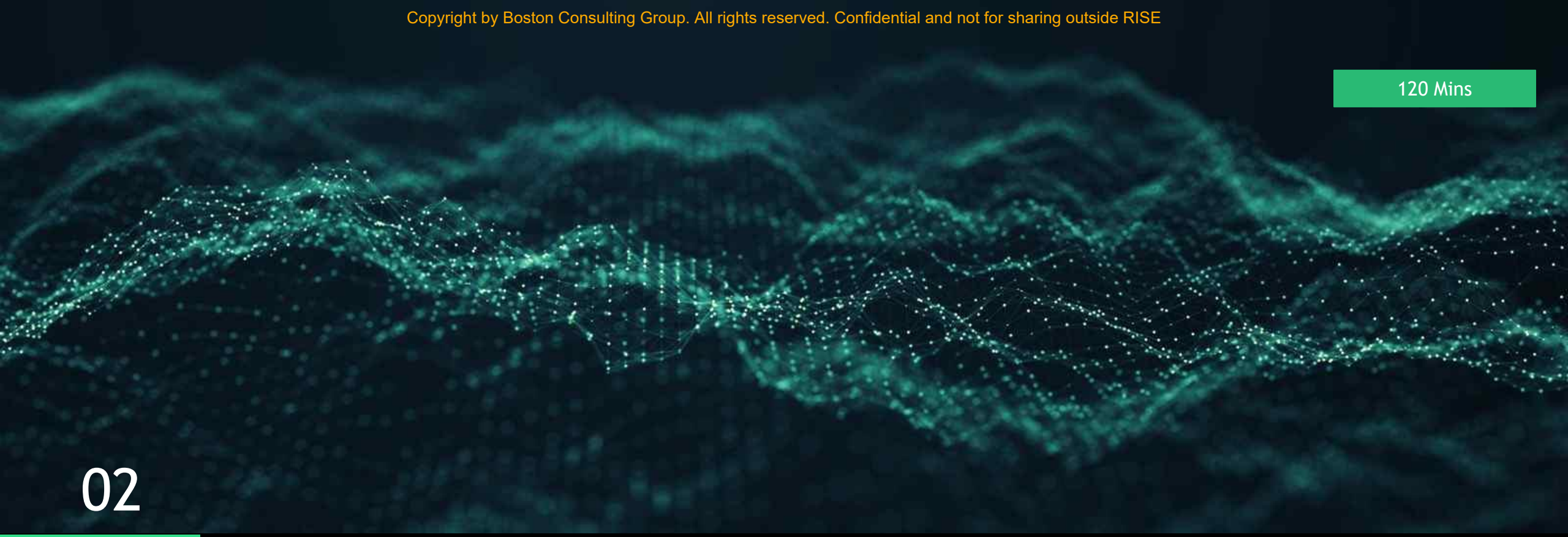
20 Mins



02

Questions?

120 Mins



02

Working session - Group activity

RISE BY DIGITAL BCG ACADEMY