**Natural language processing (NLP), PaleoDeepDive, and “dark data”**

## Problem Set 1:

+Which taxon names appear most often? Print the top 20 taxa. +What are the commonly used or colloquial names for these 20 taxa, if any? (Look them up online if you need to.) +Which of these are probably not good data? Why? +Conversely, which taxa are likely candidates for pyritization, and why?

> counts[order(-counts[,"matches"]),][1:20,]

string matches

10607 Lar 719

10574 Ave 552

10782 Pic 209

11726 Fera 200

12505 Fera 200

10892 Rall 132

10647 Alca 104

15001 Calcare 103

10969 Scop 73

3894 Ostracod 63

10267 Boida 60

6545 Gastropod 44

4617 Acar 43

11694 Theri 39

10345 Annulat 37

126 Trilobit 31

13581 Gale 31

5611 Articulat 28

14492 Articulat 28

4926 Brachiopod 24

> taxa[10607,] Lari (type of gull)

> taxa[10574,] Aves (birds)

> taxa[10782,] Pici (suborder of Piciformes, includes toucans, woodpecker, barbets, honeyguides)

> taxa[11726 ,] Ferae (cats, dogs, seals,…)

> taxa[12505,] Ferae (cats, dogs, seals,…)  
> taxa[10892,] Ralli (type of bird – rails and others)

> taxa[10647,] Alcae (suborder of Charadriiformes – includes auks, puffins, murres)

> taxa[15001,] Calcarea (class of calcareous sponge)

> taxa[10969,] Scopi (suborder of bird)

> taxa[3894,] Ostracoda (class of crustacean – also called mussel shrimp or seed shrimp)

> taxa[10267,] Boidae (a family of snakes – also called boas)  
> taxa[6545,] Gastropoda (A class - includes snails and slugs)

> taxa[4617,] Acari (a taxon of arachnids – includes mites and ticks)

> taxa[11694,] Theria (a subclass of mammals – marsupials and placental mammals including koalas and kangaroos )

> taxa[10345 ,] Annulata (a species of sea anemone – commonly called the ringed anemone / corkscrew anemone)

> taxa[126 ,] Trilobita (marine group of arthropods – trilobites)

> taxa[13581,] Galea (genus of rodents – includes cavies)

> taxa[5611,] Articulata (a class of brachiopods - includes bivalves - or a subclass of crinoids)  
> taxa[14492,] Articulata (a class of brachiopods or a subclass of crinoids)

> taxa[4926 ,] Brachiopoda – brachiopods or lamp shells

All of terrestrial taxa are probably not good data. I am including bird taxa in the terrestrial category. So, Lari, Aves, Pici, Farae, Ralli, Alcae, Scopi, Boidae, Acari, Theria, and Galea are probably not good data.   
  
Convsersely, all marine taxa may be likely candidates for pyritization. From the list, this includes Calcarea, Ostracoda, Gastropoda, Annulata, Trilobita, Articulata, and Brachiopoda.

This makes sense because pyritization requires low oxygen conditions, which only occur in the marine realm. This allows anaerobic sulfate-reducing bacteria to break down organic matter, and convert it to pyrite via oxidation if there is a sufficient concentration of reactive iron in the water.