CMPINF 2100

Introduction to Data Centric Computing

Midterm Exam Data Overview

Problem overview, goals, and context

So far this semester you have learned…

- Essential base Python programming
    - Data types, methods, attributes, for-loops, comprehensions, slicing, etc…

- Manipulation of NumPy arrays and Pandas DataFrames
    - Slicing – selecting columns and filtering rows
    - Concatenation
    - Reshaping
    - MERGE/JOIN DataFrames by column values
    - Summarize: mean, standard deviation, size, count, number of unique values
    - SPLIT-APPLY-COMBINE to summarize by GROUPS

- Data visualization to visually explore columns (variables) in DataFrames
    - Visualizations DEPEND on data type
    - Visualizations explore MARGINAL behavior (one variable at a time)
    - Visualizations explore CONDITIONAL behavior (group a variable by another)

- Cluster analysis to help find patterns in the data.

You must use <u>ALL</u> aspects learned so far to explore a realistic data application

- You will work through an application very similar to many applications I worked on as a Data Scientist in the manufacturing industry.

- Multiple data sets are provided to you as CSV files.
  - midterm_machine_01.csv
  - midterm_machine_02.csv
  - midterm_machine_03.csv
  - midterm_supplier.csv
  - midterm_test.csv

- You will explore the variables within each file, JOIN them appropriately, and then explore the combined data to identify important patterns.

The data provided to you is based on the following manufacturing scenario:

- A company manufactures cell phone cases. The cases are made from a high-density plastic.
- The company buys the plastic from **2 SUPPLIERS**.
  - The plastic is purchased in **BATCHES** from a **SUPPLIER**.
  - A single **BATCH** is represented by the plastic **DENSITY**.
- The cell phone cases are made with injection molding machines.
  - The company uses **3 MACHINES** to manufacture the cell phone cases.
  - Each **MACHINE** consists of **4 OPERATING VARIABLES** that define how the machine produces the cell phone case. The machines can be operated differently.
  - A single **BATCH** of plastic can be used across multiple **MACHINES**.

The data provided to you is based on the following manufacturing scenario:

- The company wants to produce HIGH QUALITY cell phone cases. The cases are DROP TESTED to ensure they do NOT break or shatter under reasonable use.
  - If the cell phone case breaks during the DROP TEST, the test result is a FAIL.

