Vegetation classification – lesson's learnt from analysing a 30,000-plot database.

ALCOA

Dr Andrew Grigg

Curtin University

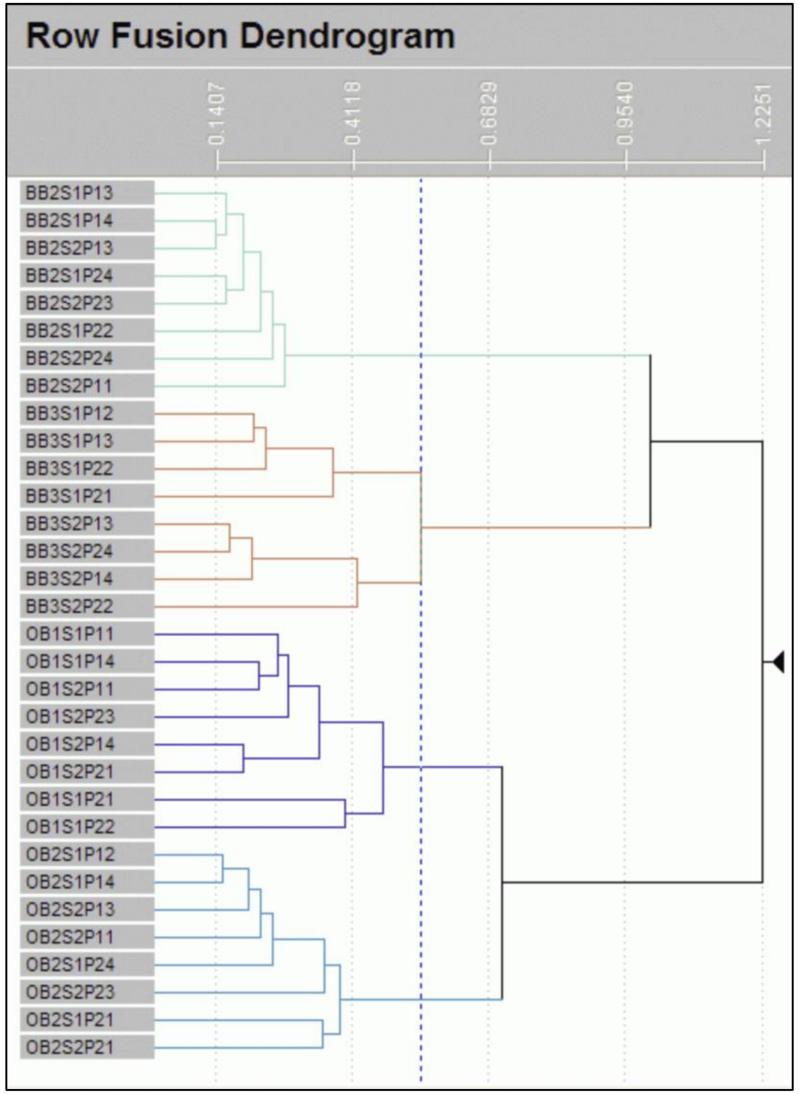
Sarah Luxton, Dr Grant Wardell-Johnson,
Dr Todd Robinson, Lewis Trotter

Arthur Rylah Institute for environmental research

Dr Ashley Sparrow

Overview

- **Re-cap** vegetation classification
- **Recent developments**
  - “Big-data”
  - Software tools for data-processing.
- **Case-study**
  - Alcoa dataset, 30 000 plots, 500 species, 25 years.
- **Going forward...**

