

EVOS data project - Plots

Jessica Couture

10/6/2017

Contents

Background	1
MS Plots:	1
1. Project Status Reporting	1
2. Data Characteristics	2
S2 Figure. Temporal trends?	6
3. Which characteristics are most important?	8
4. Reasons for not sharing	8

Background

This script produces graphics for a publication reporting the results of a two year data archiving effort by a group of researchers and students at the National Center for Ecological Analysis and Synthesis at UC Santa Barbara. The Exxon Valdez Oil Spill Trustee Council (EVOSTC) was formed following the Exxon Valdez oil spill in Alaska in 1989. Since then, the EVOSTC has funded hundreds of projects and in 2012 we initiated an effort to recover and archive the data collected through these EVOSTC funded projects. The recovery effort spanned two years.

For this paper we ask 3 main questions about the data collected from the Exxon Valdez Oil Spill Trustee Council funded projects:

1. Twenty-five years after the EVOS, for how many projects funded by EVOSTC can we collect data?
2. Are there differences in data reporting based on characteristics of the data?
 - Research field
 - Sector of researching body
 - Year data projects ended
3. Which of these characteristics are most *important* in determining if a dataset will be successfully recovered and how do the important characteristics influence the output (success)?

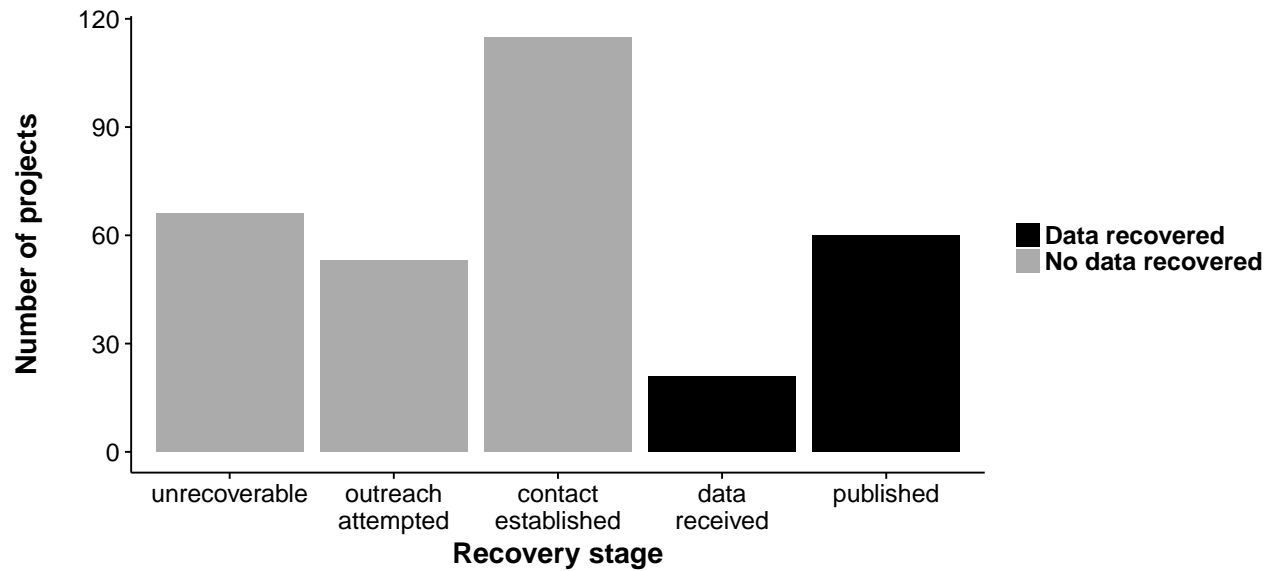
We were also interested in why data were not recovered and plotted reasons why recovery was unsuccessful.

MS Plots:

1. Project Status Reporting

Twenty-five years after the EVOS, for how many projects funded by EVOSTC can we collect data?

Final status of all projects from which data were requested. Black bars are projects for which data were successfully acquired, grey bars represent projects for which no data were acquired.

