

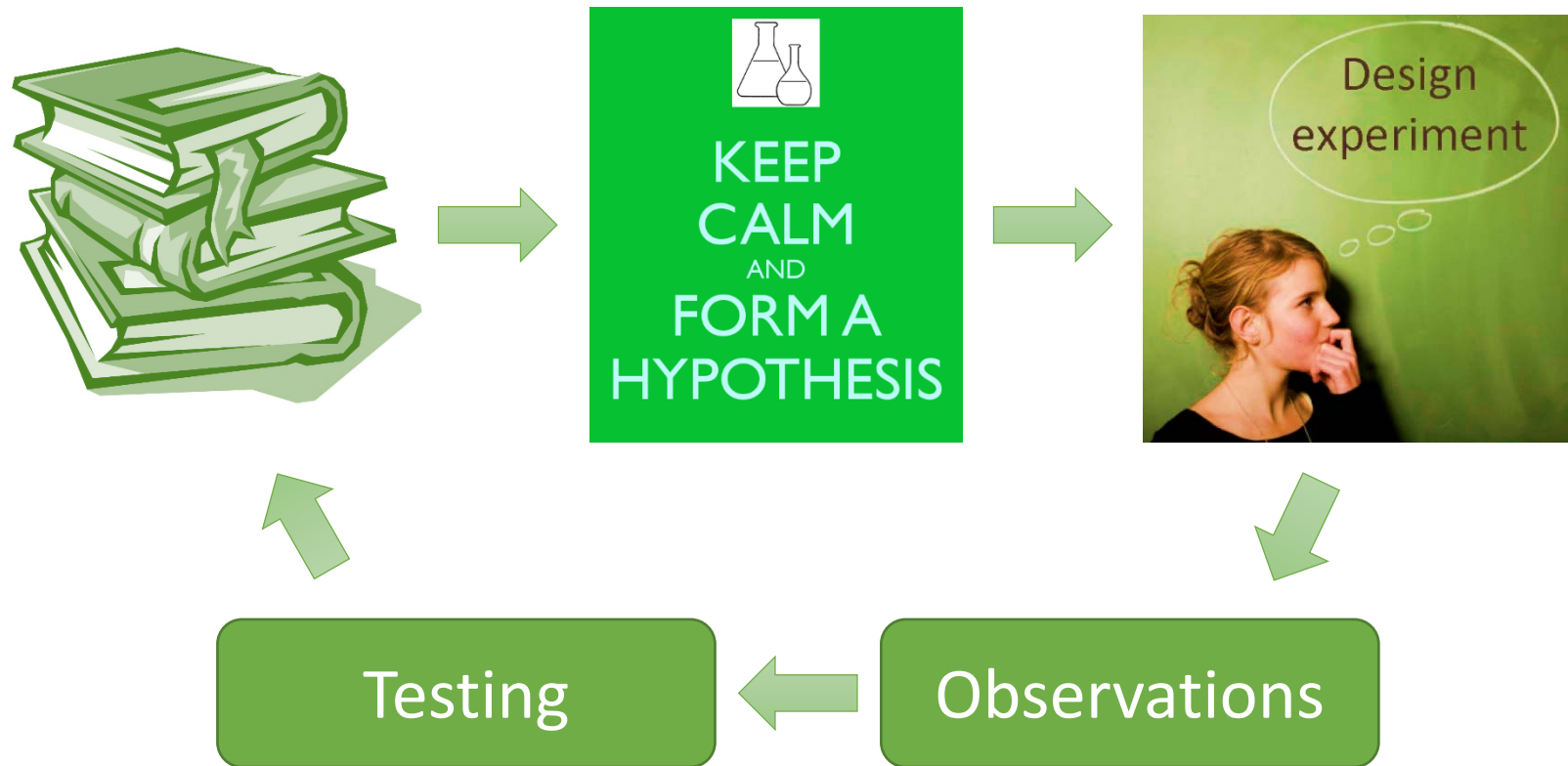
A group of approximately 15 people, including men and women of various ages, are standing in a line in a grassy field. In the background, there are trees and a large, snow-capped mountain under a cloudy sky. The image is semi-transparent, serving as a background for the text.

An infrastructure to innovate data intensive science

Jie Zhang, Data manager of KiLi Project
University of Würzburg, jie.zhang@uni-wuerzburg.de