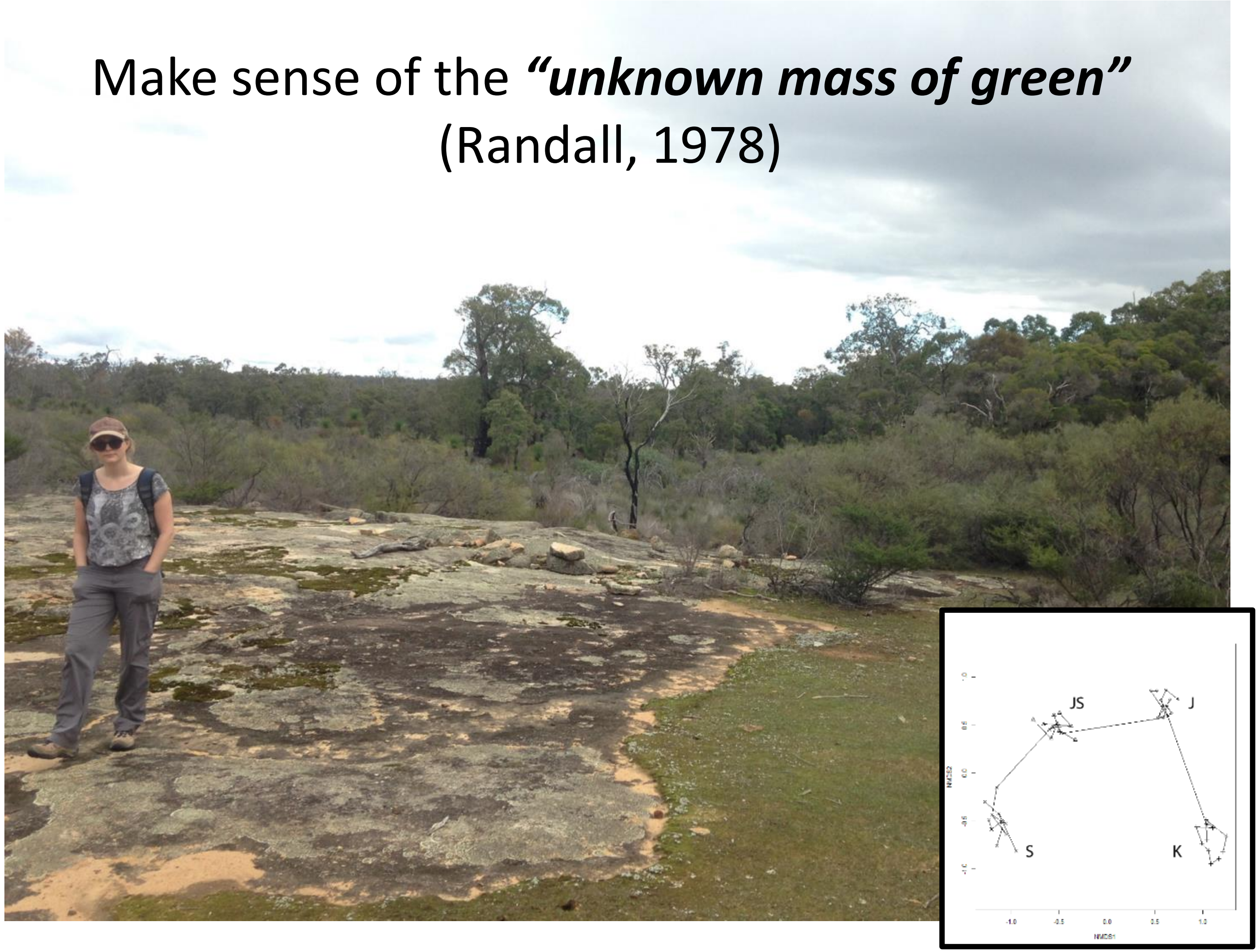


Vegetation classification - purpose

Phytosociology =

“***the science of recognising & defining different plant communities***” (Kent, 2011)

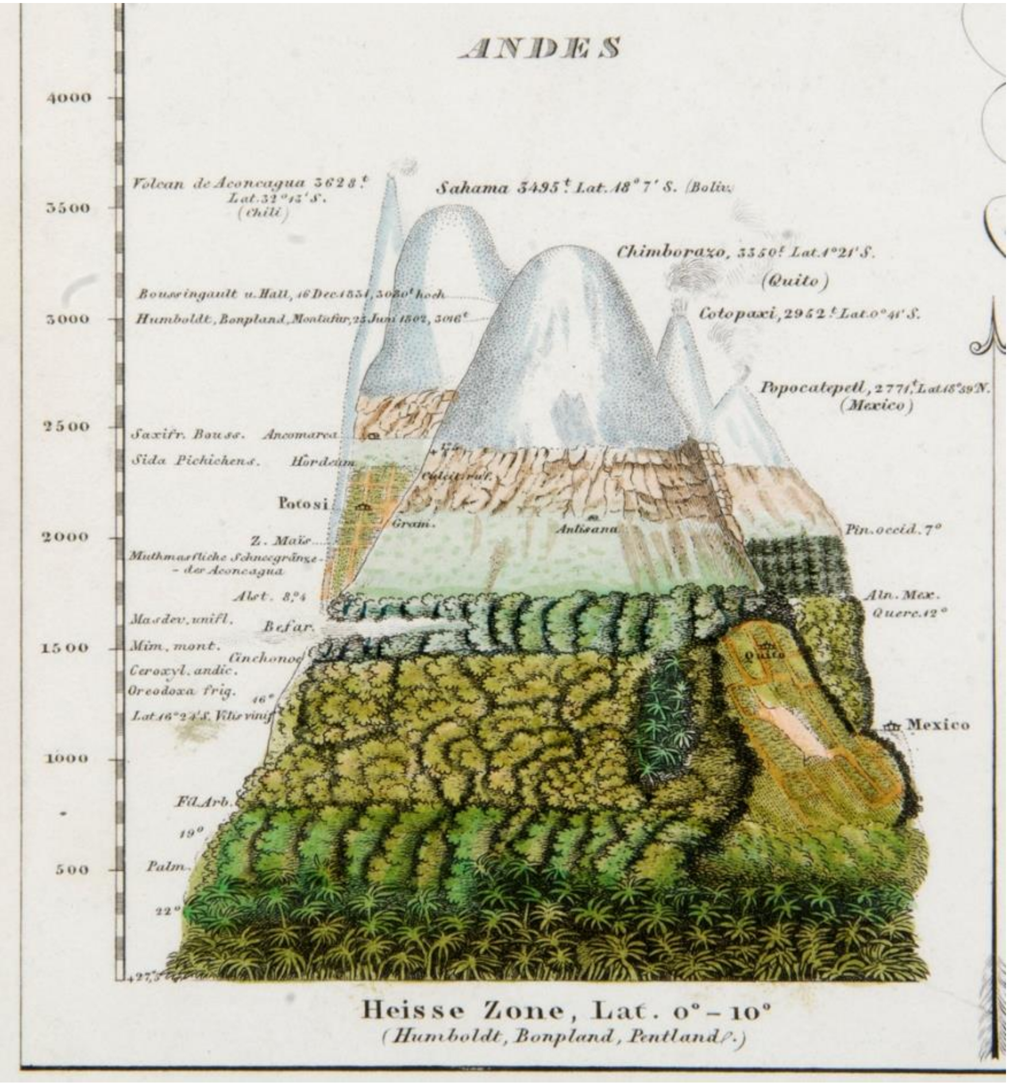
- Integrates species variation into recognisable units.
- Indicator of abiotic conditions.
- Is an *abstraction.*



