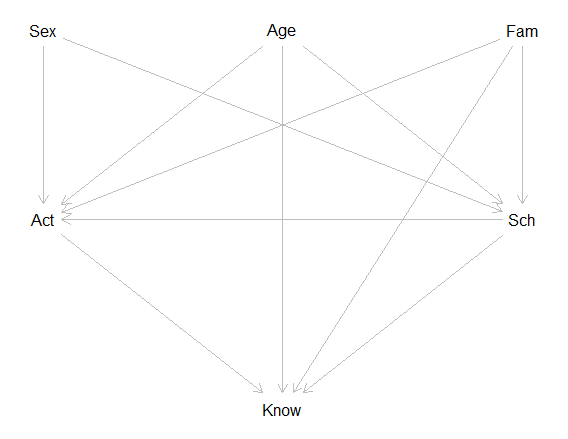
As we agreed during the TAC meeting, I focused on the first part of the project, on knowledge accumulation and change with age. This section should become the first paper and includes at the moment two preliminary subsections: one on the determinants of knowledge change, on which I did a decent amount of work and that feels quite solid, and a second one on knowledge structure and categorization of foraged goods, which is fuzzier and still requires some work. Here I explain the work done for each of the two parts and conclude with the next steps as I see them.

1. The **first part** should be relatively straightforward, now. The logic is to define the contribution of individual level traits to variation in number of species listed in the free list and in other knowledge tasks. An abstract would go about as follows: ‘Knowledge is one of the factors necessary to exploit the natural environment and to have time to learn it humans might have evolved a long pre-reproductive period. However, there is no comprehensive study on how ecological knowledge differs in children of various ages or on other determinants of individual variation. With this paper, we aim at exploring what determines the differences in the ecological knowledge of children. Children knowledge is assessed mainly through free listing, complemented by a picture recognition task. Traits evaluated include age, activities performed and family composition. Cross classified logistic regression models are used to determine the contribution of the various traits to knowledge.‘
   1. To help constructing the analysis and defining more clearly our approach, we developed a **structural causal model**, which summarizes the expected relation between the data. It basically is a graph in which arrows are drawn to indicate the correlation between the factors as expected given our previous knowledge. This helps with the analysis because it allows to define the functions underlying our variables. In this case, we say that knowledge of an individual (as evaluated in the free list) depends on age, activities performed and other factors, as described below. Have a look at the graph and at the explanation that follows.

Knowledge changes as individual get older. We expect to observe an increase with age, with individual differences that can emerge as a result of several factors. Age has a direct effect on knowledge, in virtue of the increasing cognitive abilities of human brains that allow to store and manage information (Age -> Know). Family context, such as the presence of parents or older siblings, influences what and how much children know (Fam -> Know). Knowledge also depends on the activities individual perform, because of the exposure to the relevant information and of the returns from learning these information (Act -> Know). Schooling could have an effect on the amount of information individual can manage, but most probably it reduces the contact with natural environment and thus the knowledge of it (Sch -> Know). As individuals age, they both perform new or different activities and have different probability of access to schools (Age -> Act and Age -> Sch). At the same time, the family context also influences activities performed and schooling as the absence of parents or the birth of new siblings impose to reallocate time or energies of children accordingly (Fam -> Act and Fam -> Sch). Finally, the sex of individuals determines both which activities are performed, some tasks being typically done by girls and other by boys, as well as school attendance, since boys often receive higher investment in a family (Sex -> Act and Sex -> Sch).

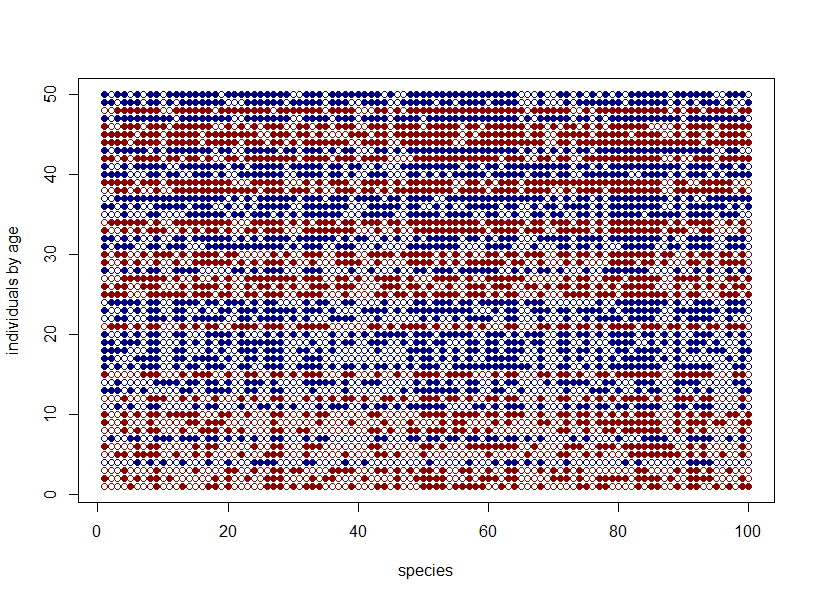
Summarizing, we expect knowledge to be determined by age, as younger children don’t know as much as older children; by activities performed, as certain activities favor the learning of ecological knowledge; by family, as access to knowledge depends on the access to older individuals able to transmit it, for example; and probably by access to schools, that provide certain types of knowledge but not others. These are the core determinants of knowledge according to our current formulation, do they seem sound to you? I would like to have your opinion and I am happy to discuss them further at any point.