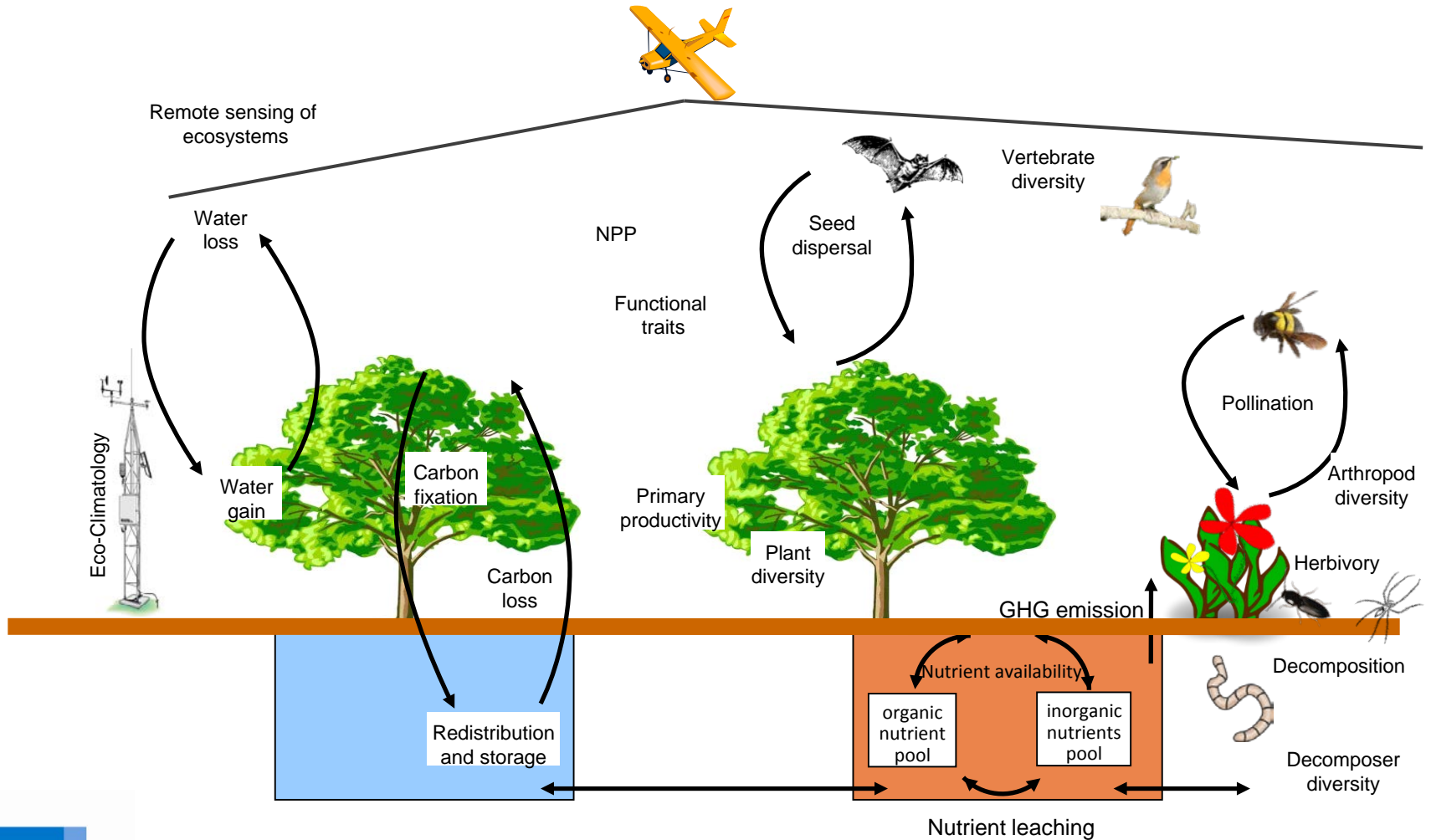


What do we usually do in Ecology

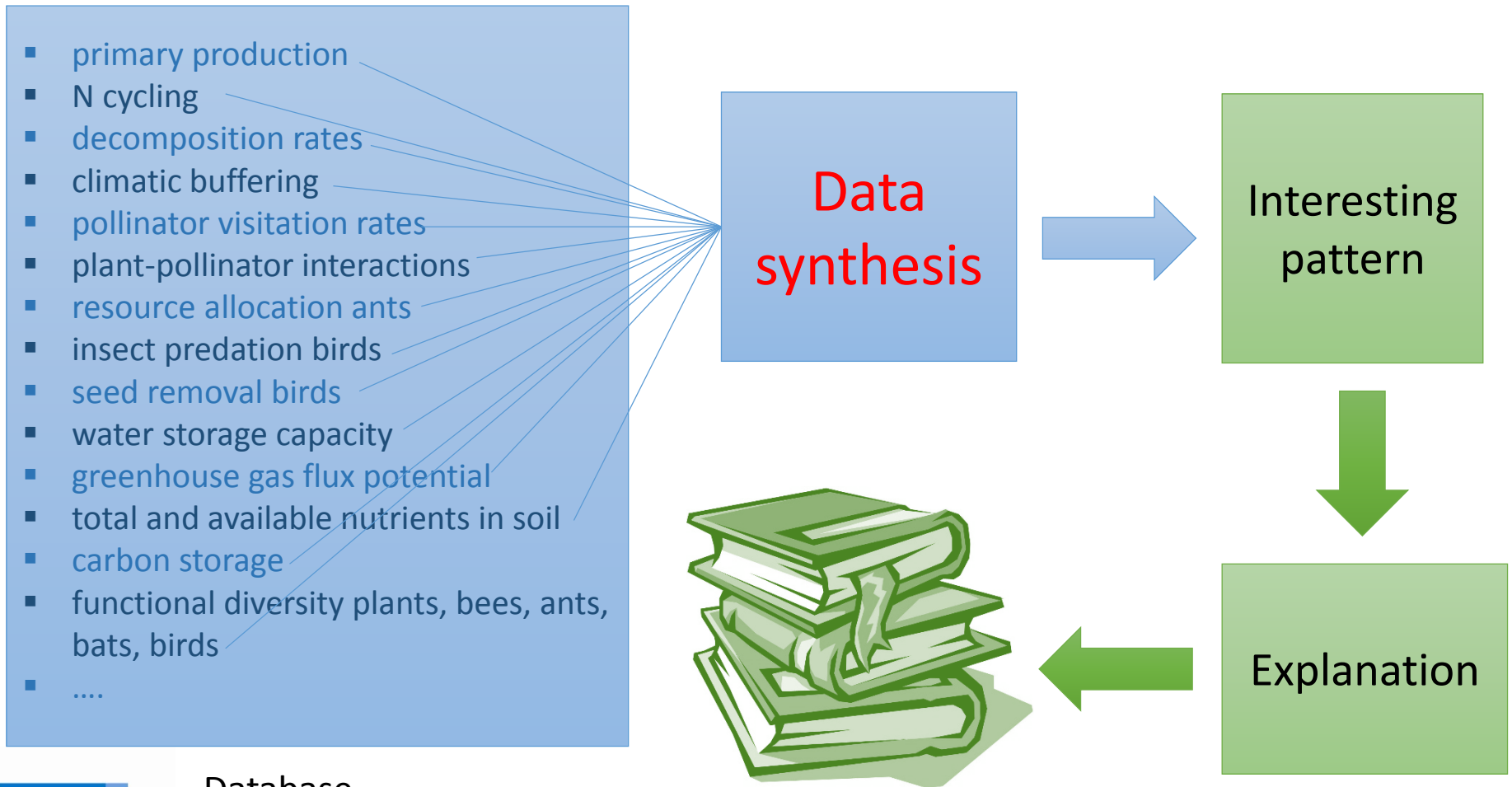


One student, one experiment, one result

What is actually needed in Ecology