Make sense of the ***“unknown mass of green”*** (Randall, 1978)

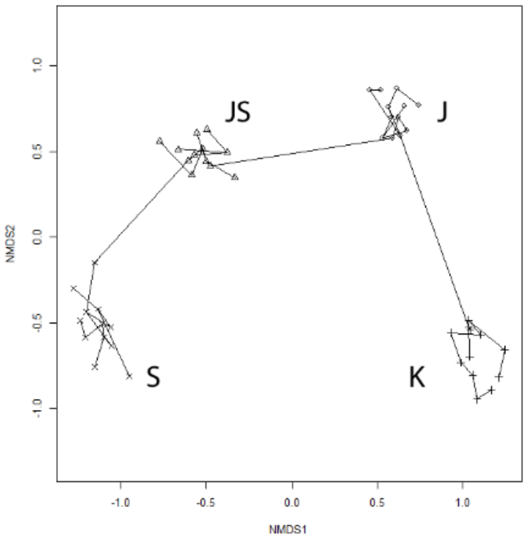
Vegetation classification – brief history

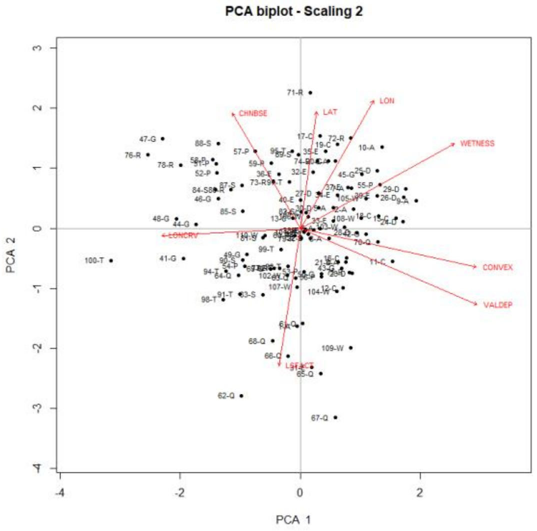
- **1800’s:** Humboldt – 1st vegetation classifications.
- **1900-20’s:**
  - Nordic (**structural**)
  - Braun-Blanquet (**floristic**) classification
  - Clements – “super-organism”
  - Gleason – individual responses
- **1980-90’s:** the problem of **scale**


ANDES
4000
3500
3000
2500
2000
1500
1000
500
Volcan de Aconcagua 3628.t Lat.32°13'S. (chili)
Sahama 3495.t Lat.18°7'S. (Boliv.
Chimborazo, 3350.t Lat.1°21'S. (Quito)
Boussingault u. Hall, 16 Dec.1831, 3080t hoch
Humboldt, Bonpland, Montufar, 23 Juni 1802, 3016t
Cotopaxi, 2952.t Lat.0°41'S.
Popocatepetl, 2771.t Lat.18°59'N. (Mexico)
Saxifr. Bouss.
Ancomarca
Sida Pichichens.
Hordeum
Potosi
Gram.
Antisana
Pin. occid. 7°
Z. Maïs
Muthmassliche Schneegränze - der Aconcagua
Alst. 8,°4
Aln. Mex.
Querc. 12°
Masdev. unifl.
Befar.
Mim. mont.
Cinchonoe
Quito
Ceroxyl. andic.
Oreodoxa frig.
16°
Lat.16°24'S. Vitis vinif.
Mexico
Fil. Arb.
19°
Palm.
22°
27,5°
Heisse Zone, Lat. 0° – 10°
(Humboldt, Bonpland, Pentland.)

Vegetation classification – brief history

- **Traditionally**: two schools of thought
  - European (discrete)
  - American (continuous)
- **2000's**:
  - Recognise that pattern is complex (Austin 2013)
  - Computing and the arrival of the "big-data" era

