Challenges and trends in really big insect collection datasets

Entomological Society of America
Big Data and Bugs Symposium
November 7, 2017

Katja C. Seltmann, PhD
Cheadle Center for Biodiversity and Ecological Restoration
University of California Santa Barbara
seltmann@ccber.ucsb.edu

NCEAS

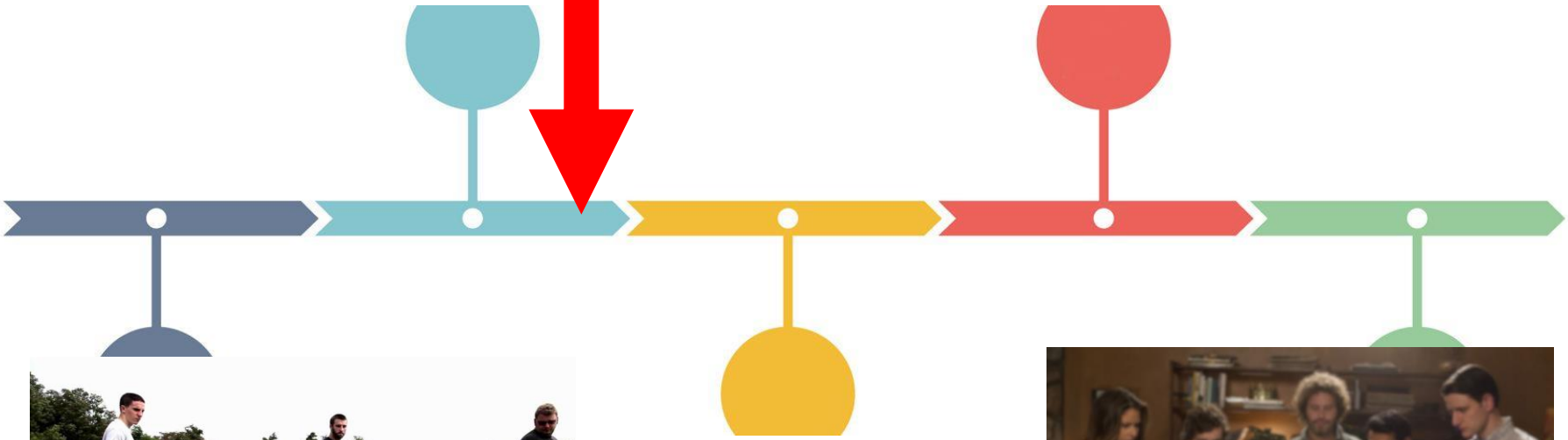
NSF

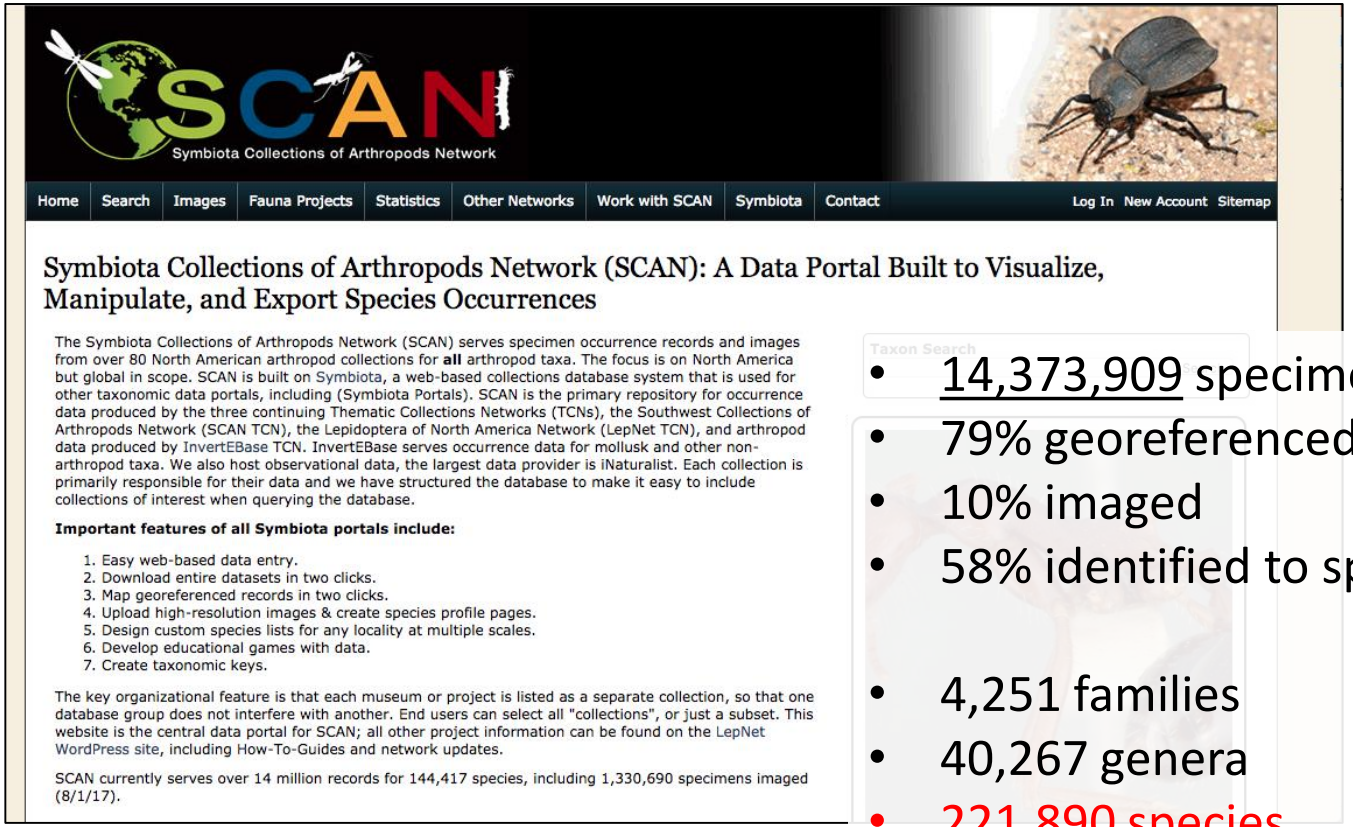
GBIF

iDigBio