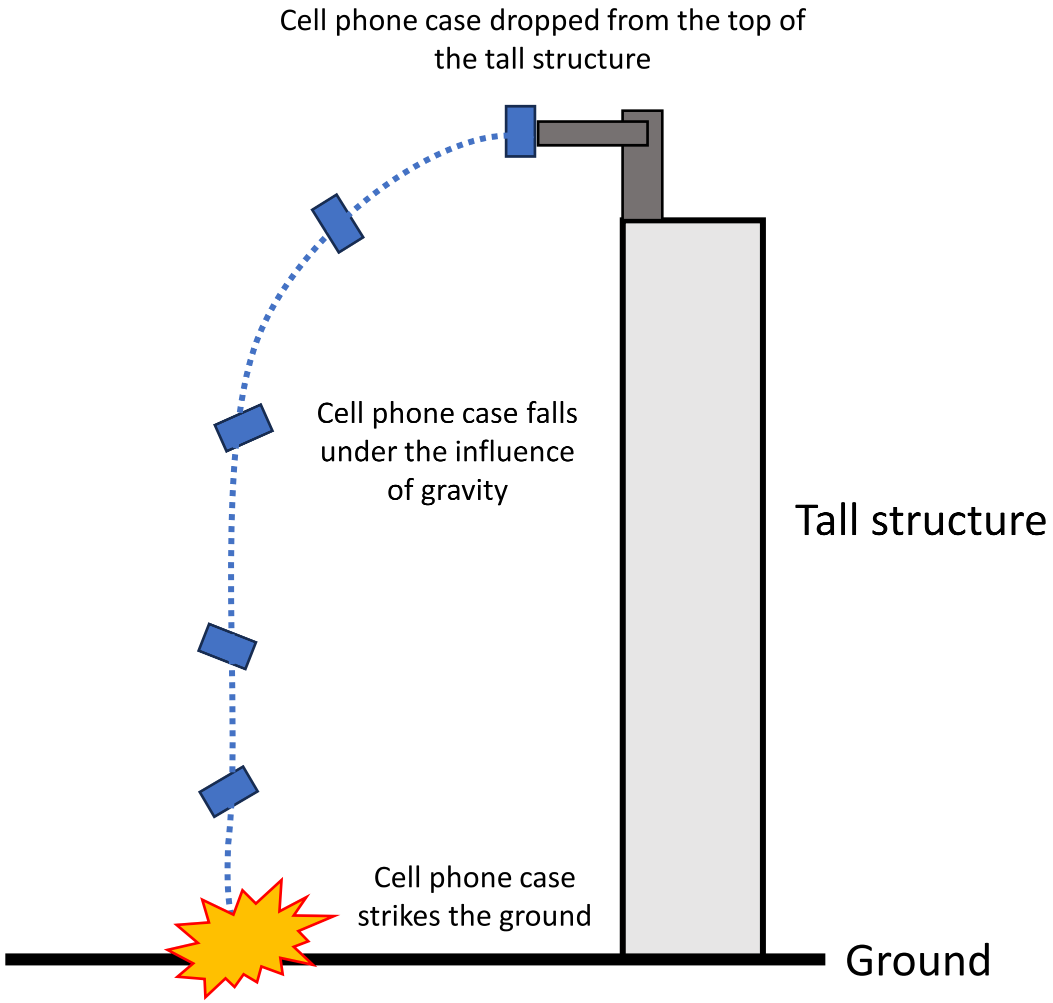
- The company uses Data Science and Machine Learning techniques to examine if:
  - the DROP TEST FAILURE RATE varies across the plastic **SUPPLIERS**.
  - the DROP TEST FAILURE RATE varies across the **MACHINES**.
  - the **OPERATING VARIABLES** impact the DROP TEST FAILURE RATE.

DROP TEST overview

- The DROP TEST works by dropping a cell phone case with a representative phone inside it from the top of a tall structure.

- It falls under the influence of gravity until it strikes the ground.