A new paradigm in Ecology – data driven approach

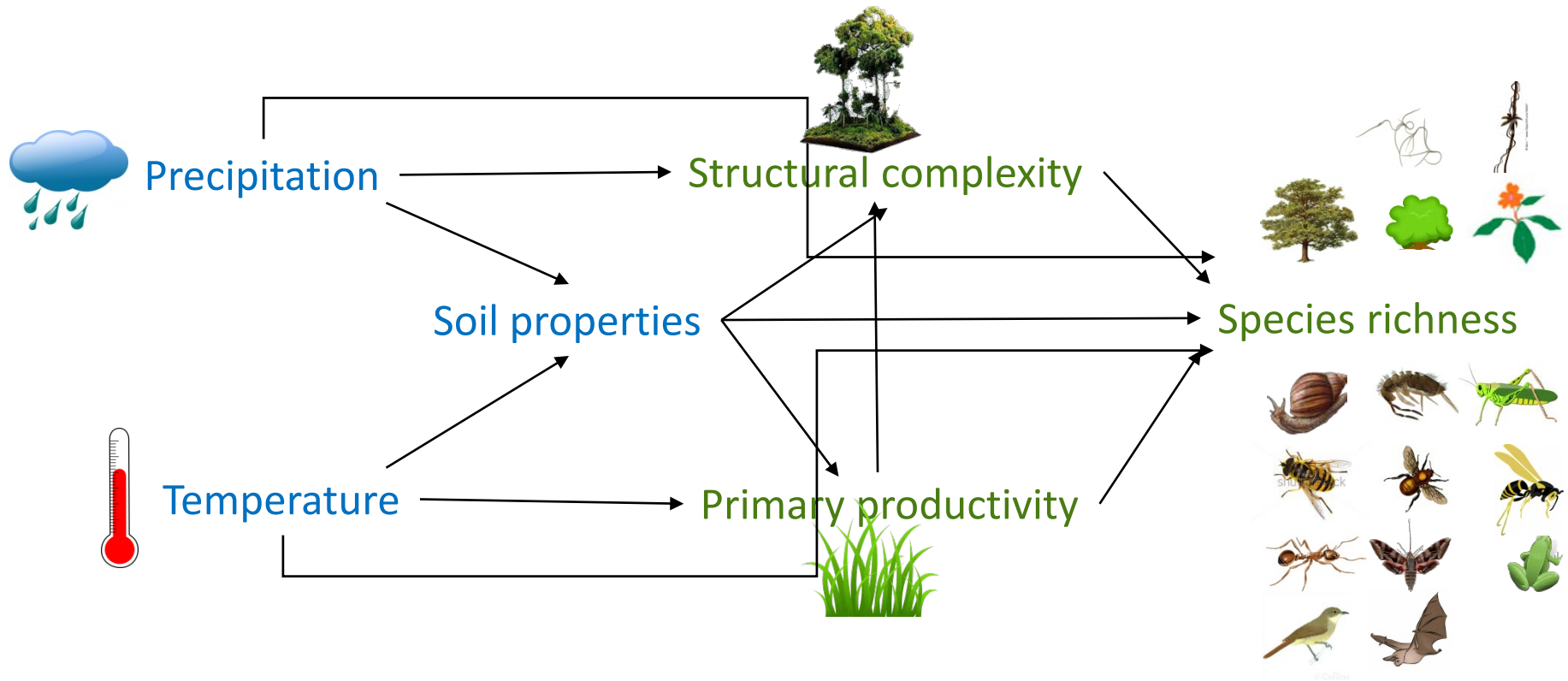


Database



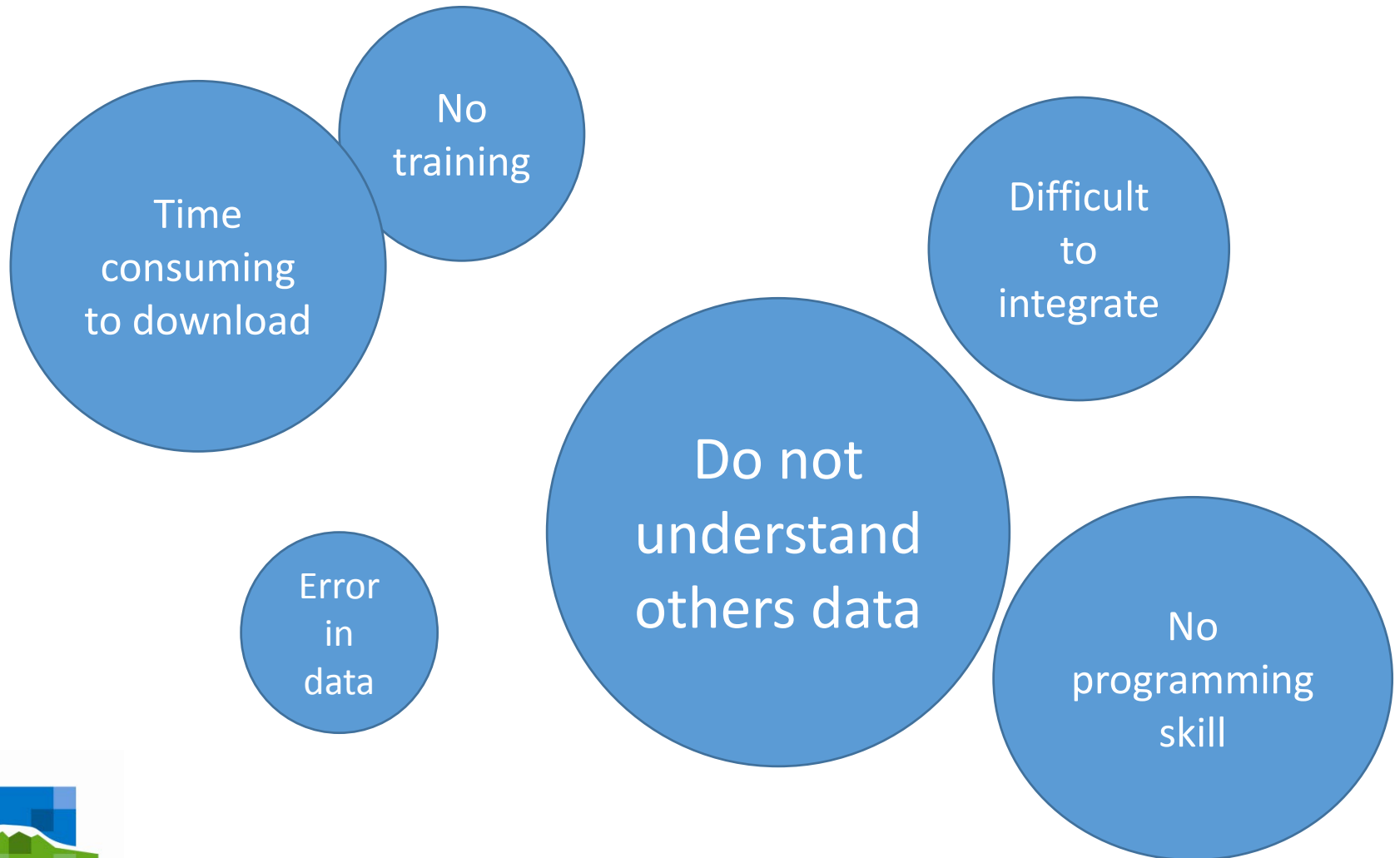
One example

Patterns in species richness:



