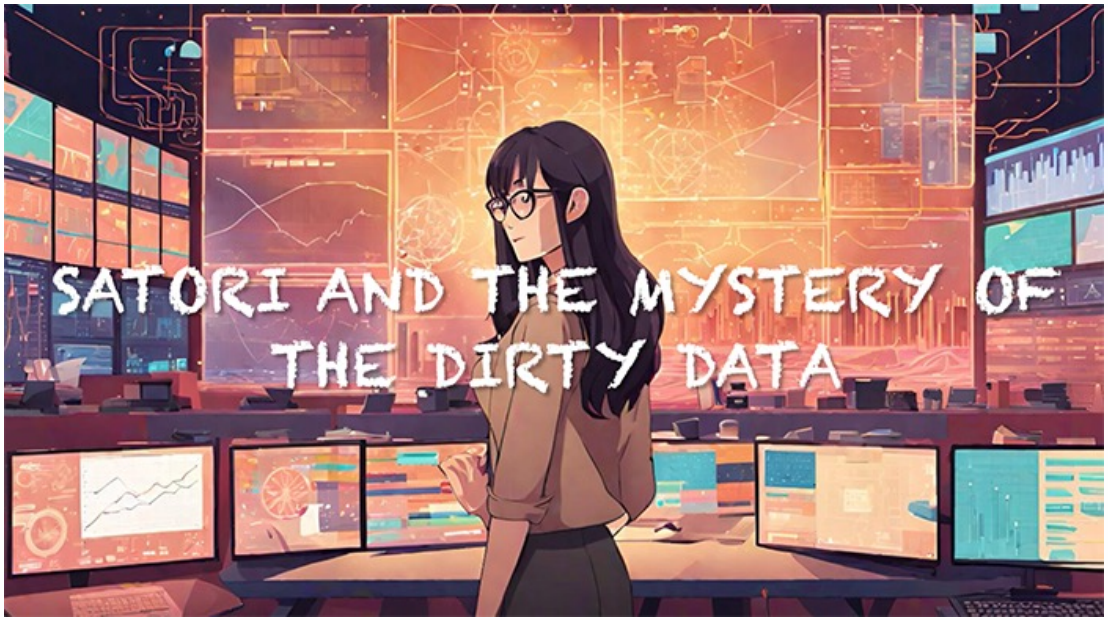
“Satori and the Mystery of the Dirty Data”

GCOOS Product Developer and Co-Data Manager Bob Currier created “Satori and the Mystery of the Dirty Data,” as a training aid to teach non-data scientists about the daily lives of data scientists. In the video, his character Satori takes viewers through the entire cycle of data processing — from finding and cleaning data, to creating and training artificial intelligence (AI) models, and finally, to visualizing the data. Here, we offer a quiz to test your new data skills!



GCOOS
GULF OF MEXICO
COASTAL OCEAN
OBSERVING SYSTEM

IOOS
Integrated Ocean
Observing System

1) What is the primary goal of Exploratory Data Analysis?

a. To turn the data into a beautiful piece of modern art.
b. To prepare a delicious data-based recipe for dinner.
c. To thoroughly understand and clean the dataset, identifying patterns and anomalies.
d. To find the best font for presenting the data in a report.

2) What is the primary goal of data cleaning?

a. To make data look shiny and new, like a freshly washed car.
b. To ensure data accuracy and completeness, making it reliable for analysis.
c. To make sure the data goblin doesn't have dirty underwear..
d. To confuse data scientists with complex and unsolvable puzzles.

3) What is the primary goal of a box plot in outlier detection?

a. To create a colorful graph that looks nice in presentations.
b. To play a game of statistical hide and seek, where outliers always lose.
c. To identify outliers by displaying how data is spread and where it is outside the norm .
d. To calculate the exact numerical value of each data point.

4) What is one-hot encoding used for in data handling?

a. Heating up your computer to keep your coffee warm.
b. Turning numerical data into a hot singles chart.
c. Converting categorical data into a binary format for easier processing by computers.
d. To find your perfect blind data-data.

5) Which of the following best describes a ‘key’ in merging?

a. A tool to unlock encrypted data.
b. A unique identifier used to match and merge data from different sources.
c. The main subject of a database.
d. The musical scale in a K-Pop video.

6) What is the purpose of sampling in data analysis?

a. To create backup of the entire dataset.
b. To reduce the size of the data for easier analysis.
c. To add more data to the existing dataset.
d. To fill up your stomach after you’ve skipped lunch.

7) Which visualization tool could Satori use to plot ocean observing data?

a. The bar graph, standing tall with each category.
b. The scatter plot, spreading dots at depth like phytoplankton in the ocean.
c. The pie chart, slicing data into perfect pieces.
d. The line graph, drawing the highs and lows of data.