Integrated Digitized Biocollections

Georeferencing for Research Use Workshop, Oct 4-7, 2016

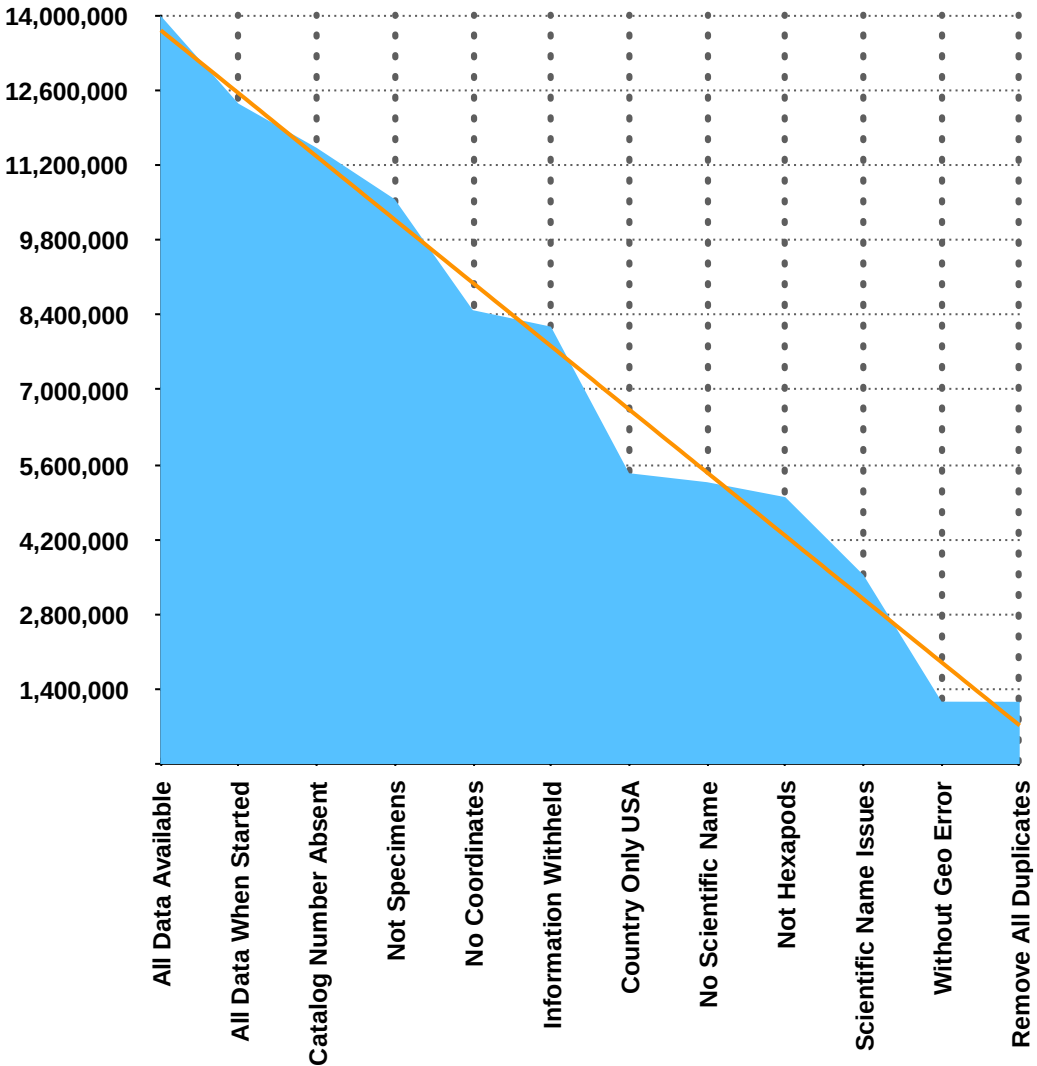
You are Here



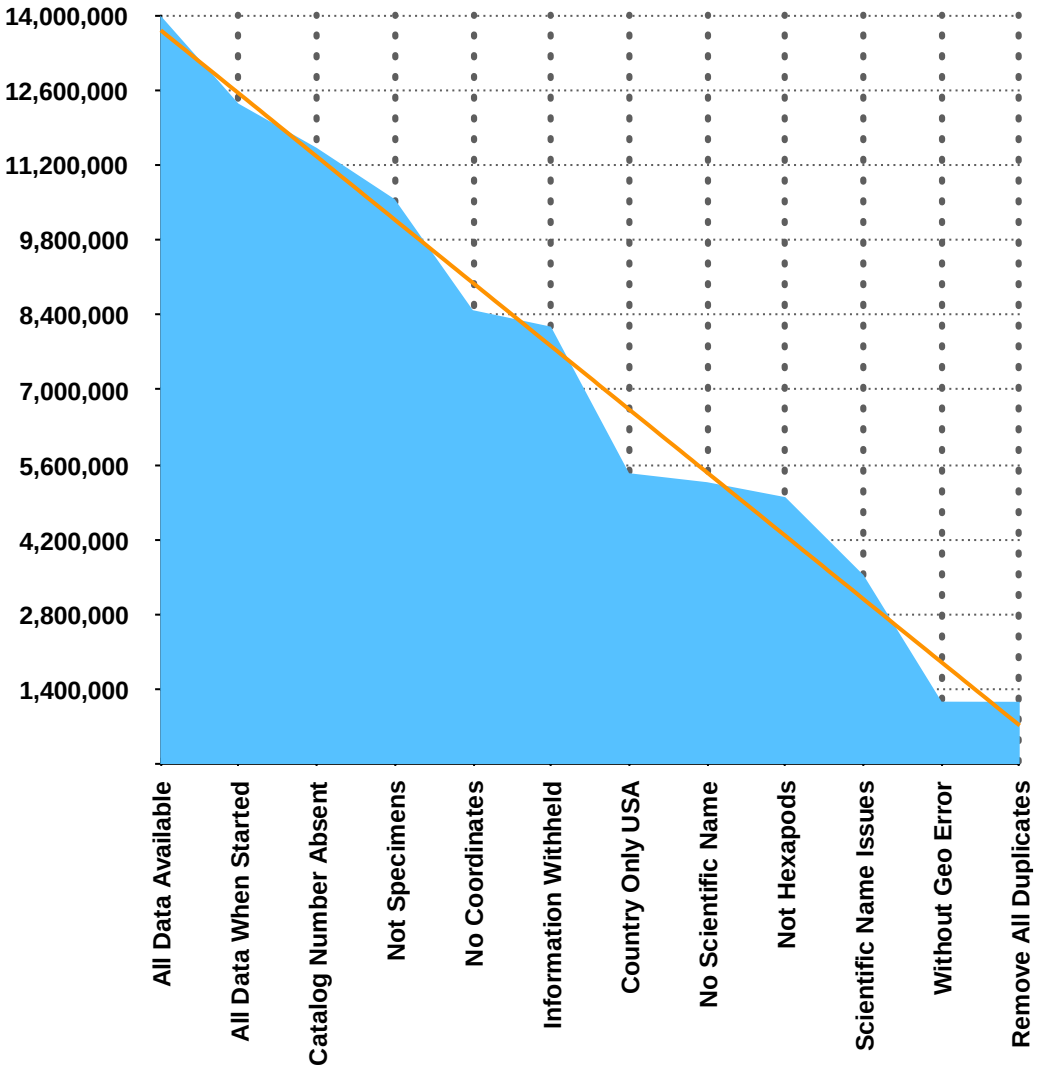


- 14,373,909 specimen records
- 79% georeferenced
- 10% imaged
- 58% identified to species

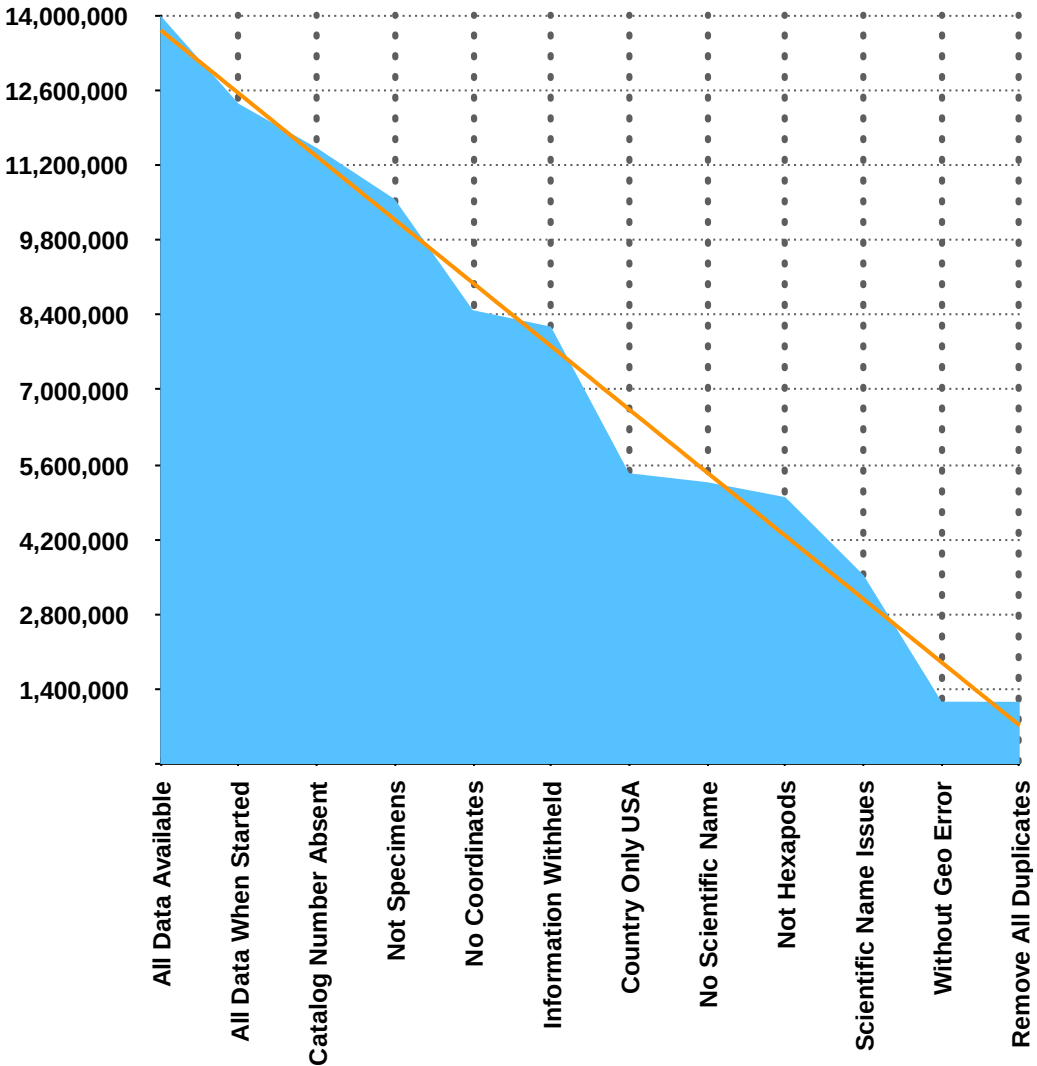
- 4,251 families
- 40,267 genera
- 221,890 species
- 228,206 total taxa (including subsp. and var.)



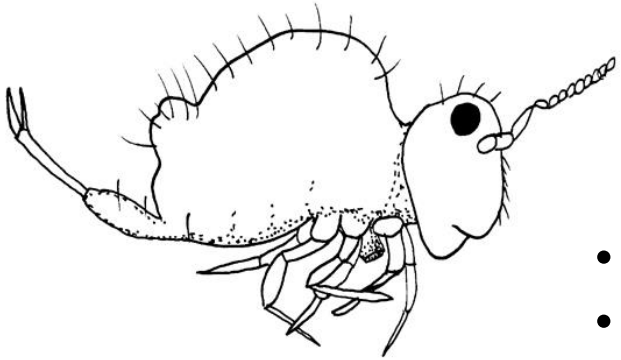
Actions Taken	*Data Available*
All Data Available	14,000,000
All Data When Started	12,347,785
Catalog Number Absent	11,523,828
Not Specimens	10,559,989
No Coordinates	8,485,230
Information Withheld	8,176,992
Not USA	5,435,585
No Scientific Name	5,268,055
Not Hexapods	4,989,736
Scientific Name Issues	3,532,798
Without Geo Error	1,162,978
Remove All Duplicates	1,161,064