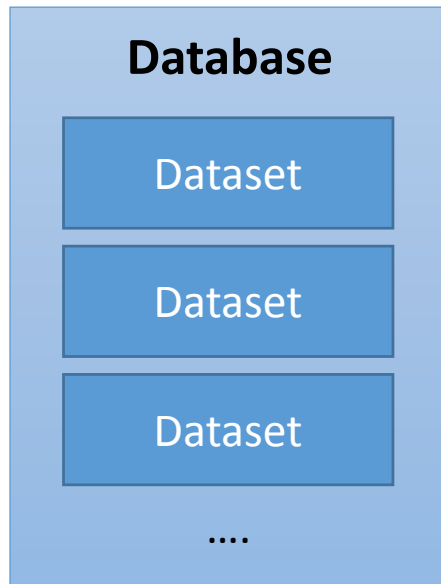


The difficulties we all share



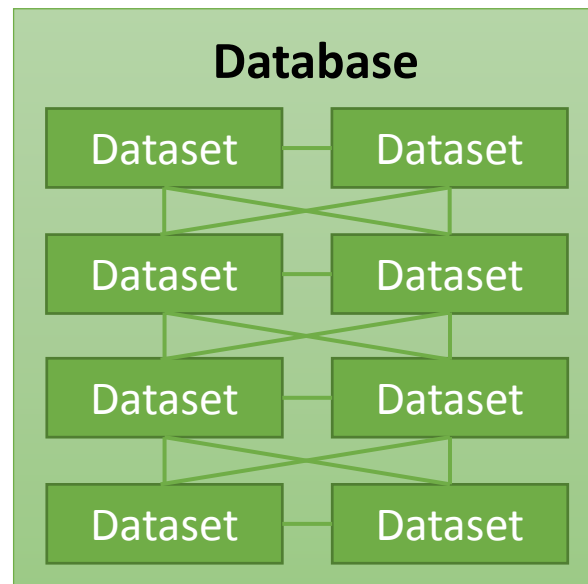
The new paradigm for database – data intensive database