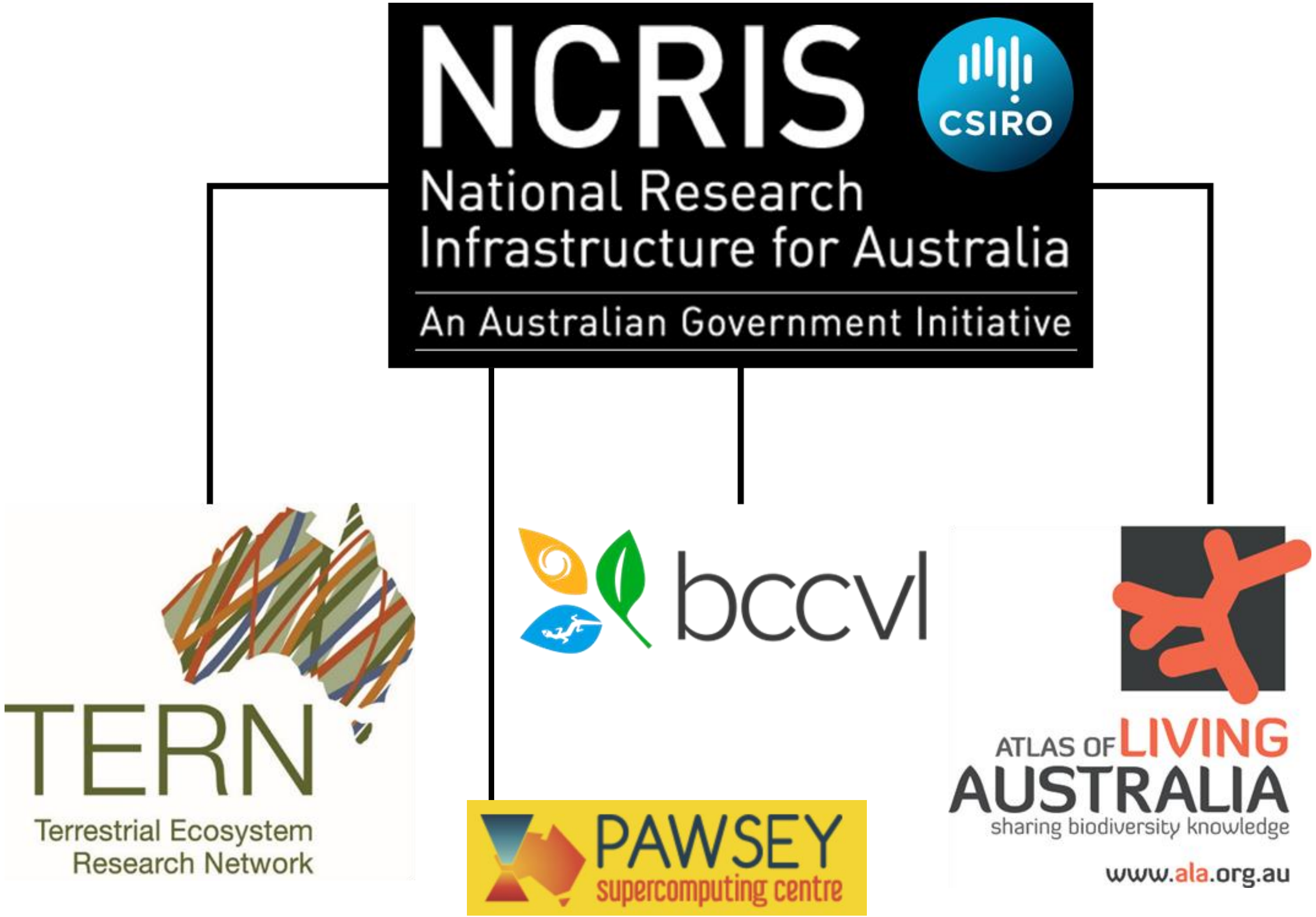




Recent developments – “Big data” – data sharing

Data Source	# of Plots
European Vegetation Archive	**1,027,376**
Global Index of Vegetation-Plot Databases	3,362,775
Terrestrial Ecosystem Research Network	**22,000**
NZ Vegetation Databank	94,000