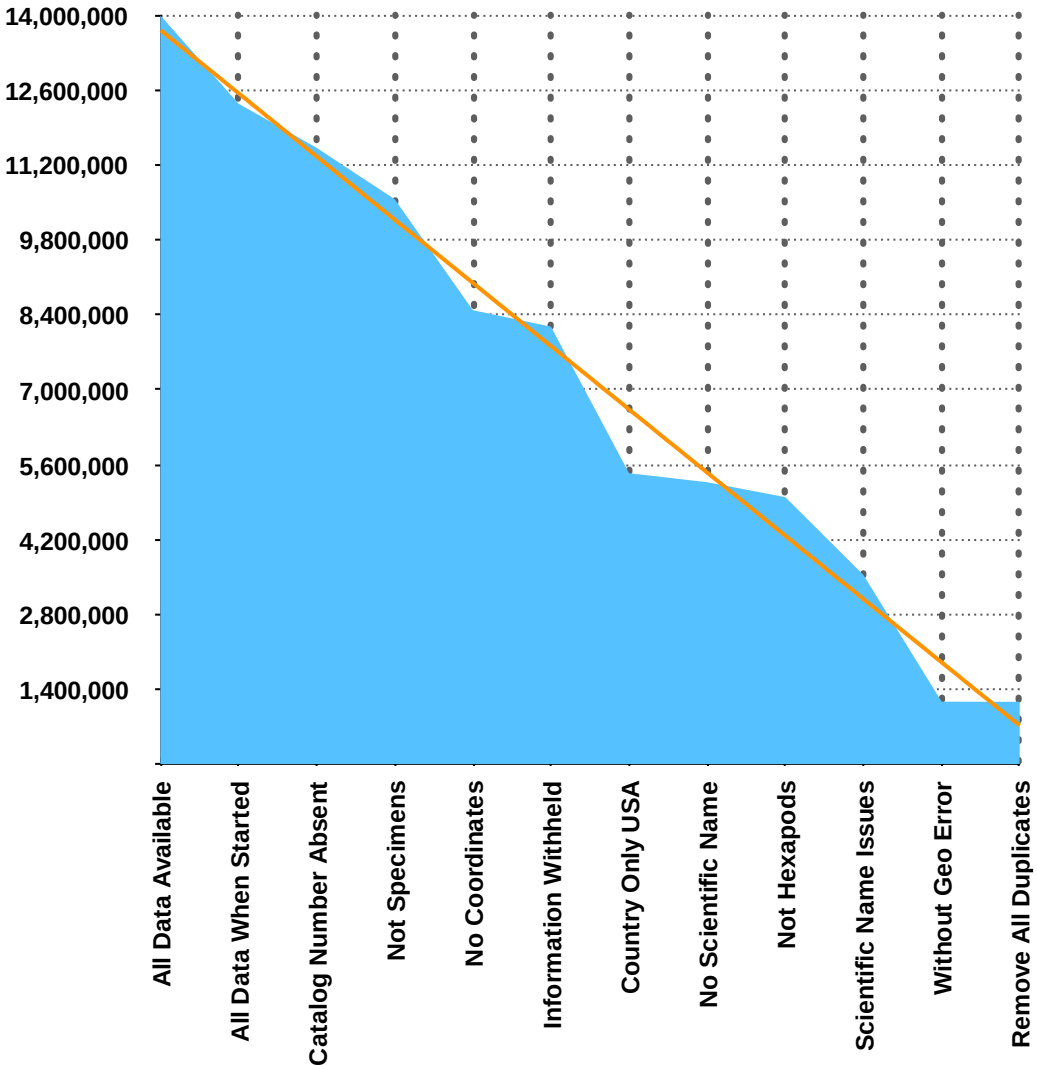
Actions Taken	*Data Available*
All Data Available	14,000,000
All Data When Started	12,347,785
Catalog Number Absent	11,523,828
Not Specimens	10,559,989
No Coordinates	8,485,230
Information Withheld	8,176,992
Not USA	5,435,585
No Scientific Name	5,268,055
Not Hexapods	4,989,736
Scientific Name Issues	3,532,798
Without Geo Error	1,162,978
Remove All Duplicates	1,161,064



Actions Taken	*Data Available*
All Data Available	14,000,000
All Data When Started	12,347,785
Catalog Number Absent	11,523,828
Not Specimens	10,559,989
No Coordinates	8,485,230
Information Withheld	8,176,992
Not USA	5,435,585
No Scientific Name	5,268,055
Not Hexapods	4,989,736
Scientific Name Issues	3,532,798
Without Geo Error	1,162,978
Remove All Duplicates	1,161,064



