Homework assignment #10

Due in class Mon 11/26

1. Looking back at the papers we read (5 points)

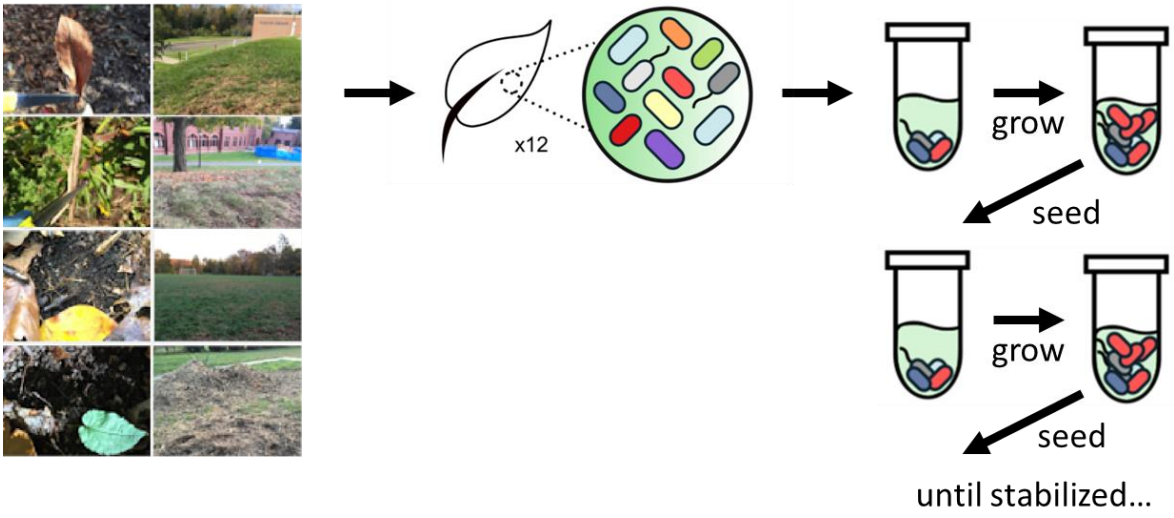
The Perusall discussions raised many excellent questions. Hopefully, our conversations in class clarified many of those points, but probably not all. Going back to the papers we talked about – name one question or point of confusion that remained open for you, and to which you would like to know the answer. This may be something you (or someone else) asked on Perusall, something that came up in class but was glossed over quickly, or something you thought about later. The question can be about any of the modules we discussed: imaging, sequencing, neural.

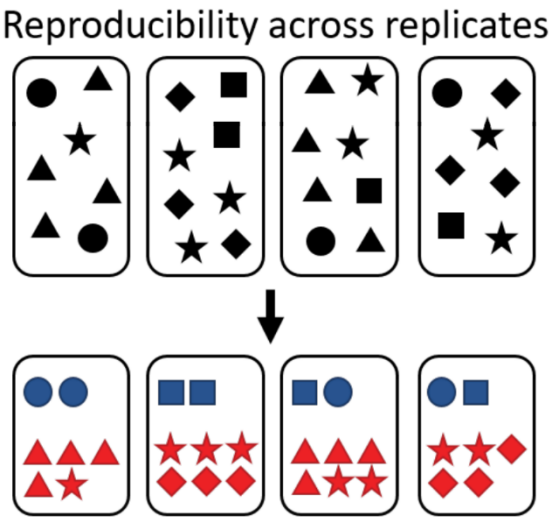
2. The right variables for microbial ecology? (15 points)

Introduction & Background

Microbial communities, even those growing in identical conditions, can be hugely variable in composition. This is because lots of different bacteria can run the same metabolic reactions, or occupy virtually the same "niche". So a set of communities "doing the exact same thing" (e.g. converting the same sugar *S* into the same waste product *W*) might be composed of species with very different names. But behind this apparent variability, something is reproducible. Can we extract this from data?

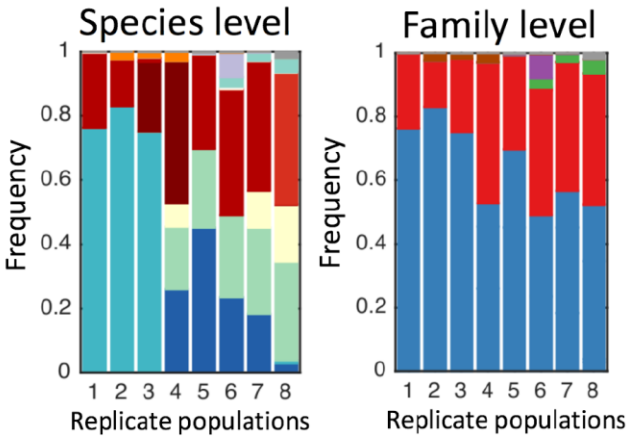
Imagine an experiment where Jeff collects samples of bacteria from the leaves in his back yard (each harboring a complex community), grows them up in identical test tubes with a simple, well-controlled "food", e.g. glucose, and transfers them into fresh tubes every day for a few weeks until they stabilize. Many original species go extinct, but the remaining ones form a stable community that successfully metabolizes the source of carbon Jeff provides.