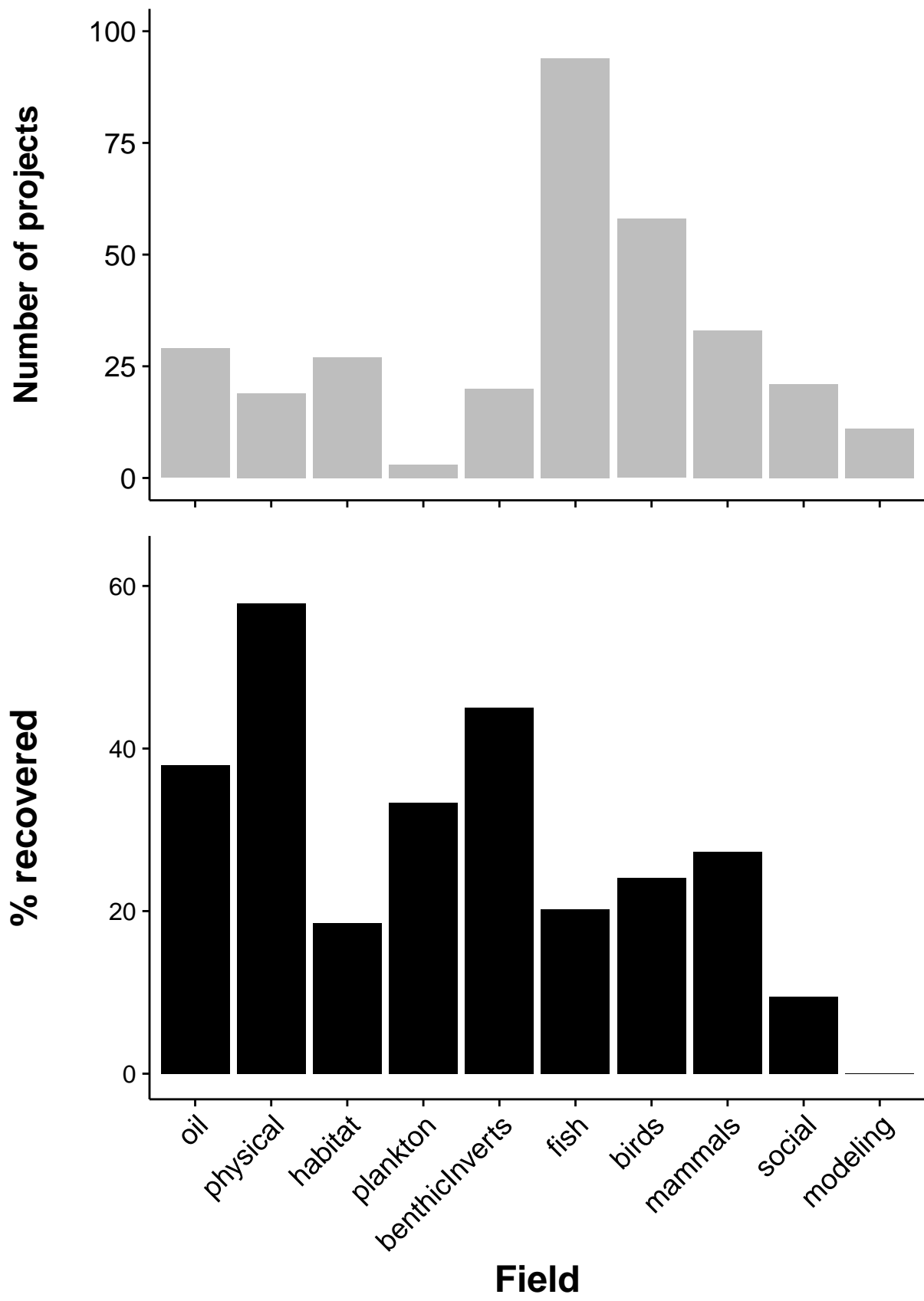


2. Data Characteristics

Are there differences in data reporting based on characteristics of the data?