Recent developments – Ecoinformatics – Australia



NCRIS

CSIRO

National Research Infrastructure for Australia

An Australian Government Initiative

TERN

Terrestrial Ecosystem Research Network

bccvl

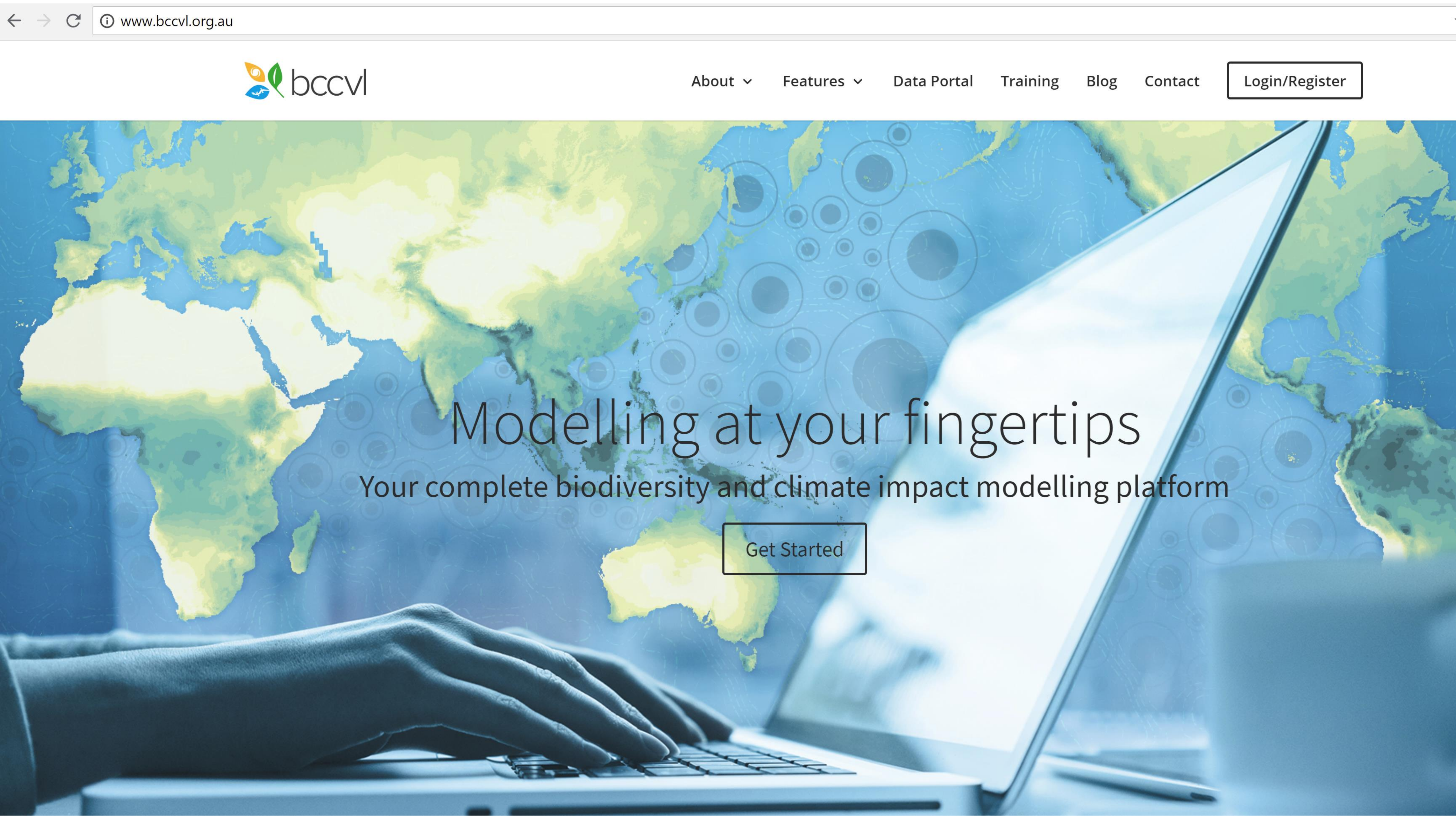
PAWSEY

supercomputing centre

ATLAS OF LIVING AUSTRALIA

sharing biodiversity knowledge

www.ala.org.au



Modelling at your fingertips

Your complete biodiversity and climate impact modelling platform

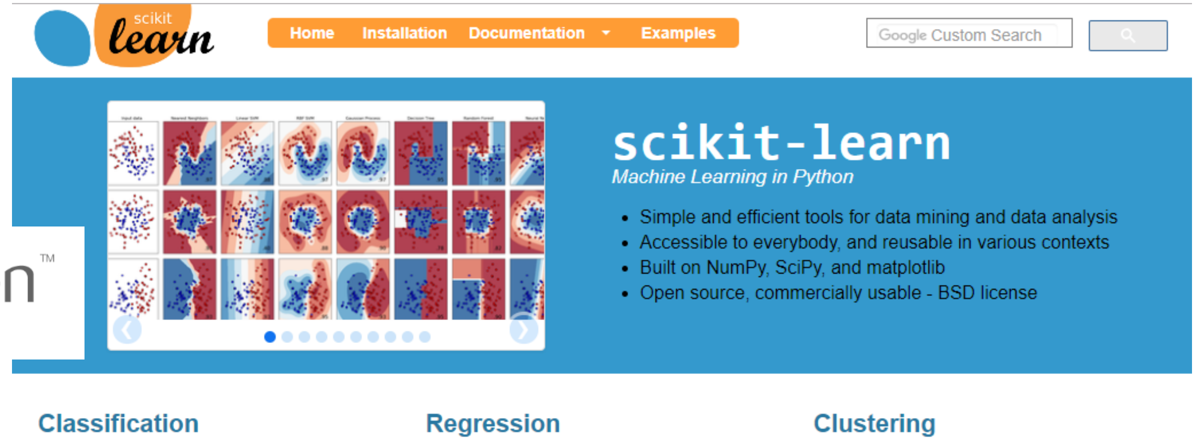
Recent developments – Ecoinformatics - General

- **JUICE** vegetation analysis software.

- **R** statistical software.

- **Machine Learning** online (scikit-learn).

python™



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Regression

Clustering

Lesson’s Learnt…

1. We’re not limited by computing power or analysis tools…

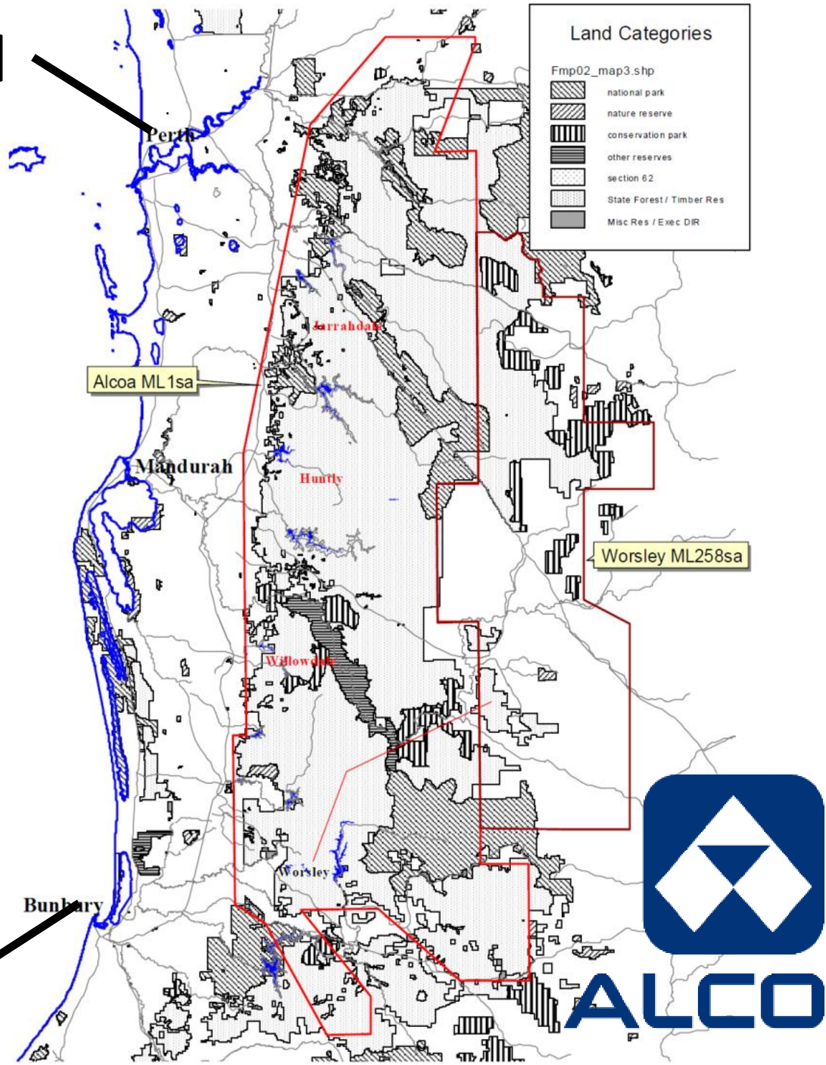
2. **Data availability & quality.**

3. Fit for purpose?

Case study – Alcoa of Australia

- Bauxite mining since 1963 in Northern Jarrah Forest.
- **Vegetation surveys (plots)**
  - **1991 – ongoing.**
- Rare flora & mapping community types.
- Species lists > define seed mixtures for restoration.