- Hexapoda
- USA
- Need Coordinates
- Need Scientific Names
- Museum Specimens

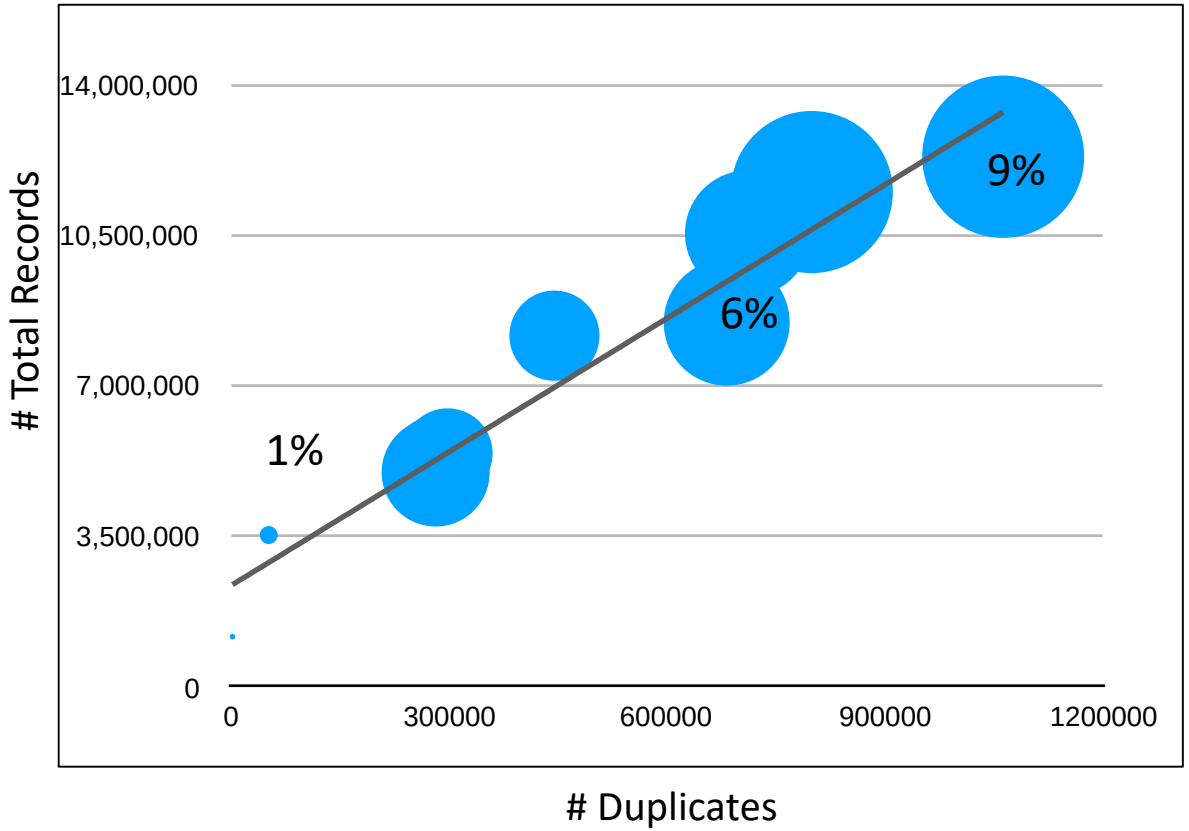


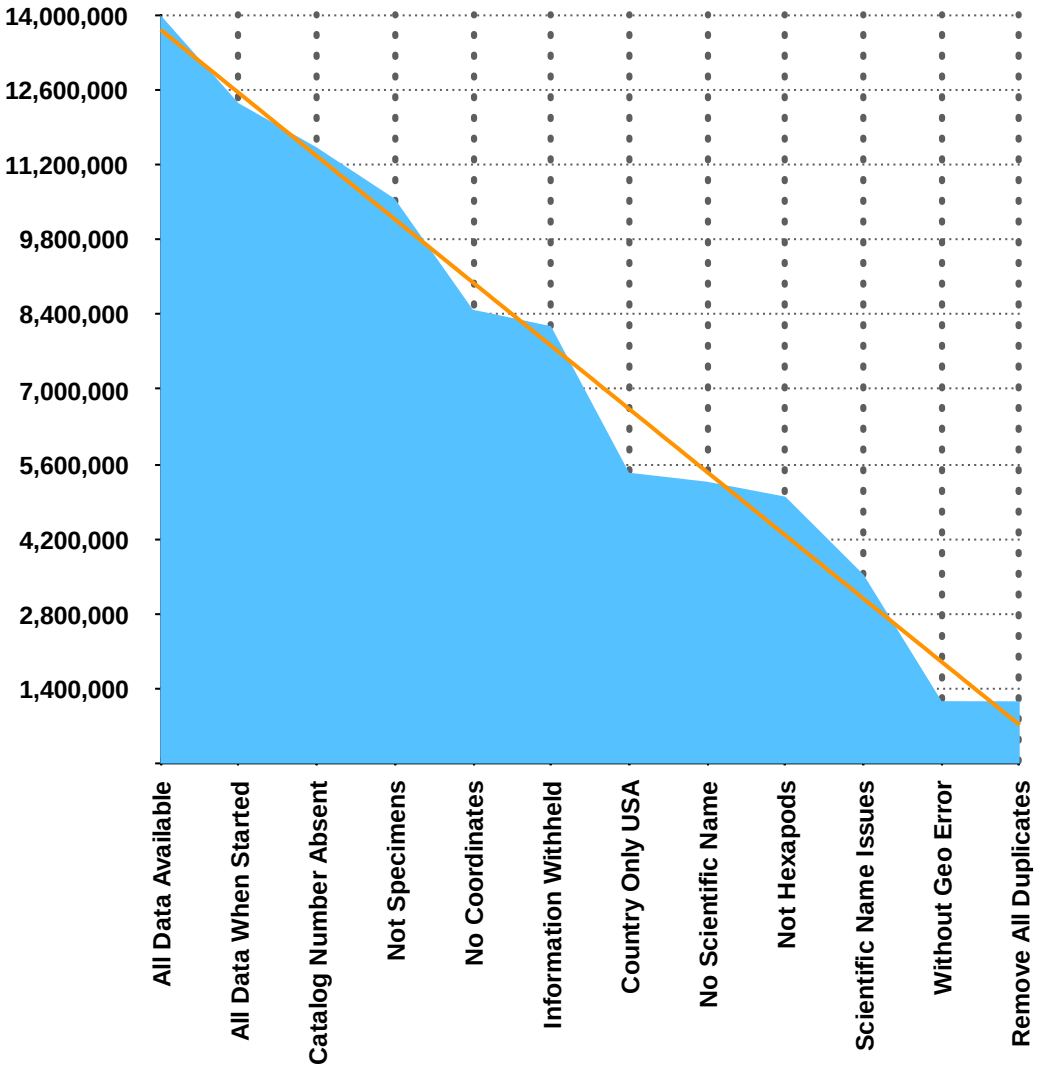
Actions Taken	*Data Available*
All Data Available	14,000,000
All Data When Started	12,347,785
Catalog Number Absent	11,523,828
Not Specimens	10,559,989
No Coordinates	8,485,230
Information Withheld	8,176,992
Not USA	5,435,585
No Scientific Name	5,268,055
Not Hexapods	4,989,736
Scientific Name Issues	3,532,798
Without Geo Error	1,162,978
Remove All Duplicates	1,161,064