Although Jeff takes great care to provide the same sugar and do the experiment the same way every time, he finds that the resulting communities assembling in these "replicate" experiments can be quite different. But behind this variability, Jeff starts noticing patterns, as illustrated in the cartoon. Here, four replicates appear very different (the different shapes represent different species). However, if instead of the microscopic level of detail we introduce a coarse-grained description, where we designate some shapes as "blue" and others "red", we uncover the hidden reproducibility: for all these communities, the blue-to-red ratio is roughly 1:3. This makes sense to Jeff, because it's plausible that there are some metabolic tasks to be performed -- here, a "red metabolism" might be converting the original sugar into some intermediate, and a "blue branch" might be further breaking down this intermediate to the final waste product. And although the exact details of who participates in these branches may differ, it's plausible that the amount of biomass each niche can sustain remains fixed by the environment.

But how do we get the right coarse-graining? Since Jeff knows a lot about taxonomy, he manages to guess a coarse-graining that seems to work quite well: on the left, the species-level description; on the right, a coarse-grained description where species are grouped by their taxonomic "family", which seems to be much more reproducible.



But Jeff's friend Ben knows a lot about metabolism and proposes a different grouping that he is convinced should work better. He wants to quantitatively prove to Jeff that his approach is superior! But how can we quantitatively evaluate how "good" a particular grouping is?

This question is the subject of this homework.

Your mission

A first hunch could be to quantitatively assess reproducibility across a set of replicates. But this can't be the only thing we want from a good metric. Otherwise we'd just color all species "black" and declare immediate victory: all replicates are 100% black – perfect reproducibility!

So to make the problem well-posed, in addition to a set of replicates that we'd like to see as "similar", we also need a set of examples we would like to see as "different". For instance, Jeff might try a different carbon source, and obtain a second set of replicates. He would then look for a coarse-grained description that makes all communities in the first set look similar, all communities in the second set look similar, but preserving the differences between the two.

Specifically: Looking back at the red-and-blue cartoon above, the second set of replicates could have a different ratio of red-to-blue, or perhaps could have a third color – a third category of species ("green") that does not appear in the first set of replicates, or shows up only rarely.

1. Inspired by the cartoons above, make a mock "data table" to work with. Let's have 10 species, and two sets of 3 replicates (the "A" and "B" set). Some species may only appear in a few samples. Design your dataset so that a good coarse-graining into two categories exists: one that makes replicates from the same set look similar, but A and B remain distinct. Show your dataset and the coarse-graining you designed.

   As you continue with this assignment, you may discover you should go back and revise your "mock dataset" to better illustrate a particular issue. Allow yourself to do that – no need to become wedded to the particular (arbitrary) choice you made at the start!

2. Qualitatively describe several ways (at least two) in which a coarse-graining can be "poor". Illustrate them on your mock dataset. (Show two "bad" coarse-grainings.)

3. Can you think of a quantitative scoring system that would assign a high score to the "good" coarse-graining you constructed in (1), but low scores to the "bad" ones in (2)?

   Once you think of some scoring system, can you think of a way to "fool it" -- trick it into reporting a high score on some other example of a poor coarse-graining? If so, add it to the list of "bad examples" in (2), and refine your scoring criteria!

4. Real life is often trickier than mock examples. Let's think about possible complications.
   a. Does your scoring metric generalize to 3-category coarse-grainings?
   b. Imagine that your communities have many "rare species", each present at very low abundance. The issue with such species is that their detection can be very stochastic: if you didn't detect it, it doesn't necessarily mean it wasn't there... Does this create a problem for your scoring metric? For instance, imagine adding 6 more species to your dataset, each unique to a sample. Can your scoring metric be fooled by some coarse-graining that places too much weight on these rare species? If so, can you propose any solution to this issue?

Some comments:

- Useful things to think about: how does this problem relate to the question of "dimensionality reduction"? To "supervised learning"?

- "Morally speaking", a good coarse-graining is one where different sets of replicates are mapped into tight, well-separated clusters. And your scoring metric is a way to quantify the terms "tight" and "well-separated" in this sentence…

- You may find it useful to recall the concept of mutual information from earlier in this class. (You don't have to! But this could be one avenue for designing the scoring metric.) Intuitively, a good coarse-graining is an example of *compression* – we try to forget as much information as possible, while retaining the information that matters – namely, whether a replicate is from set A or set B… Is there a way to make this intuition precise?

- An optional extra question.
  Rather than using your metric to **evaluate** coarse-graining someone proposed, one could imagine using it to **construct** the "best" coarse-graining – one that maximizes your metric. That way we could "learn" the correct coarse-graining directly from the data, without needing to know any biology at all! We could "rediscover" the classical taxonomy, or even propose corrections to it... or not? Are there any difficulties you foresee with this approach? (Intentionally open-ended)