# Field data analysis

## Introduction

Let’s look into a typical dataset coming from a small plot field trial. A field is divided into small plots, each of which can be about 10’ x 40’ large and contains either only the plants (i.e. the control) or the plants with one of the microbes we’re interested in (called symXXX). Within a field, each sym is tested 4 times (called reps – for repetitions) and the control 8 times. It is interesting to note that fields are planted following this rep pattern. First all syms on rep 1 are planting, then rep 2, etc…

Here is an example field, with yield shown on the left, the layout of sym placements and the repetition number.

A close up of a device

Description automatically generated

## Data

The provided csv file contains data from one field. It contains the following information:

* Field\_id, the unique code describing a field
* Sym, the code of the sym or the control
* Row, the row position in the field
* Range, the column position in the field
* Value, the yield within that plot
* Rep, the rep number

## Discussion

1. As a biological dataset, the yield data can be pretty noisy. What are your thoughts on possible sources of noise?
2. How would you approach estimating the improvement (or uplift) to yield brought by a particular sym?

## Deliverable

* Prepare a Jupyter notebook detailing your approach and results
* Feel free to use any library you deem useful such as pandas, sklearn, statsmodels, …
* Try to encapsulate your code into functions for readability (you can also put your functions in a separate module to be loaded from the notebook).
* Start your notebook by including the potential query that would collect the data from the provided csv from our database. The database is set up to have two tables, the *yield* table and the *sym* table.

The *yield* table contains the following attributes:

* + Field\_id, the unique code describing a field
  + Row, the row position in the field
  + Range, the column position in the field
  + Value, the yield within that plot

The *sym* table contains the following attributes:

* + Field\_id, the unique code describing a field
  + Sym, the code of the sym or the control
  + Row, the row position in the field
  + Range, the column position in the field
  + Rep, the rep number