PERTH

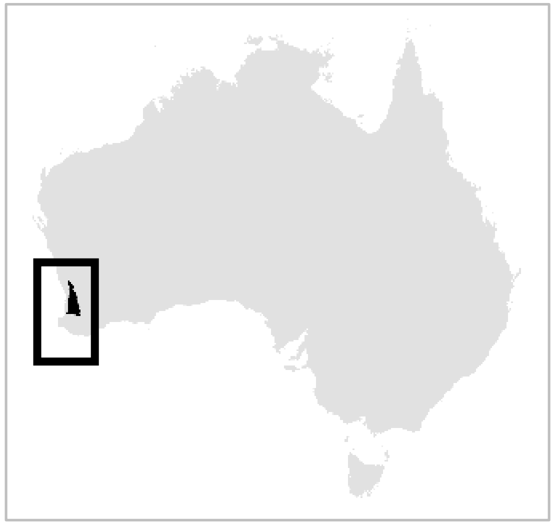


BUNBURY

ALCOA

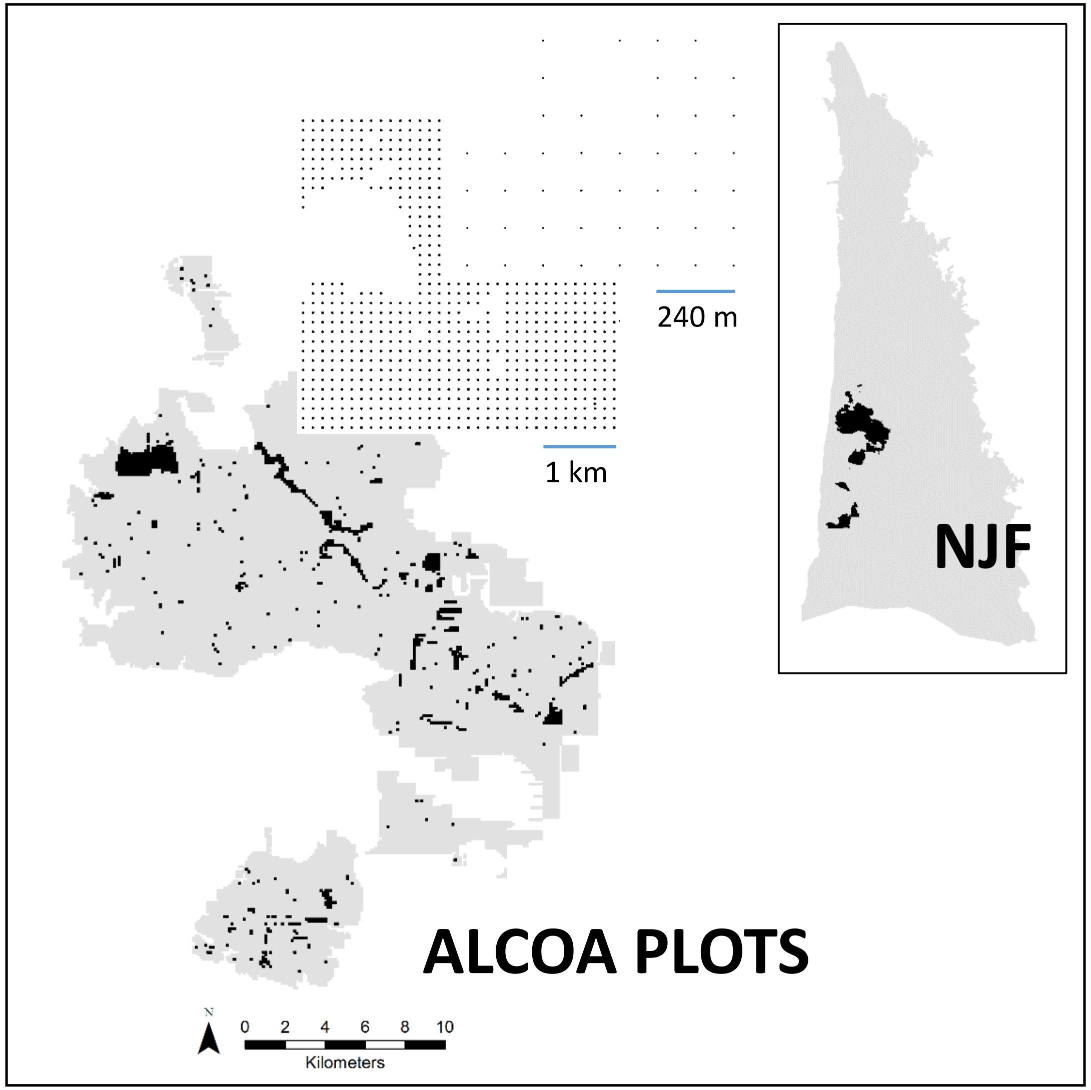
Northern Jarrah Forest

- Southwest WA
  - 3089 species
  - 138 families
- Mediterranean climate
- Subdued topography
- Fire, logging, mining
- Drying ~ 1970's