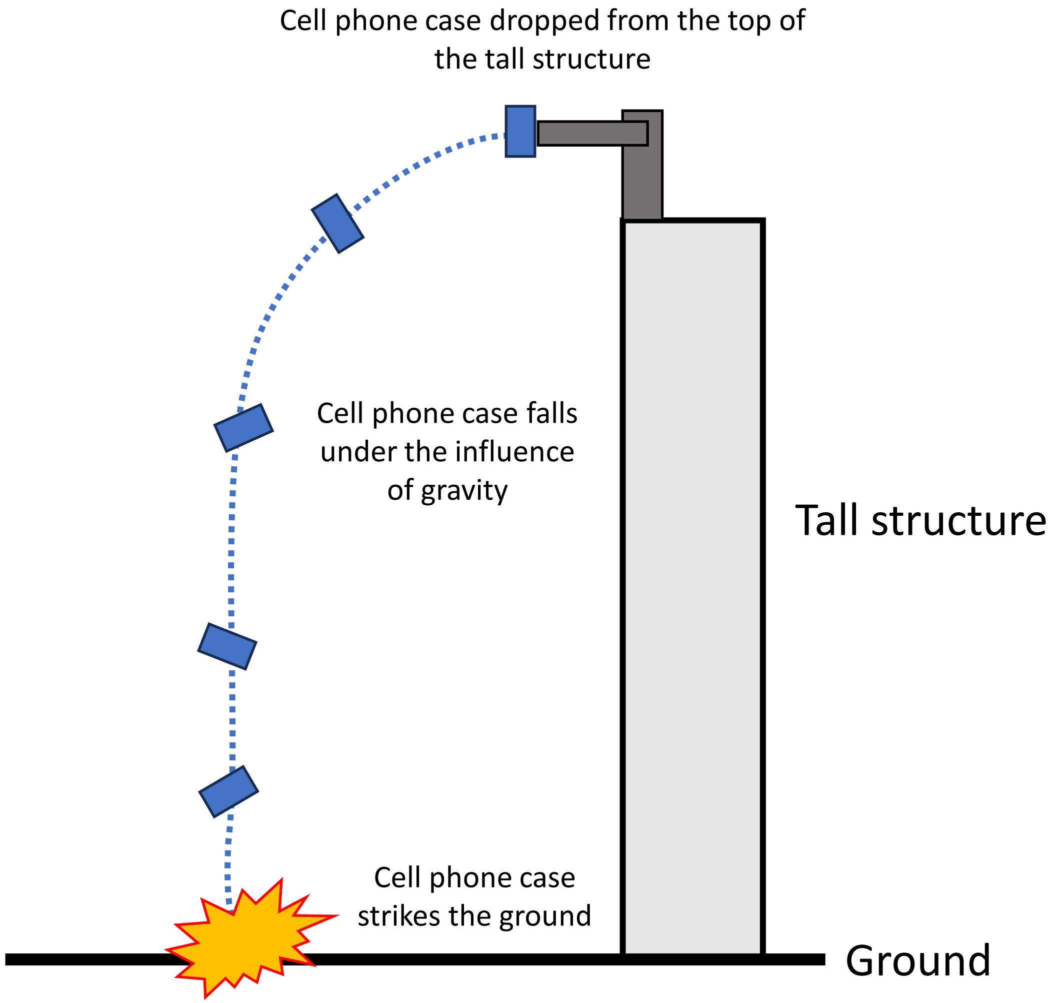
- The case is inspected to see if it survived the fall and protected the phone.



Cell phone case dropped from the top of the tall structure
Cell phone case falls under the influence of gravity
Tall structure
Cell phone case strikes the ground
Ground