Back in the days



Data storage

Nowadays

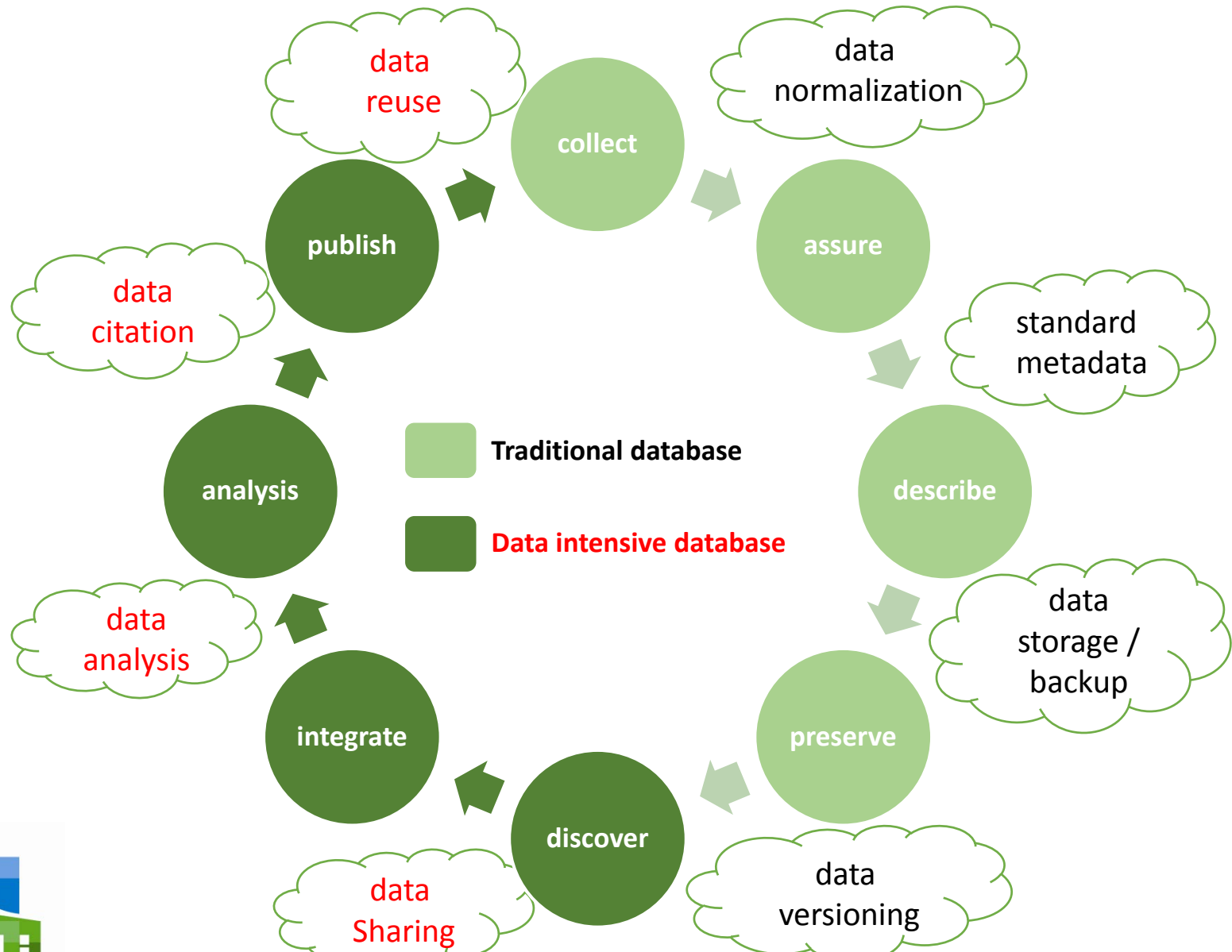


Innovate
collaboration

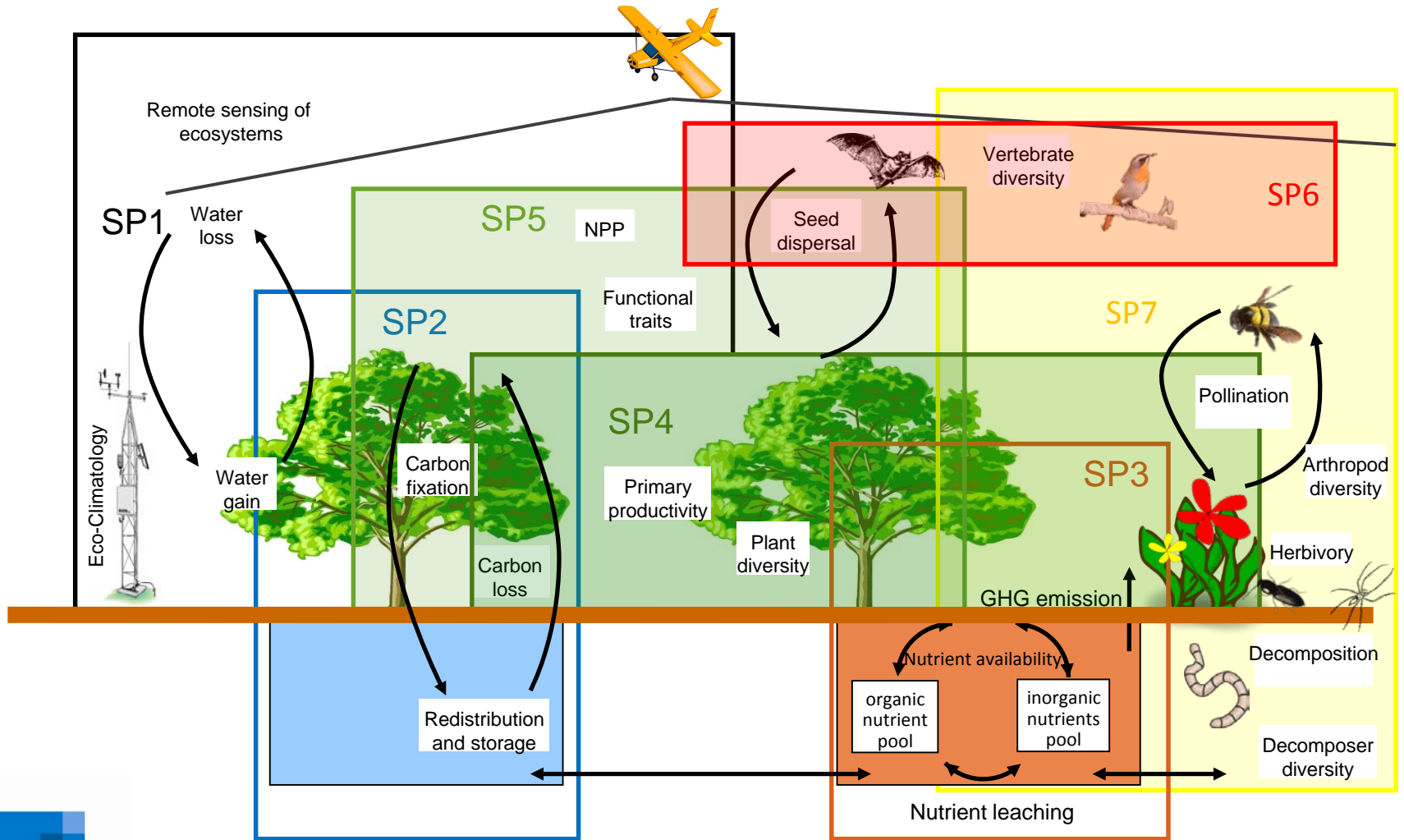
- Data sharing
- Data exchange
- Data merging
- Data synthesis