Alpha Numeric Catalog Numbers

1,061,954 duplicates

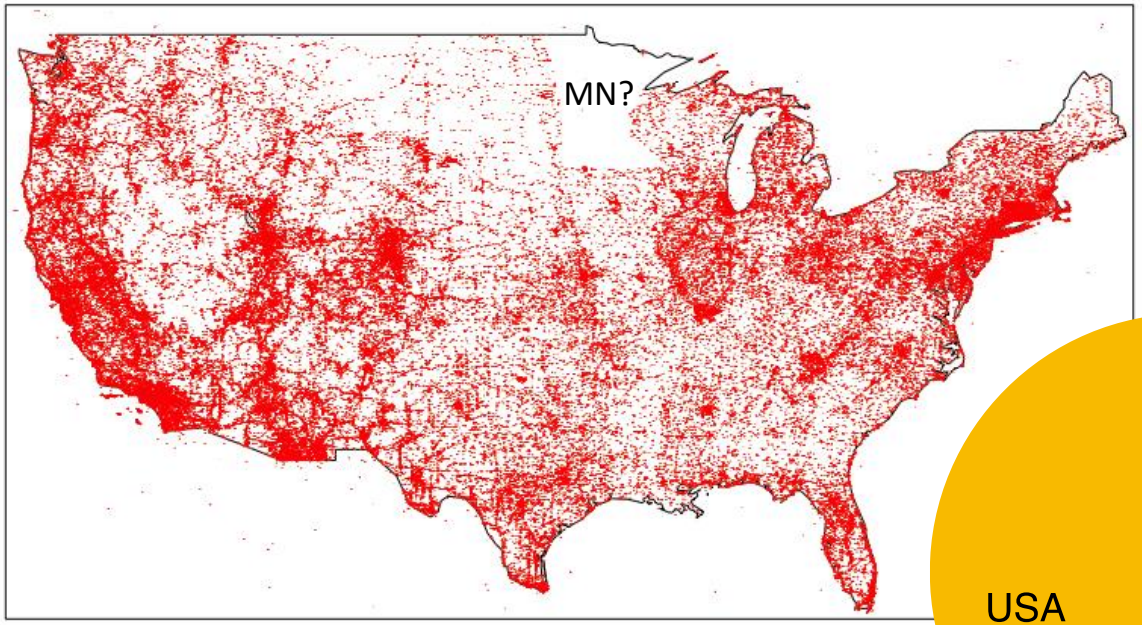
=491 days to delete

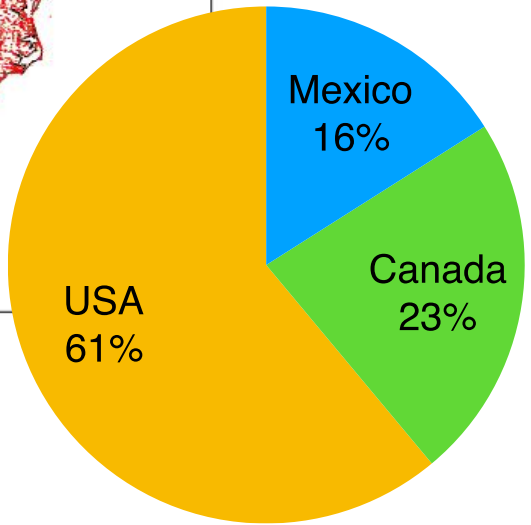




Actions Taken	*Data Available*
All Data Available	14,000,000
All Data When Started	12,347,785
Catalog Number Absent	11,523,828
Not Specimens	10,559,989
No Coordinates	8,485,230
Information Withheld	8,176,992
Not USA	5,435,585
No Scientific Name	5,268,055
Not Hexapods	4,989,736
Scientific Name Issues	3,532,798
Without Geo Error	1,162,978
Remove All Duplicates	1,161,064

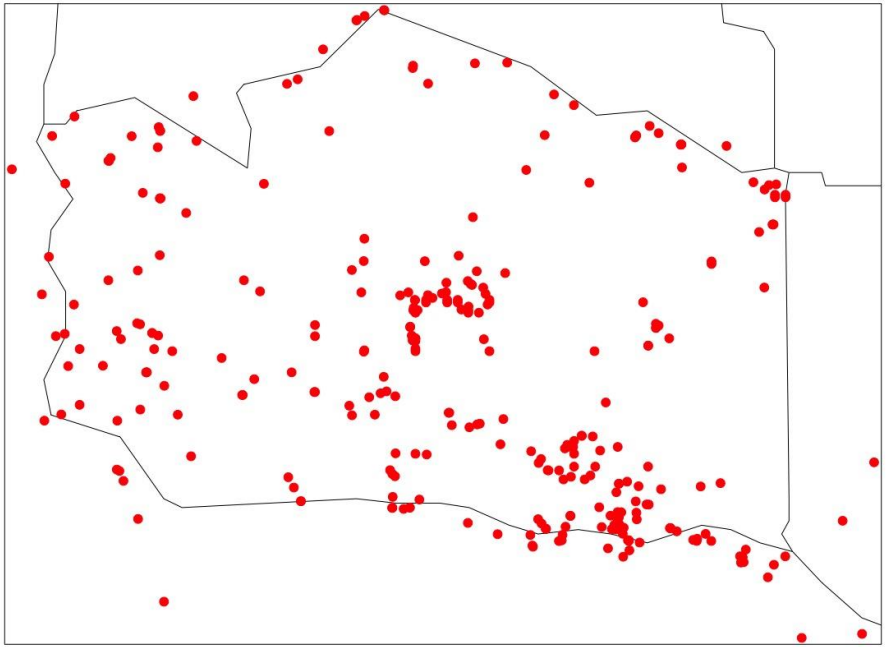
No Coordinates
Information Withheld
Without Geo Error