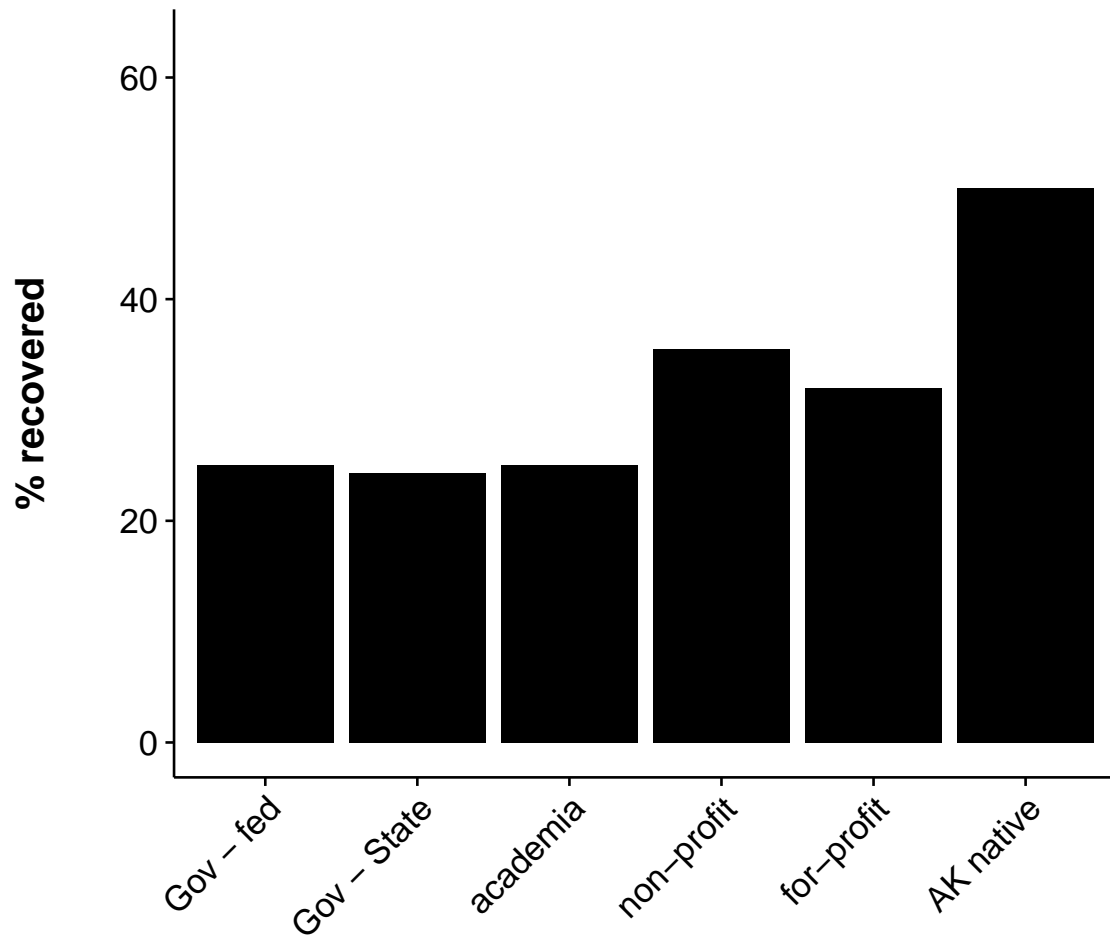
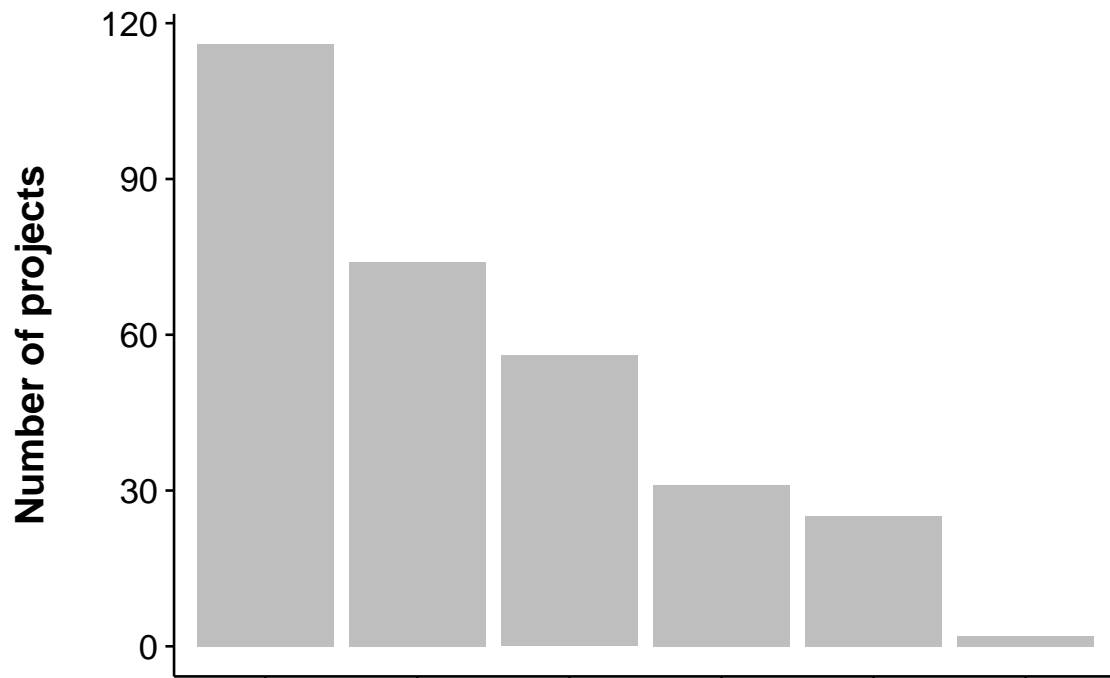
Research Field