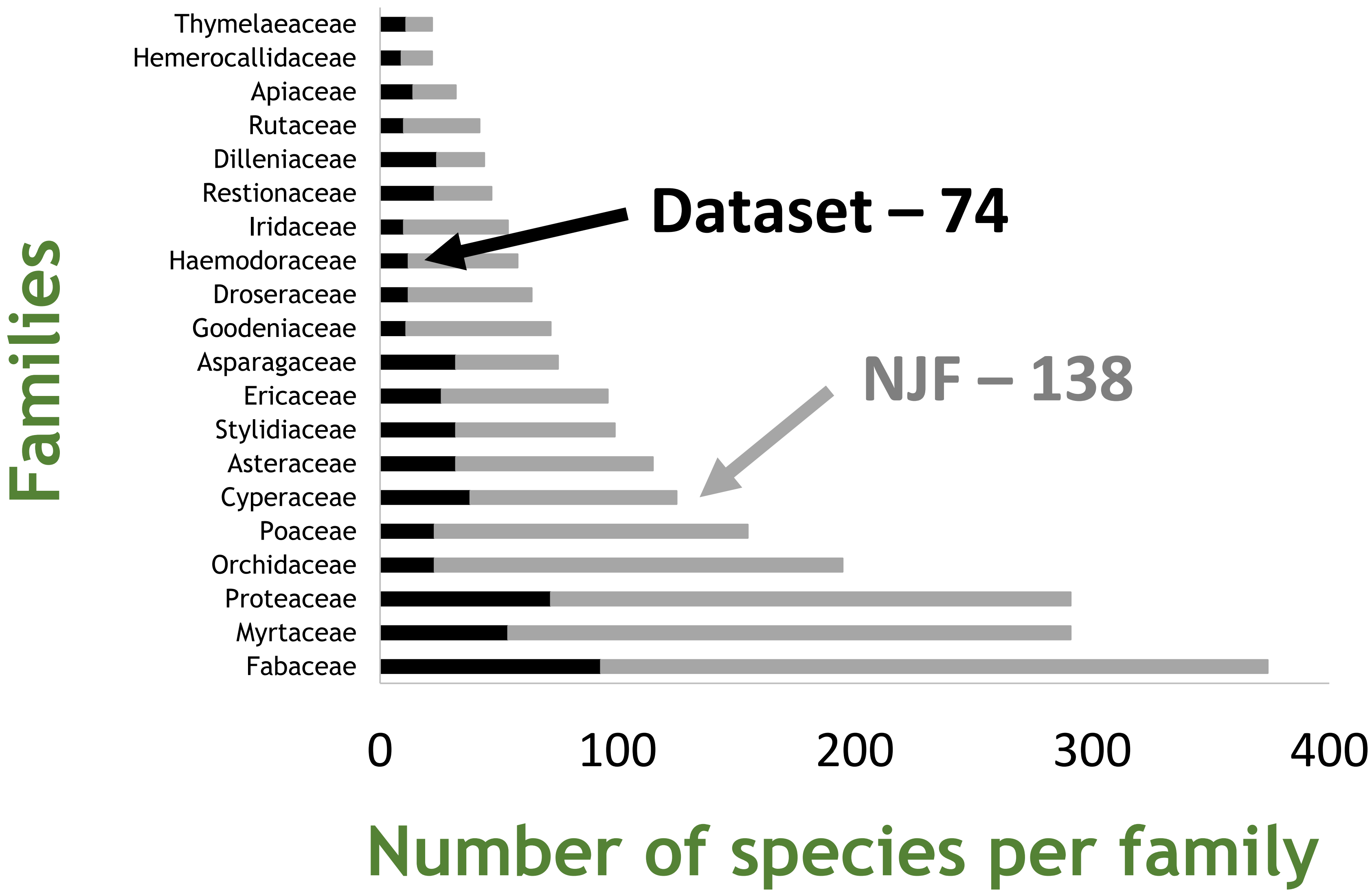


The Data – Vegetation Survey Plots

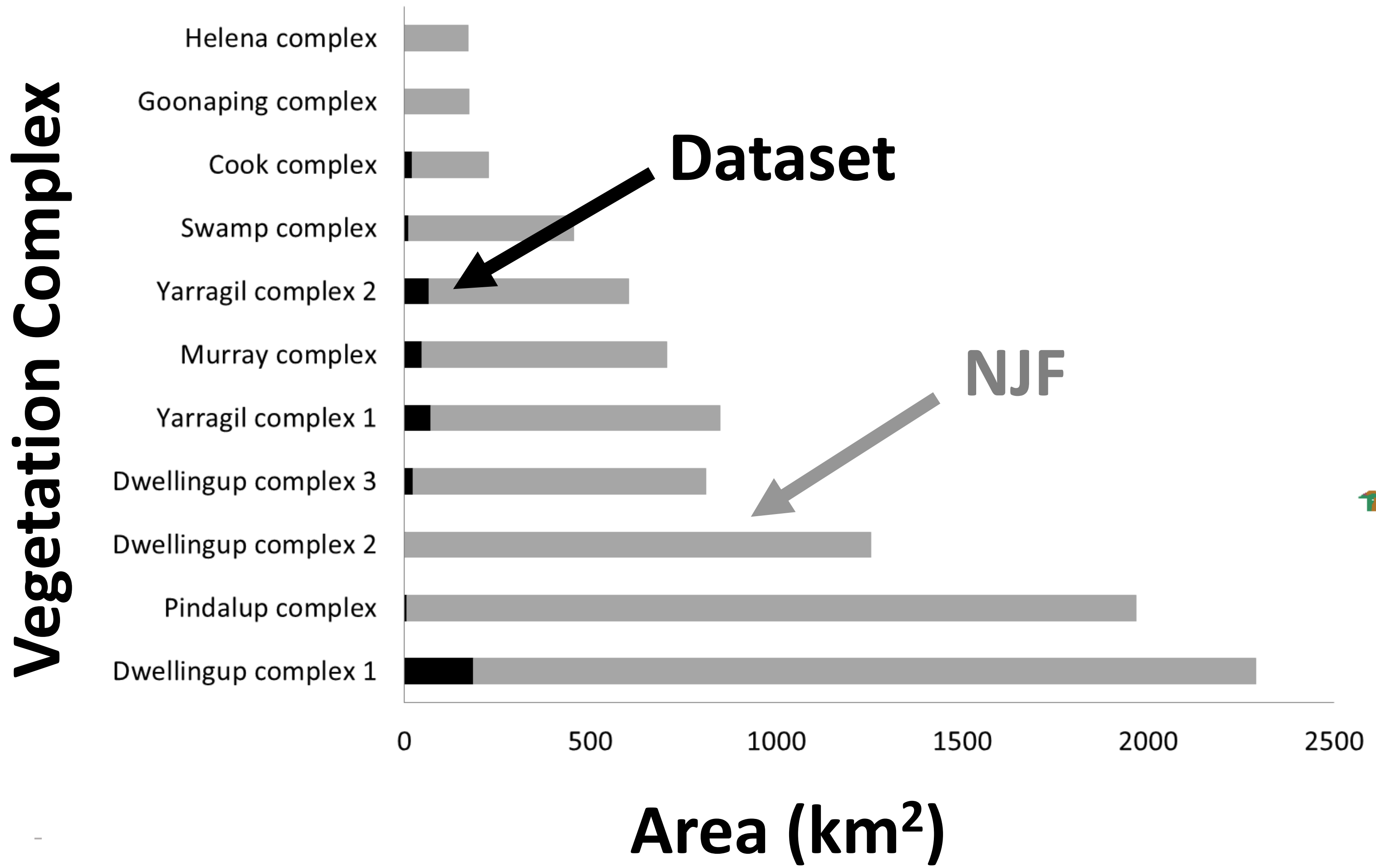
- **120 m^2 grid**
- 20 m – tree species