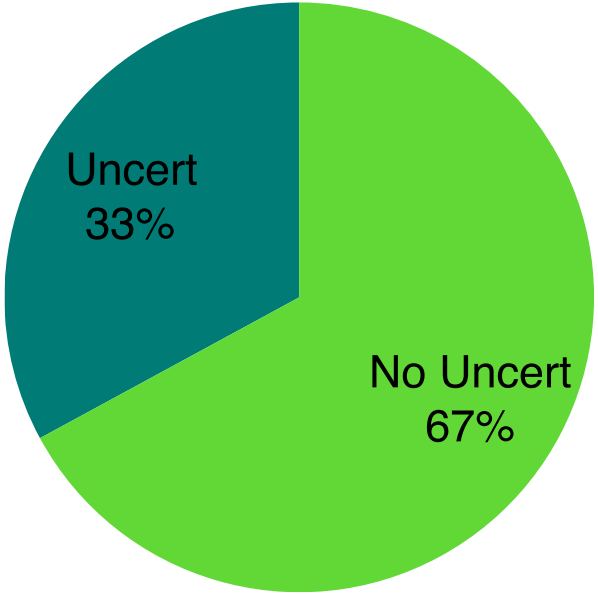




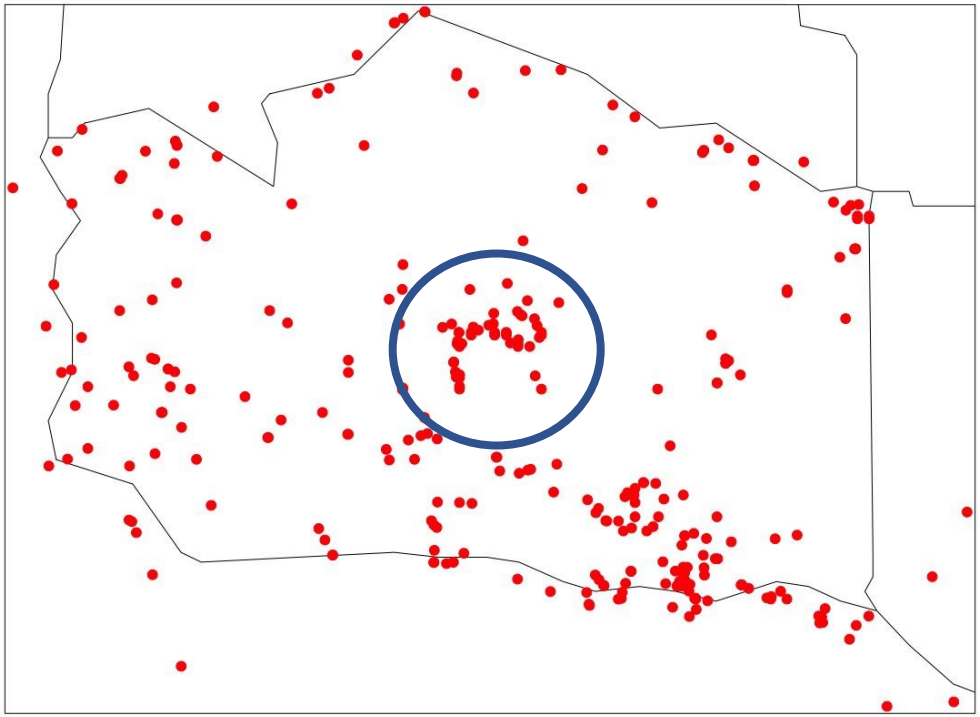
Term Name: coordinateUncertaintyInMeters	
Identifier:	http://rs.tdwg.org/dwc/terms/coordinateUncertaintyInMeters
Class:	http://purl.org/dc/terms/Location
Definition:	The horizontal distance (in meters) from the given decimalLatitude and value empty if the uncertainty is unknown, cannot be estimated, or is n
Comment:	Examples: "30" (reasonable lower limit of a GPS reading under good co having 100 meter precision and a known spatial reference system). For
Details:	coordinateUncertaintyInMeters
Term Name: coordinatePrecision	
Identifier:	http://rs.tdwg.org/dwc/terms/coordinatePrecision
Class:	http://purl.org/dc/terms/Location
Definition:	A decimal representation of the precision of the coordinates given in the
Comment:	Examples: "0.00001" (normal GPS limit for decimal degrees), "0.0002 http://terms.tdwg.org/wiki/dwc:coordinatePrecision
Details:	coordinatePrecision
Term Name: pointRadiusSpatialFit	
Identifier:	http://rs.tdwg.org/dwc/terms/pointRadiusSpatialFit
Class:	http://purl.org/dc/terms/Location
Definition:	The ratio of the area of the point-radius (decimalLatitude, decimalLongi representation of the Location. Legal values are 0, greater than or equa the given point-radius does not completely contain the original represer representation is a point without uncertainty and the given georeferenc the same point, the pointRadiusSpatialFit is 1.
Comment:	Detailed explanations with graphical examples can be found in the "Guic http://terms.tdwg.org/wiki/dwc:pointRadiusSpatialFit
Details:	pointRadiusSpatialFit

No Coordinates
Information Withheld
Without Geo Error





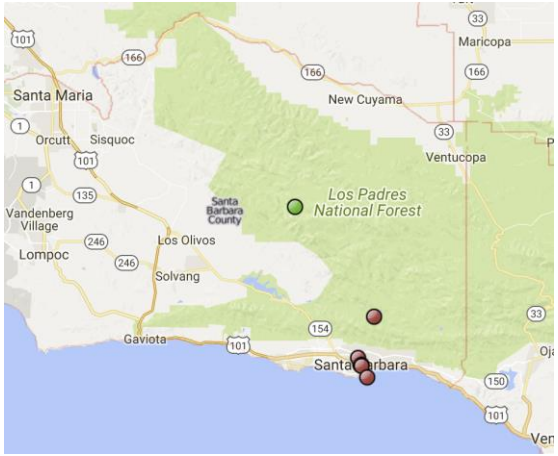
No Coordinates Information Withheld Without Geo Error



The problem with centroids without error

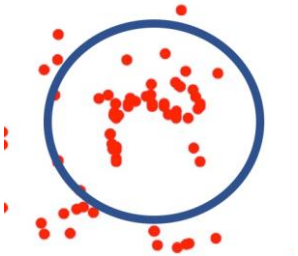
No Coordinates Information Withheld
Without Geo Error

