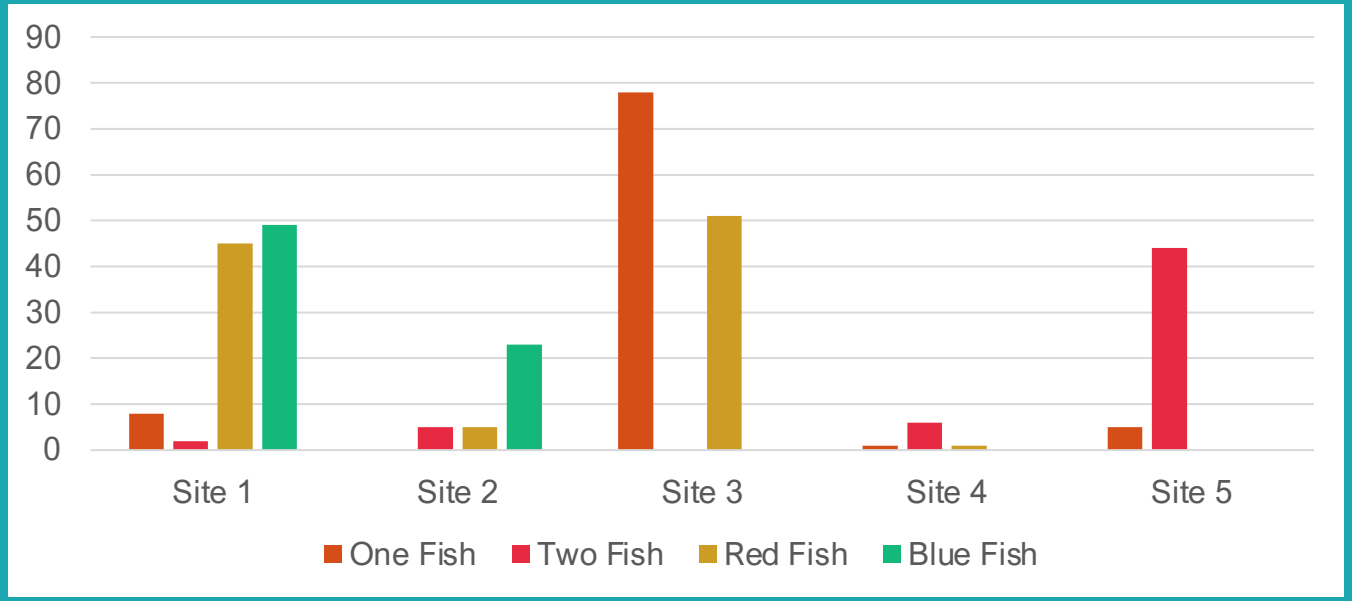
Big Data in R

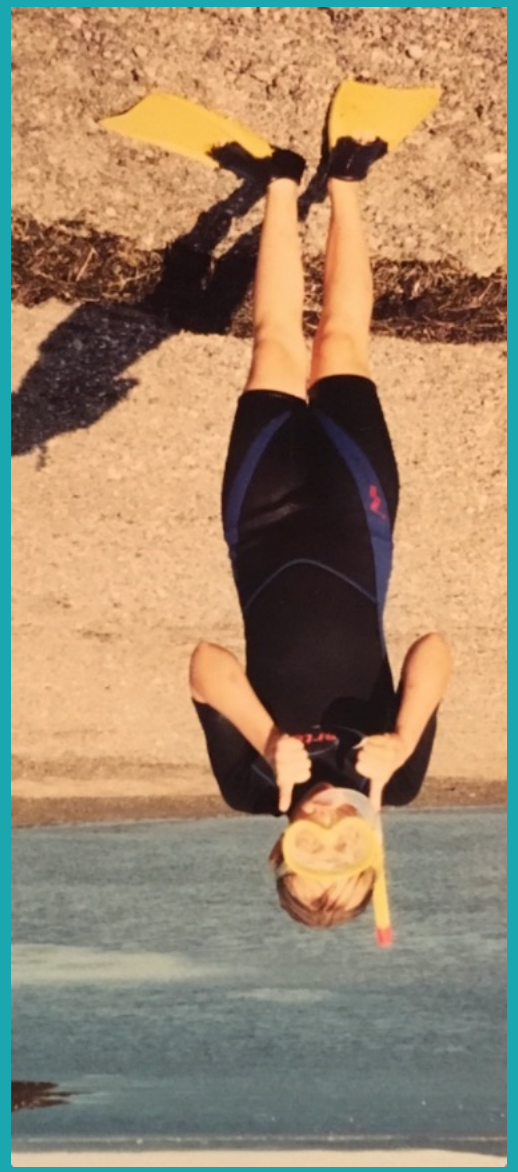
ZOË KITCHEL

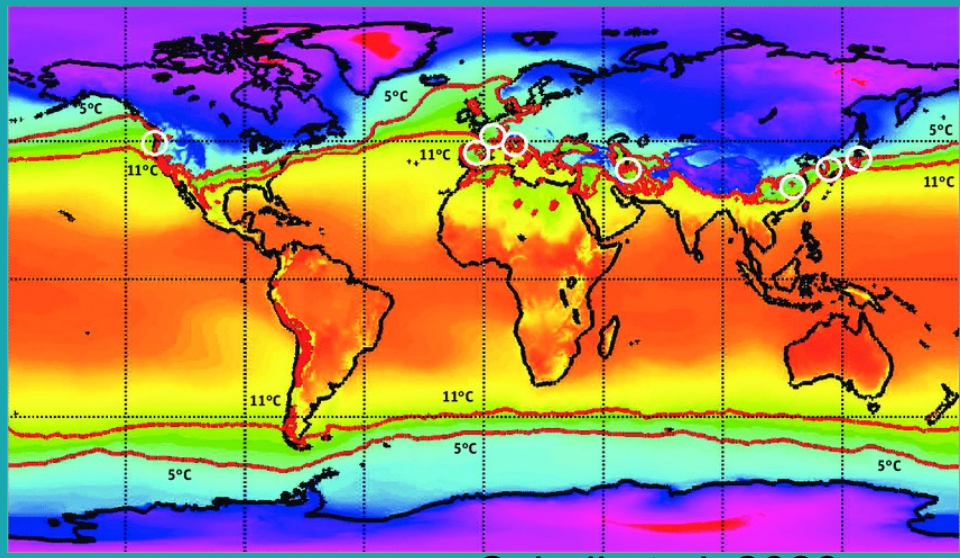
SPRING 2022

STATISTICAL PROGRAMMING FOR ECOLOGY, EVOLUTION, AND ENVIRONMENTAL SCIENCE

Species	One	Two	Red	Blue
Site				
1	8	2	45	499
2	0	5	5	23
3	78	0	51	0
4	1	6	1	0
5	5	44	0	0







Sajadi et al. 2020

How do I know if my data are 'big?'

- 1 M records = good to go, 1 M – 1 B = can work in R with some extra help, > 1 B = need to be analyzed with map reduce algorithms with help from Hadoop etc.

More practically…

- If R doesn't work for you because you have too much data
- What can get more difficult when data is big?
  - The data may not load into memory
  - Analyzing data may take a long time
  - Visualizations get messy

Okay, my data are too ‘big,’ what now?

- Check if you’re using 64-bit version of R
- Allocate more memory (if you have it) to R
- Reduce # of objects stored in memory

Still too big?