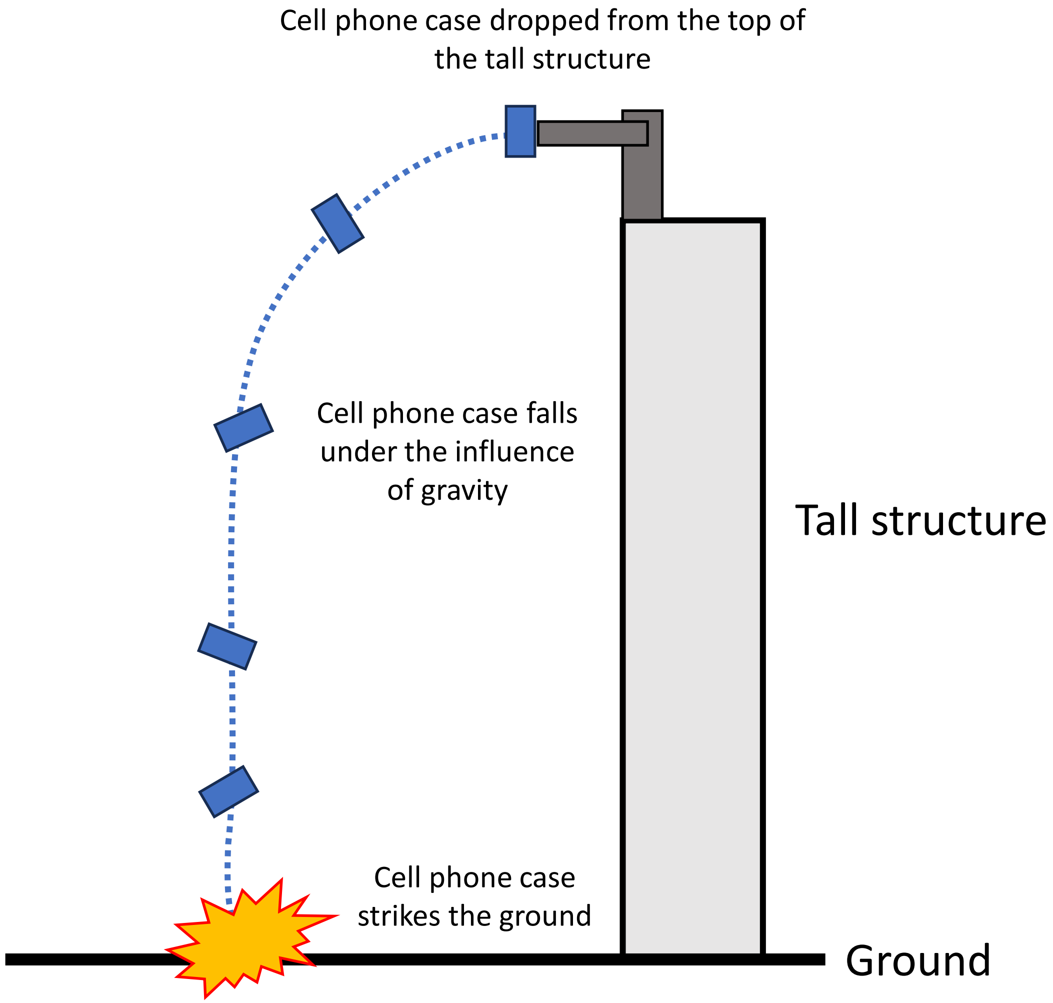
DROP TEST overview

- The DROP TEST may destroy the cell phone case!
- Testing is also time consuming!
  - The company produces THOUSANDS of cases per day.
  - It would take a VERY LONG time to test every case.
- Thus, NOT all cases are tested!
- A SAMPLE of cases are collected and DROP TESTED. tested.


Cell phone case dropped from the top of the tall structure
Cell phone case falls under the influence of gravity
Tall structure
Cell phone case strikes the ground
Ground

DROP TEST overview

- The SAMPLING PLAN requires 10 out of every 100 cases manufactured per **MACHINE** to be DROP TESTED.
- The **OPERATING VARIABLES** used to produce the case on the **MACHINE** are recorded in a data base.
- The **SUPPLIER** that provided the plastic for each case is also recorded in a data base.


Cell phone case dropped from the top of the tall structure
Cell phone case falls under the influence of gravity
Tall structure
Cell phone case strikes the ground
Ground

The data provided to you is based on the following scenario:

- Ultimately, the TEST data are used to TRAIN a classifier to predict if a case will **FAIL** the drop test.
- The variables associated with the production of each case are known in a data base:
  - **MACHINE**, **OPERATING VARIABLES**, and **SUPPLIER**.
- Those variables are used as INPUTS to the classifier.
- However, you are NOT training the classifier in the midterm!
- You are EXPLORING the data. You will learn how to train classifiers AFTER the midterm.

The manufacturing data are stored in separate CSV files for each machine

The **MACHINE** data are stored in 3 CSV files:

- midterm_machine_01.csv, midterm_machine_02.csv, midterm_machine_03.csv
- The file name tells you the machine ID the data come from:
  - For example, midterm_machine_01.csv is associated with Machine 1.

Each CSV file consists of 7 variables:

- ID: The unique unit ID for the cell phone case
- Batch: The batch index that denotes the plastic the cell phone case is created from
  - **NOTE**: the Batch is an INTEGER data type but is a CATEGORICAL variable.
- s_id: The sequential production index for a single cell phone case within a Batch on a machine.
- 4 **OPERATIONAL VARIABLES** that describe the behavior of injection process: x1, x2, x3, and x4.