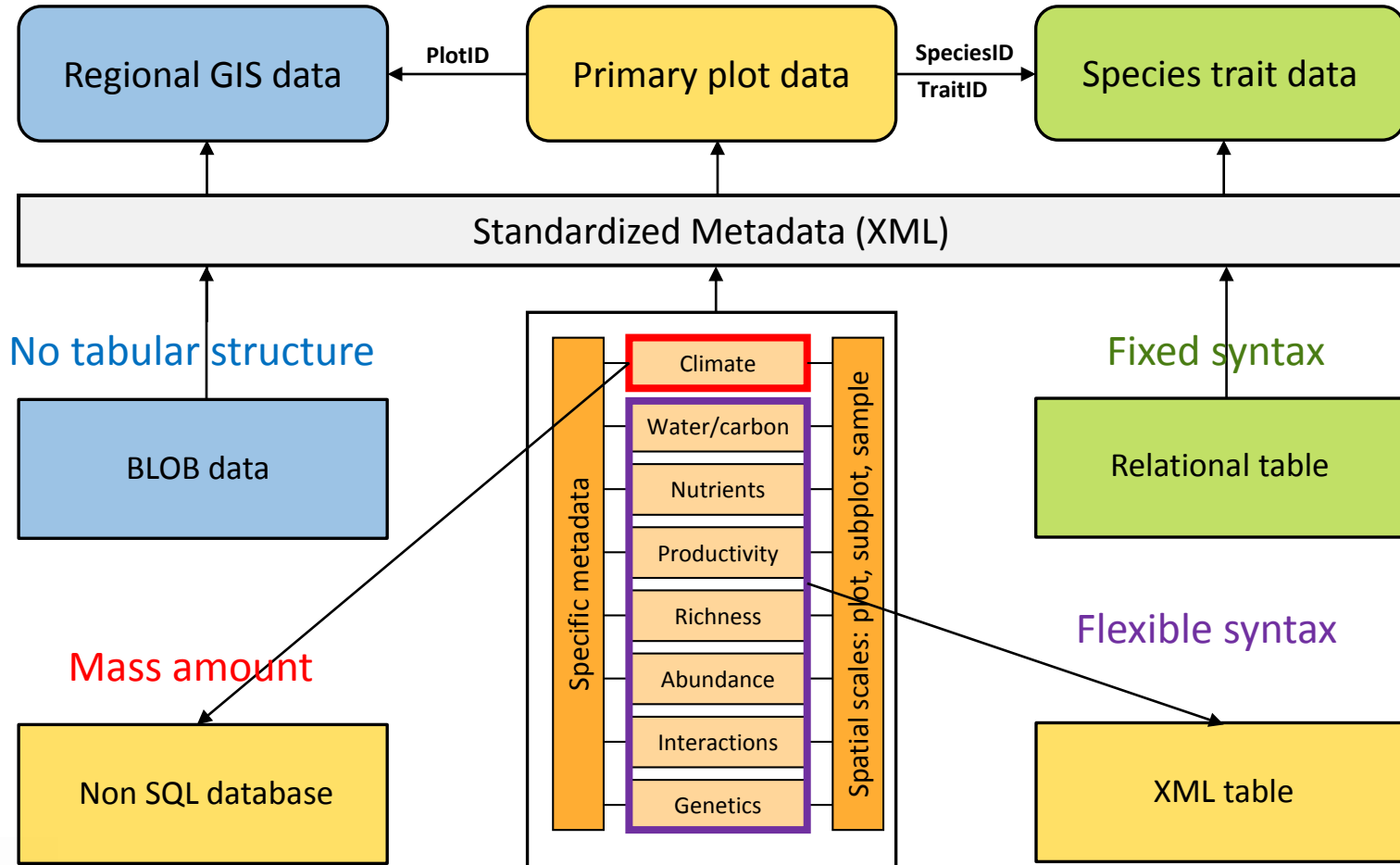
Ecological data cycle

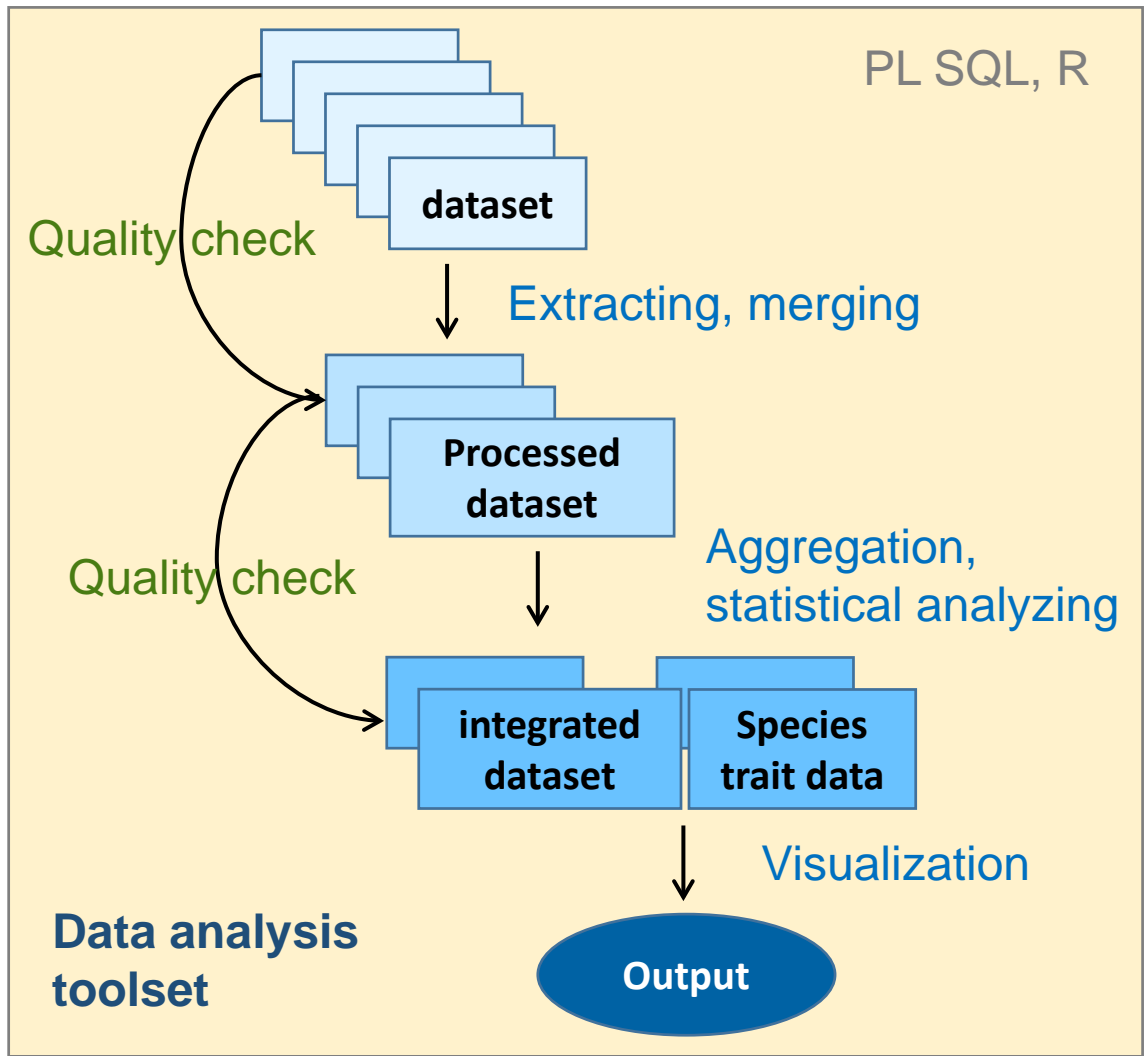


Best practice: KiLi Project



Best practice: KiLi Project





New mission for next generation database:





- Detect **outlier**
- Detect **duplicate**
- Display error table
- Error report to owner

Choose the type of your quality check: Outlier detection

Outlier detection

species_occurrences < 5 Define outliers

species_occurrences < 5

Number of outliers: 17

Outlier table:

obsld	PlotID	date	h2o	nacl	su	glu	suglu	oil	species
64890	flm6	10.11.2011 00:00:00	0	1	2	0	1	0	4
64888	flm4	10.03.2011 00:00:00	0	0	0	0	1	2	3
64886	flm2	30.05.2012 00:00:00	0	0	0	0	0	2	2
64882	mai3	16.12.2011 00:00:00	1	0	2	0	0	0	3
64913	fod2		0	0	0	0	0	0	0
64910	foc1	05.11.2011 00:00:00	0	0	0	0	0	0	0
64911	foc2	05.11.2011 00:00:00	0	0	0	0	0	0	0
64912	fod1	02.11.2012 00:00:00	0	0	0	0	0	0	0
64914	fod3	24.10.2012 00:00:00	1	0	0	0	0	0	1
64915	fpo1	07.11.2012 00:00:00	0	0	0	0	0	0	0

Show rows: 10 Page 1

Send outlier report to data owners

Choose the type of your quality check: Duplicate detection

Duplicate detection

Choose a combination of the following variables which should be unique for each observation:

obsld
 PlotID
 date
 h2o
 nacl
 su
 glu
 suglu
 oil
 species_occurrences
 species_richness

Detect duplicates

Number of duplicated rows: 4

Duplicated rows based on unique columns combination:

obsld	PlotID	date	h2o	nacl	su	glu	suglu	oil	species_occurrences	species_richness
64910	foc1	04.11.2011 23:00:00	0	0	0	0	0	0	0	0
64911	foc2	04.11.2011 23:00:00	0	0	0	0	0	0	0	0
64916	fpo4	23.10.2012 22:00:00	0	0	0	0	0	0	0	0
64917	fpd4	23.10.2012 22:00:00	0	0	0	0	0	0	0	0

Send duplicate report to data owners





Method Column to aggregate Aggregate Columns

MAX	h2o	>> <<	MAX species_occurences
MIN	nacl		AVG species_richness
AVG	su		
SUM	glu		
	suglu		
	oil		

Based on unique value of column or column combinations:

PlotID	>> <<	PlotID
date		date
h2o		
nacl		
su		
glu		

Apply aggregation

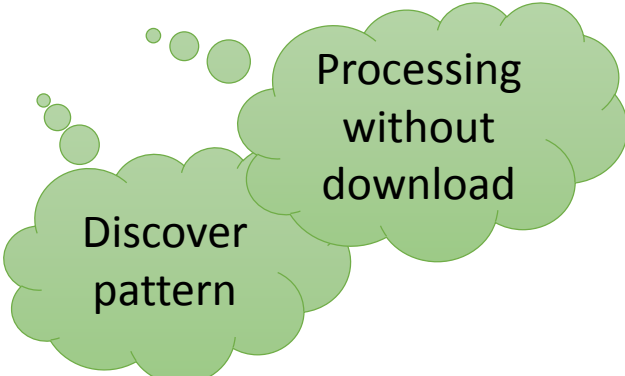
Aggregated table:

No. of rows: 47

obsId	PlotID	date	species_occurences	species_richness
64896	cof1	23.02.2011 23:00:00	15	4

Aggregation:

- Max, min, avg, sum
- Aggregate on any column or combination
- Display agg table
- Export agg table



CP2_Ants_bait_data

Merge Dataset Variable to merge Selected variables

Climate

obsId
PlotID
Avg_Precipitation
P_Source
Avg_RelativeHumidity
RH_Source

>>

Avg_Temperature
Avg_Precipitation

<<

Apply Merge

Merged table:

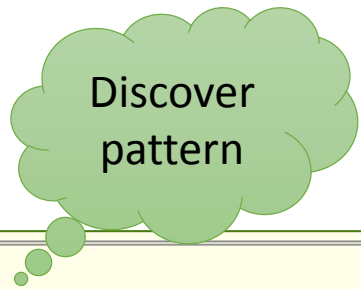
x.obsId	x.PlotID	x.date	x.h2o	x.nacl	x.su	x.glu	x.suglu	x.oil	x.species_occurrences	x.species_richness	y.Avg_Temperature	y.Avg_Precipitation
64899	cof4	22.02.2011 23:00:00	4	1	5	3	6	6	25	7	22,5965	1115,14813
64897	cof2	02.11.2011 23:00:00	0	1	1	1	1	3	7	3	19,5925	1518,028774

Merge database:

- Dataset is linked by PlotID, TraitID, SpeciesID
- Merge dataset based on shard ID.