34.726867, -119.806743



34.733639, -119.856181

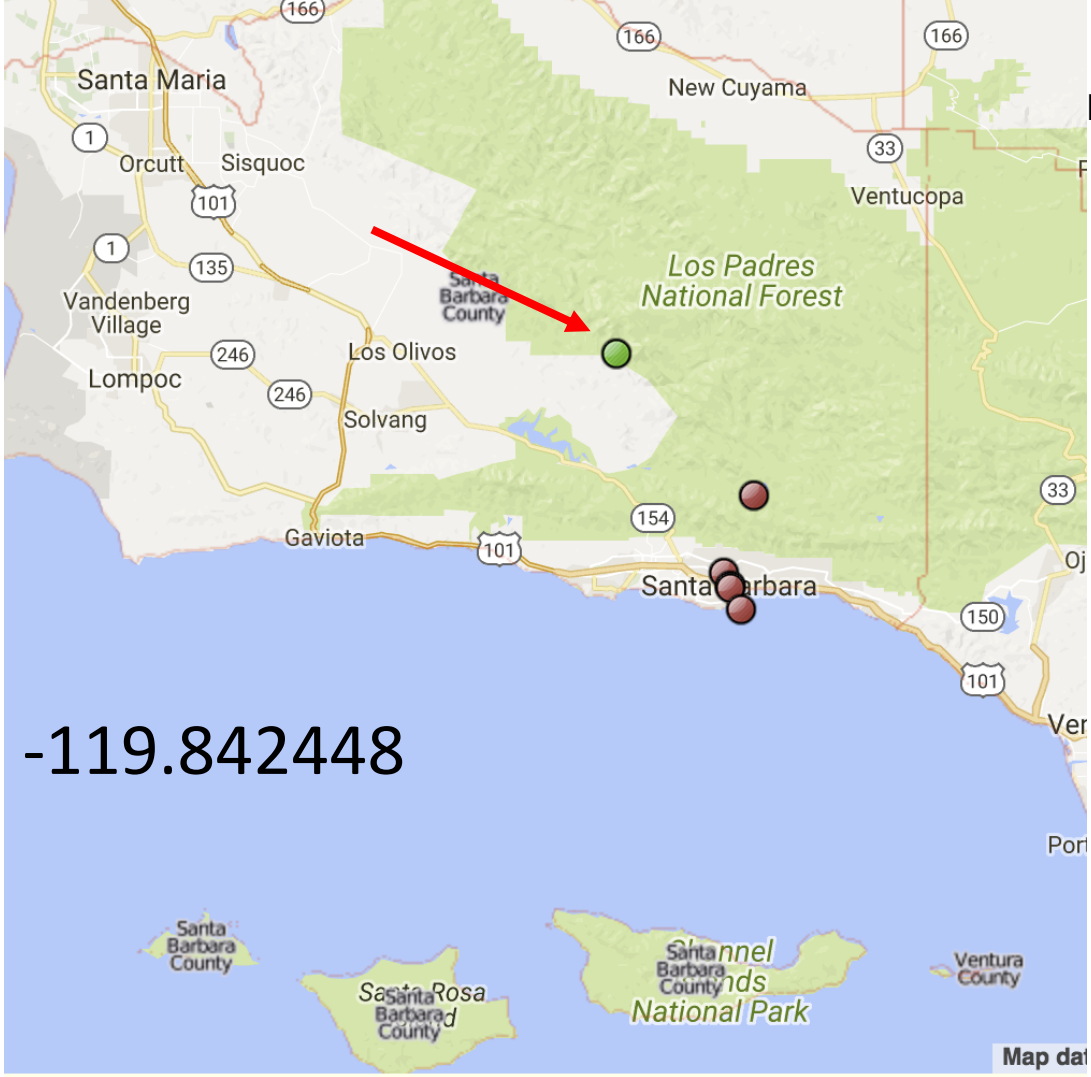
34.76072, -119.894633



Centroid method of furthest point from the boundary of a county

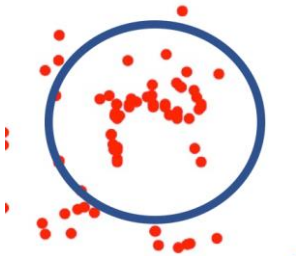
Geographic center (centroid)

34.665896, -119.842448



No Coordinates Information Withheld

Without Geo Error



How do we get to Bigger Data?

At the Data Aggregator Level

- TDWG Data Quality Working Group
- Data quality across data providers

At the Research Level

- Understand present data cleaning considerations for large data.
  - Need new methods and repeatable practices
  - Calculate a hole in the middle that is likely the radius of a centroid and delete all records without an uncertainty

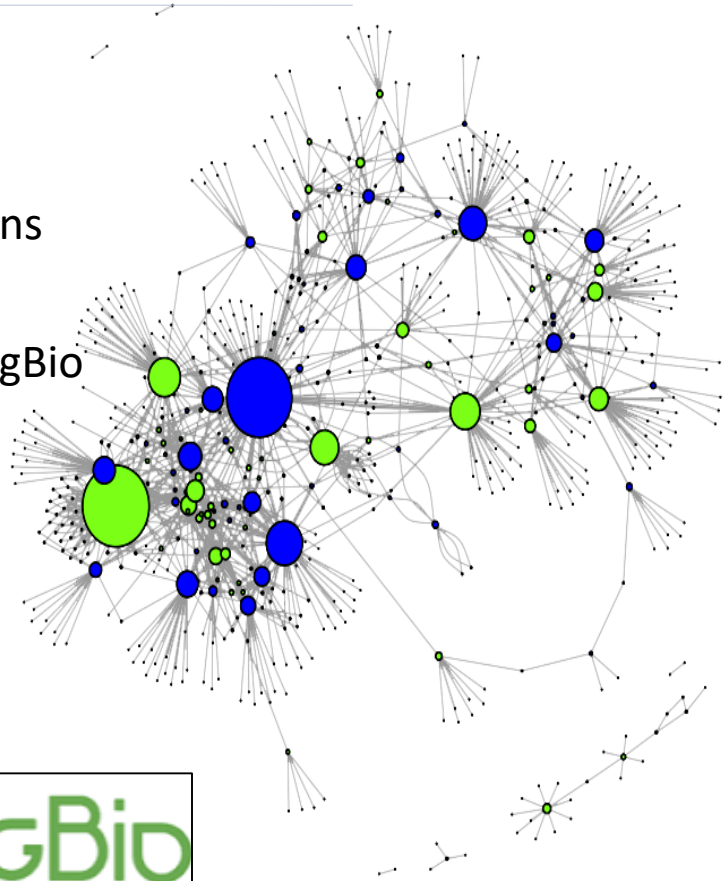
At the Data Provider Level

- Understand research practices
  - Provide extant for georeferences
  - Consider your data in the downstream context

When in doubt, we throw it out!

Acknowledgements

- Cheadle Center for Biodiversity
- NSF: Tri-Trophic TCN project PIs, digitizers, and collections
- Museum collections, herbaria, and curators worldwide
- Neil Cobb, Benjamin Brandt, Deb Paul, Jorrit Poelen, iDigBio and many others



Katja C. Seltmann, PhD
Cheadle Center for Biodiversity and Ecological Restoration
seltmann@ccber.ucsb.edu