Machine 1

ID	Batch	s_id	x1	x2	x3	x4

Machine 2

ID	Batch	s_id	x1	x2	x3	x4

Machine 3

ID	Batch	s_id	x1	x2	x3	x4

The batches of plastic material come from the 2 different suppliers

- **SUPPLIER** data are stored in the midterm_supplier.csv file.

- That data table consists of 3 variables:
  - Batch: The batch index associated with the batch of plastic from the SUPPLIER
    - **NOTE**: the Batch is an INTEGER data type but is a CATEGORICAL variable.
  - Supplier: The supplier ID
  - Density: The supplier reported density associated with the batch of plastic

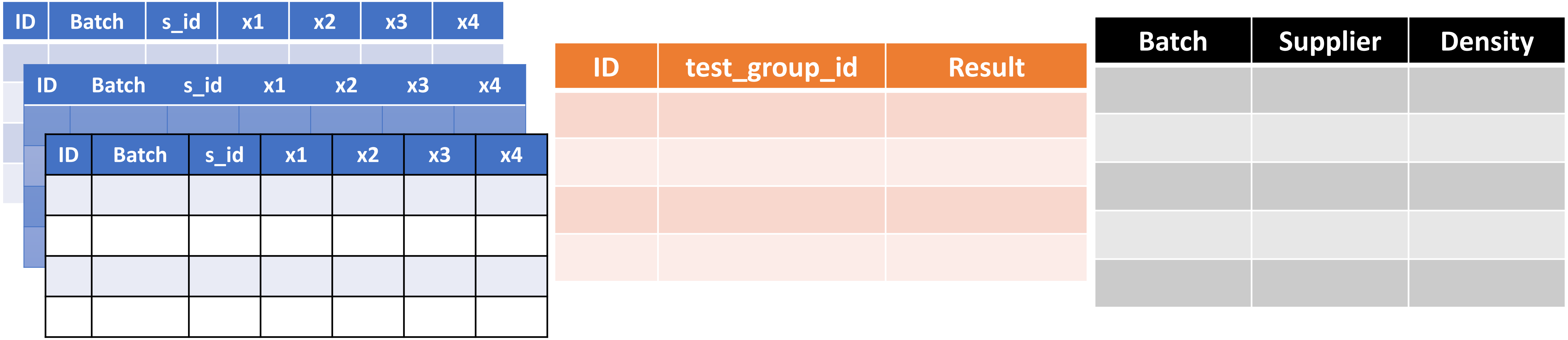
Batch	Supplier	Density

The TEST results are stored in the midterm_test.csv file

- That CSV file consists of 3 variables:
  - ID: The unique unit ID the tested cell phone case
  - test_group_id: Test grouping identification label
  - Result: The DROP TEST result which is encoded as:
    - A value of 1 corresponds to FAIL
    - A value of 0 corresponds to PASS

ID	test_group_id	Result

The 5 data tables need to be JOINED to ultimately link the INPUTS with the DROP TEST result.