- Make data smaller
- Get a bigger computer
- Access data differently
- Split up the dataset for analysis

Adapted from Big Data 2015 Columbia

Make data smaller

- Run your analyses on a smaller chunk of your overall dataset to make sure it is indeed a memory or data size issue

Adapted from Big Data 2015 Columbia

Get a bigger computer

- Convince your supervisor to buy you a new computer

When that fails

- Rutgers High-Performance Clusters
  - Amarel (DEENR Node)
  - Annotate (SEBS; Windows & Linux; get in touch with Robert Muldowney)

Adapted from Big Data 2015 Columbia

Split up analyses

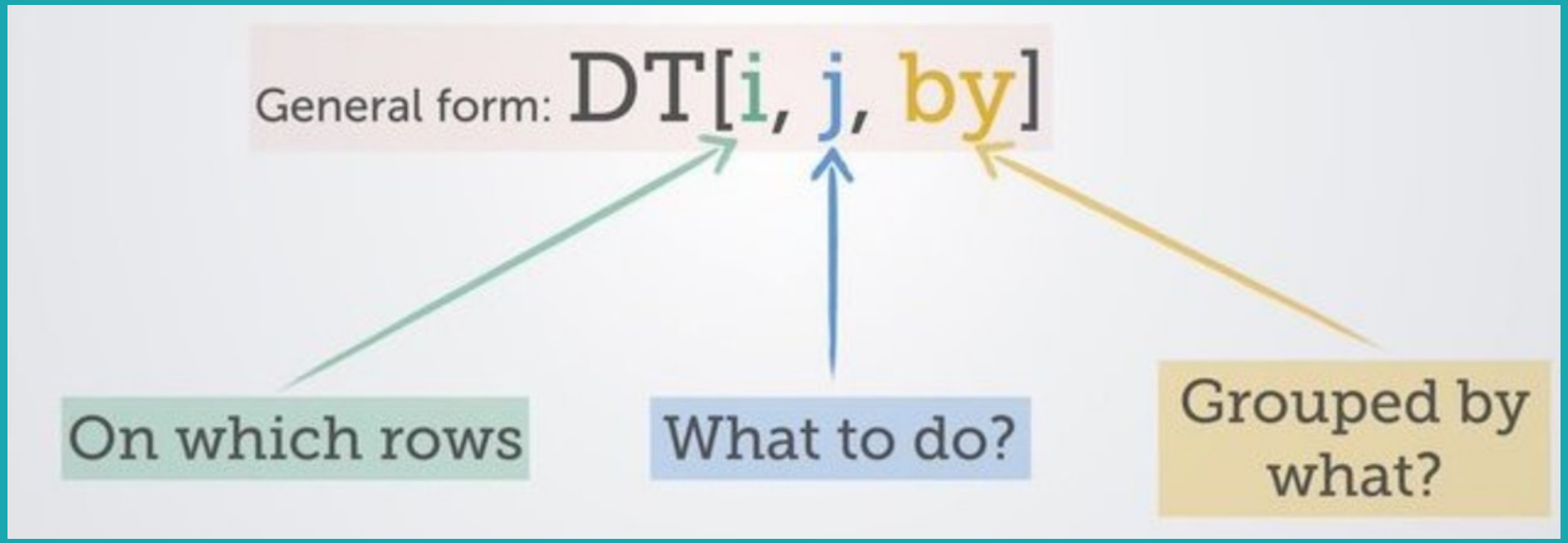
- Do analyses on *x* MB at a time
- Combine results
- Can use computing clusters to parallelize analysis:
  - Farming out subtasks to independent processors
  - MapReduce algorithms

Access data differently

- Use `**data.table**` package
  - Good for very large data files
  - High performance version of base R's data.frame
  - Offers fast subset, grouping, update, and joins
  - Makes it easy to turn an existing data frame into a data table

Adapted from Big Data 2015 Columbia

data.table



Resources

Wisconsin Data Science

Large Datasets and You

FasteR! HigheR! StrongeR! - A Guide to Speeding Up R Code for Busy People

Taking R to the Limit, Part II: Working with Large Datasets

R Bloggers: Big Data in R

CRAN-R: Intro to Data Table

CRAN-R: Keys in Data Table

CRAN-R: Assignment by Reference