8) Which tool did Satori use for creating interactive maps?

a. Web Sockets, for that real-time magic touch.
b. Leaflet, leading the way through data forests.
c. Folium, weaving Python spells into maps.
d. Google Maps.

9) What did Satori use to reveal the secret dance of variables?

a. Ice cream and sunburn charts.
b. Correlation matrices.
c. Scatter plots and heat maps.
d. Her microscope.

10) What is essential for choosing the right model framework?

a. Selecting the fanciest algorithm, because fancy is always better.
b. Picking the right tools and libraries, like sci-kit learn, TensorFlow and PyTorch.
c. Adding as many new spices (algorithms) as possible, without tasting the data.
d. Ensuring the data is as messy as possible to give the model a challenge.

11) Which of the following best describes feature engineering?

a. Using domain knowledge to transform raw data into meaningful feature.
b. Using any old data, regardless of the quality or relevance of features.
c. An automated process, where data scientists drink coffee and gossip while waiting.
d. The final step in the data science workflow, performed after training and evaluation.

12) Which of the following describes training and testing data?

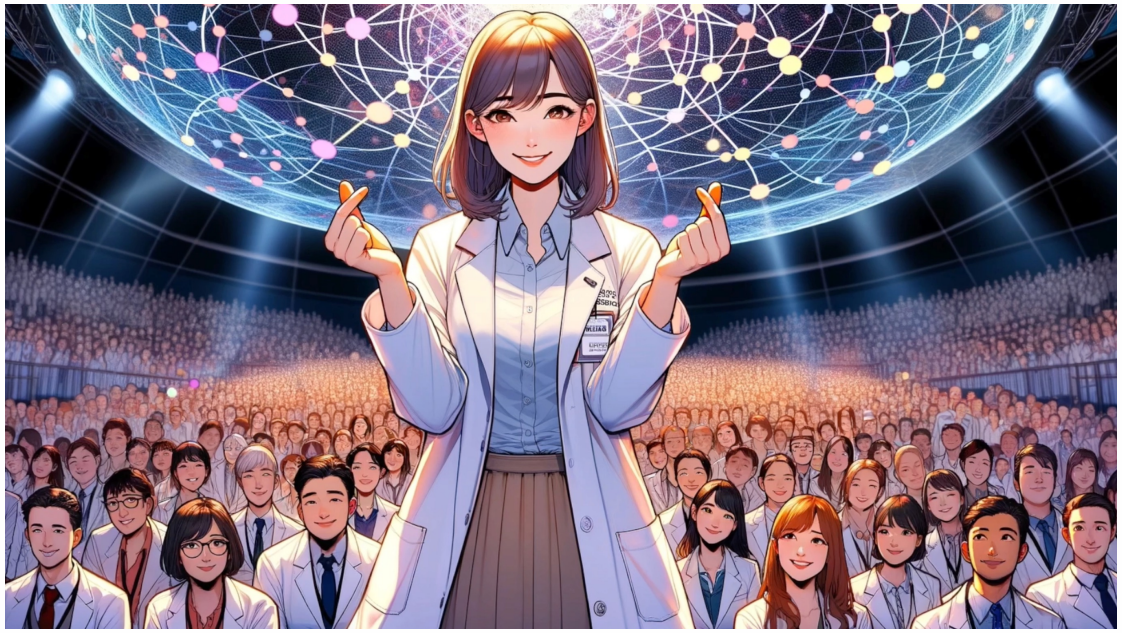
a. Training data is used to teach the model, testing data is to show it off.
b. Training data is for parameter tuning, testing data is for fun.
c. Training data is for fitting the model, testing data is for testing on unseen data.
d. Training data is to confuse the model, testing data is clear things up.

13) What is NOT a common method for assessing model performance?

a. Precision and recall.
b. Overfitting.
c. The F1 score.
d. Mean Squared Error.

14) Which aspect is crucial in the realm of data driven decision making?

a. The importance of having a large data set.
b. The role of ethical considerations and responsible data use.
c. The need for high computational power.
d. The focus on profit maximization.



GCOOS GULF OF MEXICO COASTAL OCEAN OBSERVING SYSTEM

IOOS Integrated Ocean Observing System

“Satori and the Mystery of the Dirty Data”

Answer Key

1. c
2. b
3. c
4. c
5. b
6. b
7. a,b,c,d
8. b,c
9. c
10. b
11. a
12. c
13. b
14. b

“Satori and the Mystery of the Dirty Data” was created by Bob Currier and Data Goblin Productions.

GCOOS GULF OF MEXICO COASTAL OCEAN OBSERVING SYSTEM

IOOS Integrated Ocean Observing System