ID	Batch	s_id	x1	x2	x3	x4

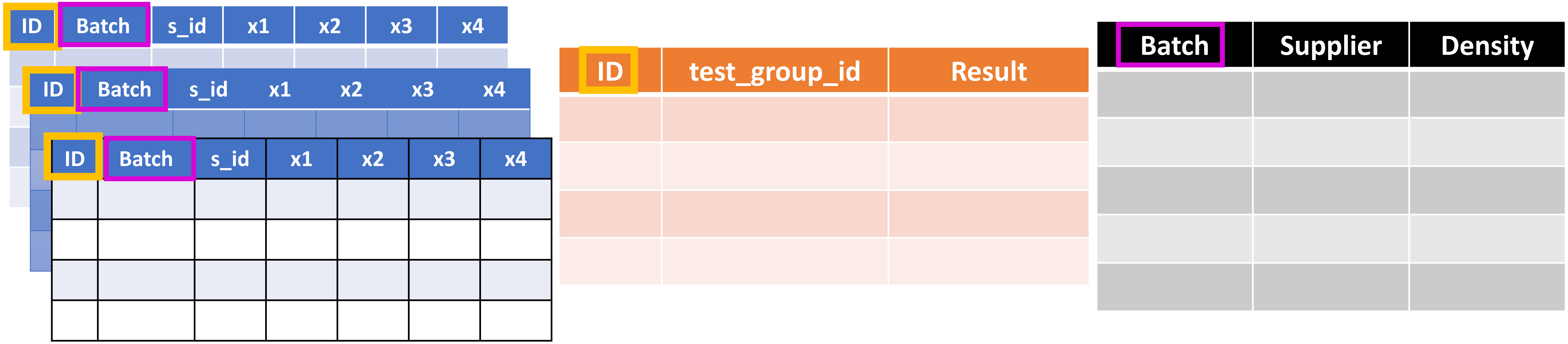
ID	Batch	s_id	x1	x2	x3	x4

ID	Batch	s_id	x1	x2	x3	x4

ID	test_group_id	Result

Batch	Supplier	Density

To do so we need to identify the common KEYS *across* the tables



ID	Batch	s_id	x1	x2	x3	x4

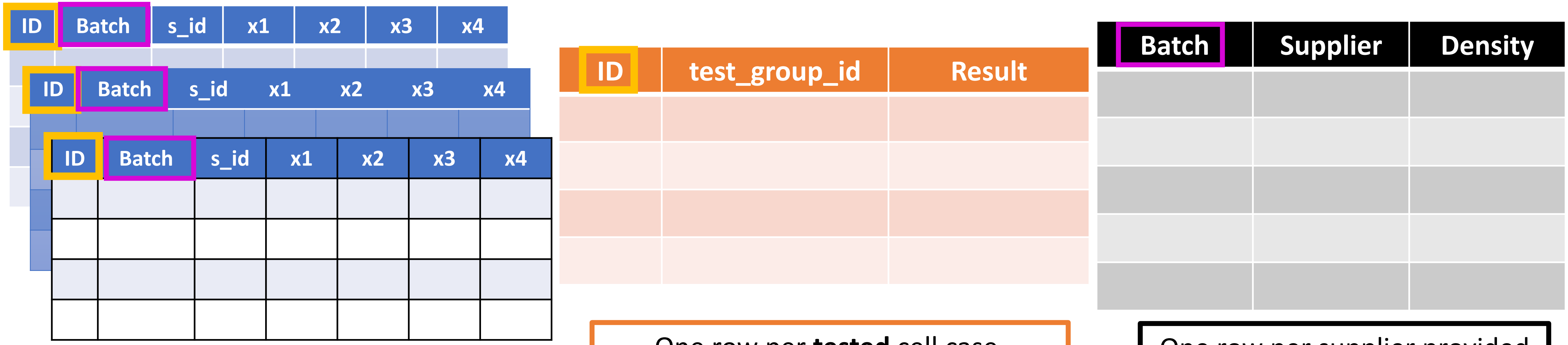
ID	Batch	s_id	x1	x2	x3	x4

ID	Batch	s_id	x1	x2	x3	x4

ID	test_group_id	Result

Batch	Supplier	Density

And make sure we understand what one row represents in each data set!!!



ID	Batch	s_id	x1	x2	x3	x4

ID	test_group_id	Result

Batch	Supplier	Density

One row per **manufactured** cell case **per machine**.

Contains every manufactured cell phone case per machine.

One row per **tested** cell case.

Only contains test results for the tested cell phone cases.

One row per supplier provided **batch**.

However, combining the machine data requires extra attention

ID	Batch	s_id	x1	x2	x3	x4

ID	Batch	s_id	x1	x2	x3	x4

ID	Batch	s_id	x1	x2	x3	x4

The machine tables do **NOT** contain any identifying information associated with the MACHINE!

The CSV file contains the machine ID!

You must **ADD** a column, machine_id, that identifies the machine as 1, 2, or 3 BEFORE the machine data sets are combined.