* 1. Based on the graph above, we simulated **mock data** to be fed to the analysis. I started by creating a bunch of people (N) with age (5 to 20) and sex (-1 or 1). Each individual belong to a household and the other individuals within the same household are its older and younger siblings. In each household lives also a certain number of people above 20 years of age. Children have spent a certain number of years going to school. All children below seven years and a certain proportion of the total population have never been to school. Seven years and older children have spent some years in school, with older children that can have more years of schooling. One of the two sexes has, on average, less years of education. A higher number of siblings also reduces schooling. Children can perform any of the possible activities. Each of these is more common among older than younger children, at different levels, and some activities are practiced more by children of one of the two sexes. School years reduce the probability of performing certain activities, whereas the number of siblings can have positive or negative effect. This way all mock children have individual characteristics.

Now I assign them species known (equivalent to listed in the free list, whereas the not known are those not listed). A certain number of species were created with a curve of probabilities of being known at different ages. Each species can be listed by each child. The probability of it being listed increases with age, as children knowledge increase. Each child's knowledge of species depends, thus on its age. Other traits of the child affect the probability of knowing each species: the activities a child practices create an offset of ‘knowledge years’. For example, by practicing an activity, a child has the same probability of knowing a species of older children (or younger, as some activities impede the acquisition of ecological knowledge). Years spent in school reduce the probability of knowing a species. The same does the number of younger siblings, while the number of older siblings increases species known (i.e. the position in the sibship has an effect). Adults in the household also increase knowledge, each additional adult adding a bit less knowledge than the previous one. At the end, each individual receives a list of species known. The probability of a species to be known depends on the age of the child, on the activities performed, on the number of years the child spent in school and on the number of adults in the household as well as on the number of younger and older siblings, as expected given the graph at point 1.1.

The result is shown in the following figure, where children are on the y axis and species on the x axis. If child y knows species x, the dot is colored, otherwise it’s empty. The children are listed in order, from the younger (bottom) to the oldest (top). Young children know less than old children in my simulated data, and some species are more widely known than others.

When simulating the data, I also thought of which type of variables would be necessary for the model and how to collect them in the survey



* 1. We made several **models** to analyze these mock data. The last and most complete one includes the effect of individual children and species, as well as children age, family, activities practiced and years of school. It is a logistic regression where the probability of a species being known by a child depends on the child, the species and the child’s characteristics. The r code is about like this:

knsp ~ dbinom( 1 , p ) ,

logit(p) <- a\_bar + a[ID]\*sigma\_a + s[spp\_id]\*sigma\_s +

bA\*age + bS\*school\_years +

bY\*young\_sib + bO\*old\_sib +

bAd\*adults + h[HH]\*sigma\_h +,

AE[ 1 , spp\_id[i] ]\*am[i,1] +

[…]

Where all the parameters are decentered for the model to run better and to make easier priors.

The model works fine on the fake data and recovers the information we know was embedded in the data we constructed

* 1. We are now planning to run the models on several batches of simulated data with different **sample sizes**. This way we can see the effect on standard deviations- for example- of interviewing 100 instead of 20 children. We should also be able to evaluate other characteristics of sample size, for example try to sample children from different families instead than multiple children from one family. When I will be back in Leipzig next week, this is priority.
  2. While creating the mock data and thinking of the model, I thought of the **data I need to collect** to feed into the model. They will be collected in two parts: household surveys, implemented with a short reproductive history to collect info about sibship composition, and children interviews, with freelist and picture recognition. The model should work with any question with yes/no or true/false answer also. I am discussing widely with Silke and Bret and everyone else, and I am changing and adapting the checksheets I started working with, last year. We are coming to an agreement on the whole pipeline for the transcription of the data, following the requirements for reliability and reproducibility, but also the needs of data collection in the field. I attached to the email the current version of the checksheet for household composition (it will probably need some small changes) and I would appreciate comments, if you think there’s something missing or that could be improved.

1. The **second part** is less defined, both theoretically and practically. The idea is to study how categories for ecological knowledge change with age. Children might change the groups in which they associate the species they forage on as they grow, for example two animals can be grouped together because they are both hunted in the same way but this happens only after the child starts hunting those animals. Children would supposedly list sequentially animals that belong to the same group, whereas animals that belong to different groups would likely be named together less consistently. We imagine that the order and the timing of the items in the freelist would reflect these latent categories.
   1. I **simulated freelist data**, where children name species that belong to different groups sequentially and we record the time at which each word is told. Species of the same group are listed in a sparse order within the group, before changing to naming the species of another group. The time lag between items in the same group is shorter than the time lag between items of a different group. In the picture, each line is the freelist of a child, and time is on the horizontal axis. The words belong to different groups, as shown by their color, and they’re named at sequential times. The time between items of different groups is on average longer.

These data are not fully representative of what we expect to have from the actual freelist. There is no effect of age or other characteristics on how children group items yet. On the contrary, all children include the species in exactly the same categories. But they are a start to start developing a model to analyze this kind of data.



* 1. We think –Richard thinks- we should use a Hidden Markov **Model** to analyze these data. The logic of this kind of models is that they can infer the state of a variable from a signal characteristic. The state here is whether a species named belongs to the same group as the species listed just before it or if it belongs to a different group. This should be inferred by the time between the two words: short time means same group, longer times mean different groups. I still don’t understand these fully, but we are working on this.

1. Summarizing the **next steps**:
   1. Simulate the model with different sample size and define optimal sample size.
   2. Finalize pipeline to transcribe and store the data from the field.
   3. Work on both simulated data and model for the first part.
   4. Hopefully, preregister study before data collection.
   5. Of these, I think that 1,2 and 4 are reasonably going to be completed before I’ll be going to the field. The second part of the project will require some more thinking, I believe.