Data
integration
without
download





• Descriptive statistics

Filter

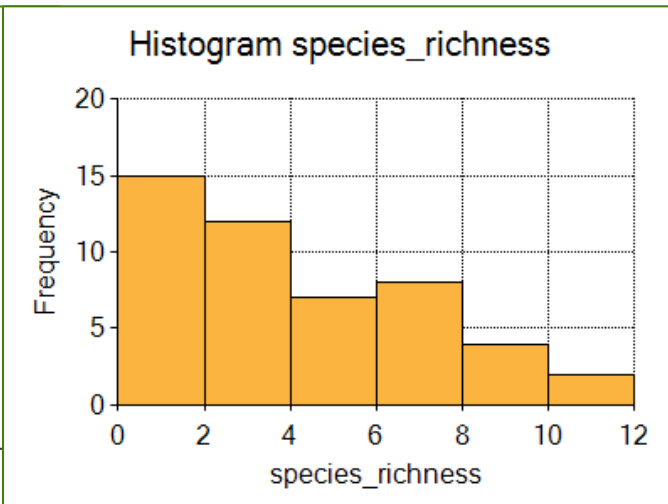
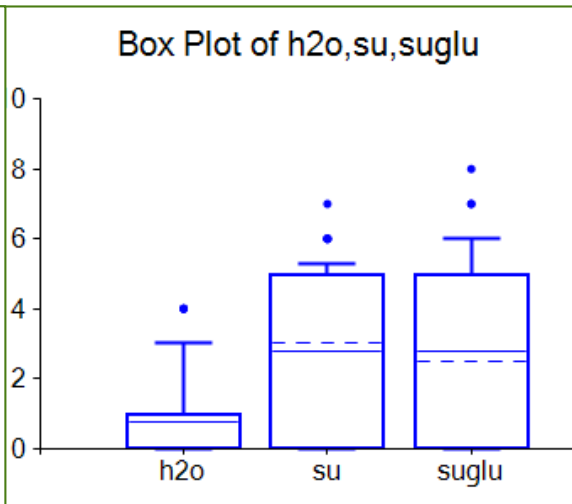
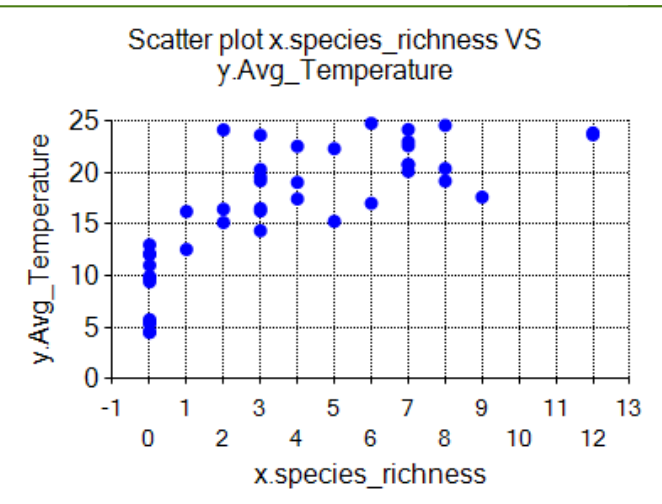
PlotID = Add filter

Mathematic operations

species_occurences All Filtered Add math operation

Stats on filtered data													
Name	Sum	Min	Max	Avg	VARIANCE	STDDEV	Standard Error	Range	Median	Mode	Confident Interval Upper Limit	Confident Interval Lower Limit	
species_occurences	516	0	27	10,75	81,9791666666667	9,05423473666696	1,30686621563018	27	11,5	0	13,3114577826352	8,18854221736484	

• Statistical plot: scatter, histogram, box plot, plot based on aggregated habitat

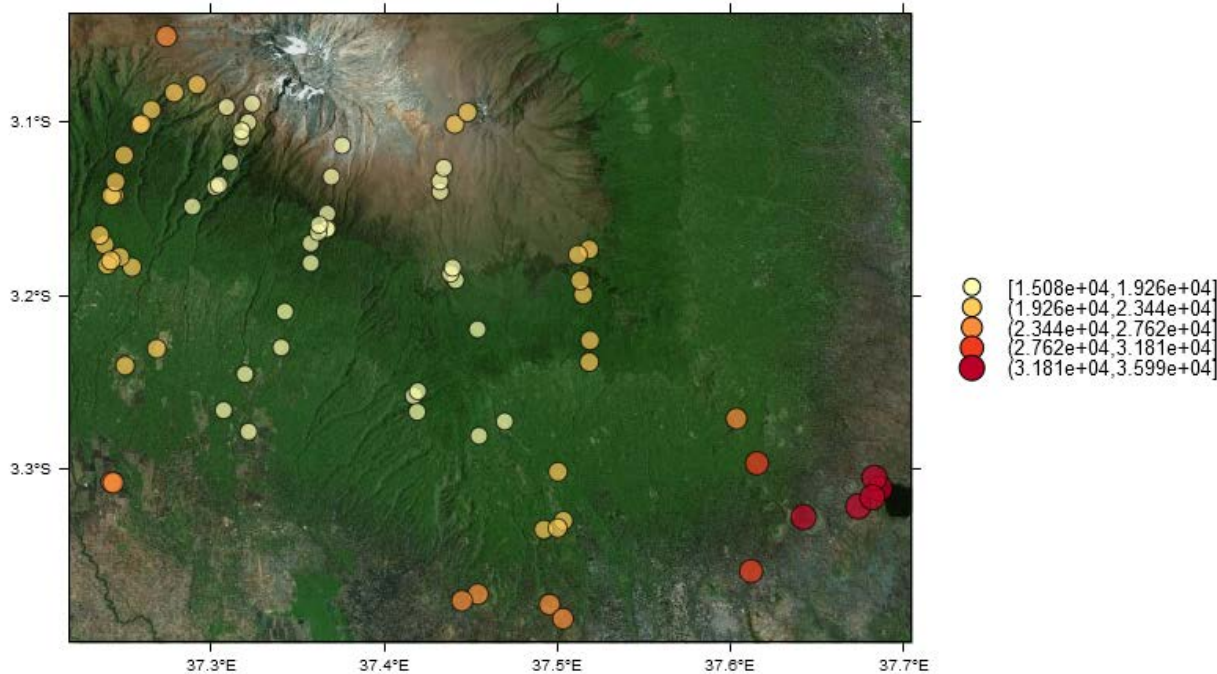


Plot filtered data on a map

Choose a variable to plot on the map:

Select aggregation method for your variable based on unique plot:

DISTANCE_2D



R.NET



Database application

Geospatial Visualization

Or more...






Data
integration
and analysis
tools

Data long
term archive

A dance between ice and fire...





Thank you for your attention!