Use cases to show that it's needed to have *key yes*, *identifying yes* and distinct yes

Q1: Why we need to distinguish the same entities using *key yes* and *identifying yes*)?

Assume the following tabe is the measurement for some plant tree at different spots.

Consider this question that a user may ask. Give me the average dbh for every *piru* tree (i.e., tree entity). First, we have three observations here. But how many tree entities here is a question.

There are several cases to consider:

- Case 1: The naive extreme way to interpret the data is that each observation is from different tree entity. Then, we have **four** tree entities. This may be too strict. People may say, well, I have some observations for the same entity.
- Case 2: The second naive extreme way is to interpret that different *spp* represent the different tree entity. That's obvious that *piru* is different from *abba*. With this constraint, we get **two** tree entities.
- Case 3: the assumption of case two has some obvious problem. People want to further limit that the same *spp* in the same *plt* should represent the same tree entity set. To achieve this, we use *identifying yes*. Now, it should return

$$(A, piru, 36), (B, piru, 33.2), (B, abba, 34)$$

plt	spp	dbh
A	piru	35.8
A	piru	36.2
B	piru	33.2
B	abba	34

(a)

plt	area	spp	dbh
A	1.0	piru	35.8
A	1.1	piru	36.2
B		piru	33.2
B		abba	34

(b)

Table 1: Dataset

Q2: Why we need to distinguish the same entities using *key yes* and *distinct yes*) to identify the same observation? For the example, we have one observation for splot A. What's the semantic purpose of this? What kind of query may need this? E.g., how many spots in this dataset?

Note 1: if one observation type is marked with *distinct yes*, all its measurements should be marked with *key yes*. Otherwise, we may have the same observation with different measurement values. E.g., what will happen for the following:

observation "o1" distinct yes
  entity "Plot"
  measurement "m1" key yes
    characteristic "EntityName"
    standard "Nominal"
  measurement "m2"
    characteristic "area"
    standard "sqft"

Will $(A, 1.0)$ and $(A, 1, 1)$ be treated as the same observation? According to the semantic meaning, they are the same observation because they have the same value on the key measurement and this observation type is marked witht *distinct yes*. However, there is something wrong here.

Table 2: Dataset

Based on this note, it seems like it is not useful to denote *distinct yes*. Basically, once all the measurements are marked with *key yes*, it automatically infers that it is distinct yes.