Percent success of data recovery, with projects grouped by research field. The top plot is total number of projects funded for each field, the bottom plot shows the percent success for the given field.



Awardee agency sector

Percent success of projects by agency sector based on the agency affiliation of the principal investigator on the project. The top plot is total number of projects funded for each sector, the bottom plot shows the percent success for the given sector.



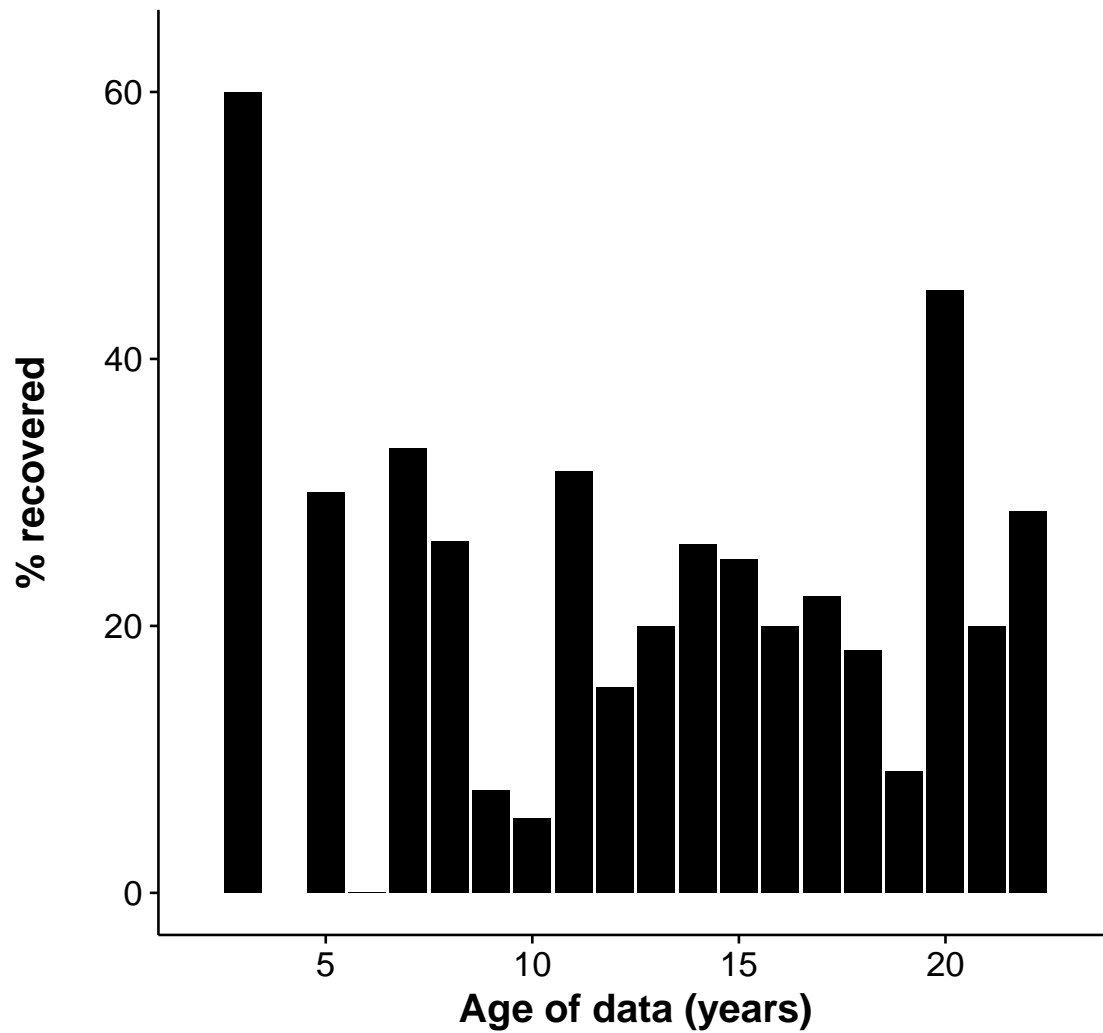
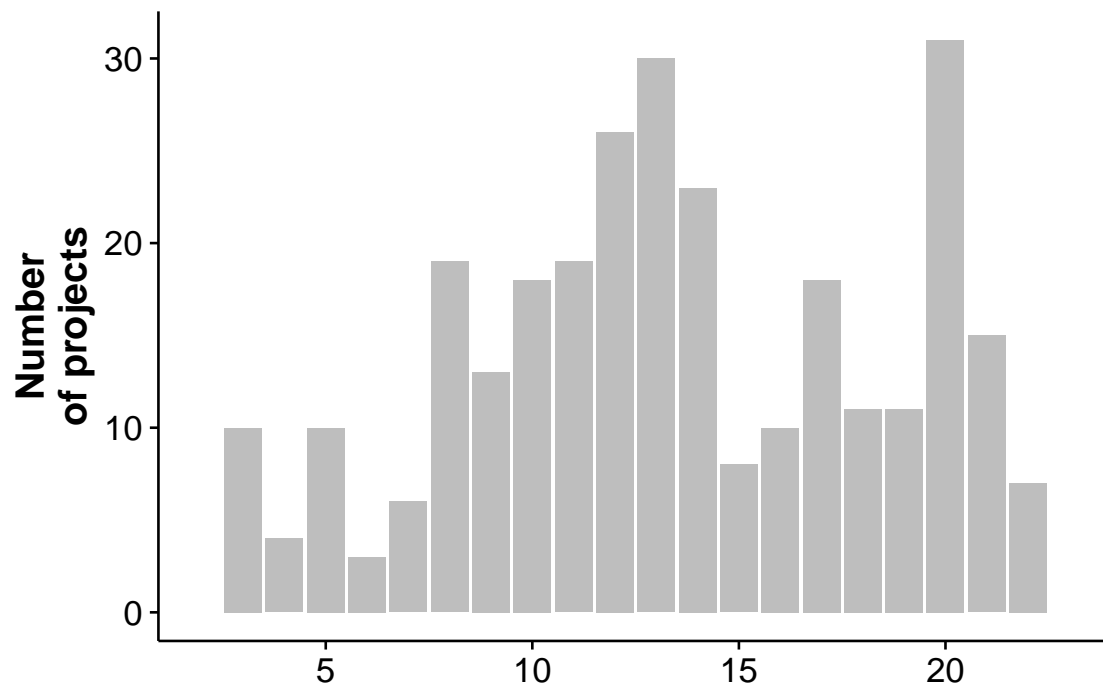
Sector

S2 Figure. Temporal trends?

Is the availability of data correlated to how old the data are?

Percent success of projects by age of data. Age is calculated based on number of years between the last year of EVOSTC funding and start of the archiving project (2012). The top plot is total number of projects that ended each year, the bottom plot shows the percent success for the given year.

The x-axis might not be initially intuitive but we are representing years since a project ended in opposite-chronological order to show any effect of increasing age of a dataset. We chose this design to be able to compare to Michener et al. 1997 - fig 1.



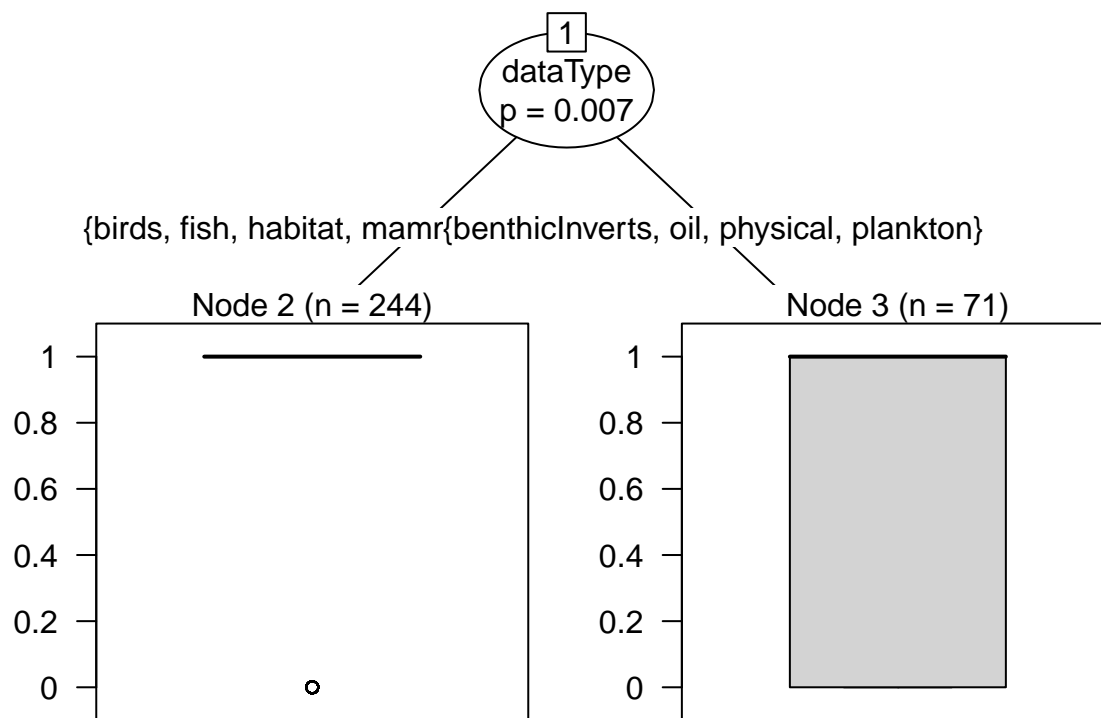
3. Which characteristics are most important?

How do the important characteristics influence the output (success)?

Classification tree

Classification tree of which data variables and characterizations predicting successful and non-successful data recovery. Data field was the only variable that could be used to predict data availability.

```
rs1t2$dataType<-factor(rs1t2$dataType)
partreeBio<-ctree(succ~agSubGrp+end+dataType,data=rs1t2)
plot(partreeBio)
```



4. Reasons for not sharing

Why don't people share data?

Reasons given for not providing data. Communication loss and lack of contact information were the main reasons data were not obtained.

Data details: Here I used all of the data that were not collected and sorted them into categories based on the redmine notes we took:

- All datasets labeled “emailed” were grouped with a other datasets that were labeled “unrecoverable” because no contact information could be found.
- All datasets labeled “replied” we put into their own category called “communication lost”.

- The “data lost” category includes all datasets in which someone confirmed that the data no longer existed either due to damage, non-persistent formats (not including printed/non-digital), etc.
- The other are self-explanatory, but represent confirmed responses as to one of these reasons.
- I tried to separate out the datasets that were not recovered but “should be on the CD” to sort these into their own category, but we asked sufficient notes to confidently add this grouping.