- 5 m – perennial herb & shrub sp
- **500 species (cleaned)**
- 260 genera
- 74 families
- **88+ botanists…**

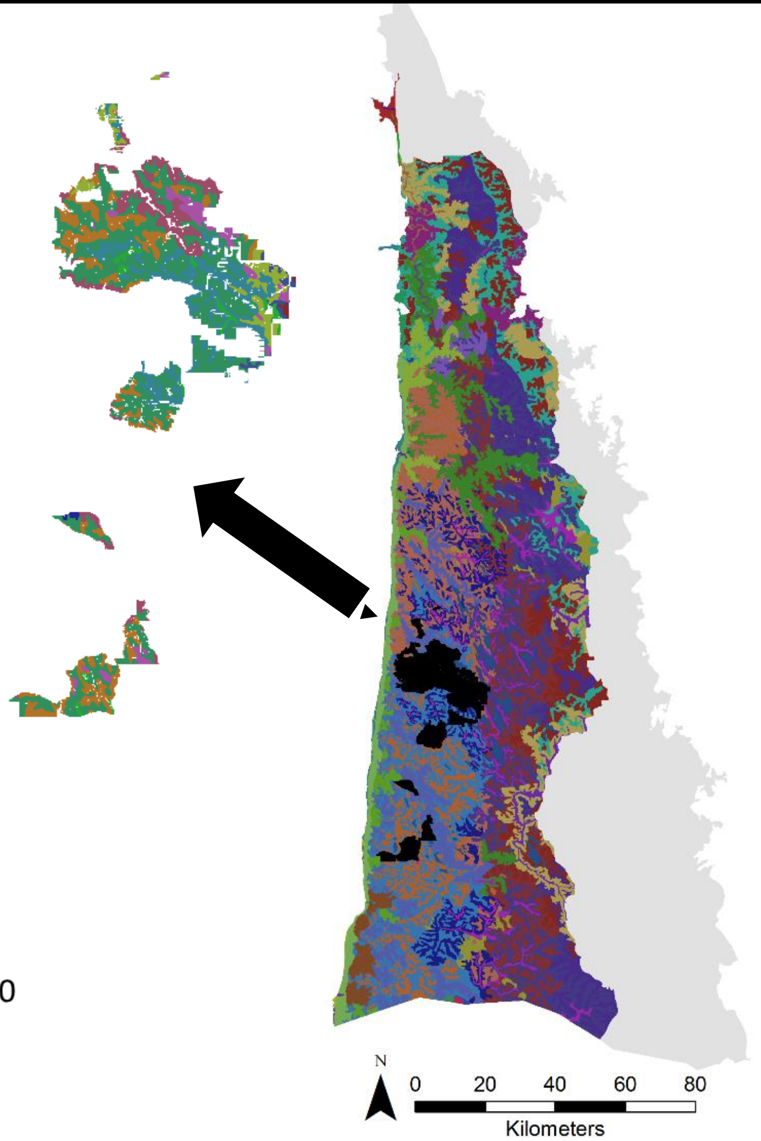


The Data – Taxonomic Context



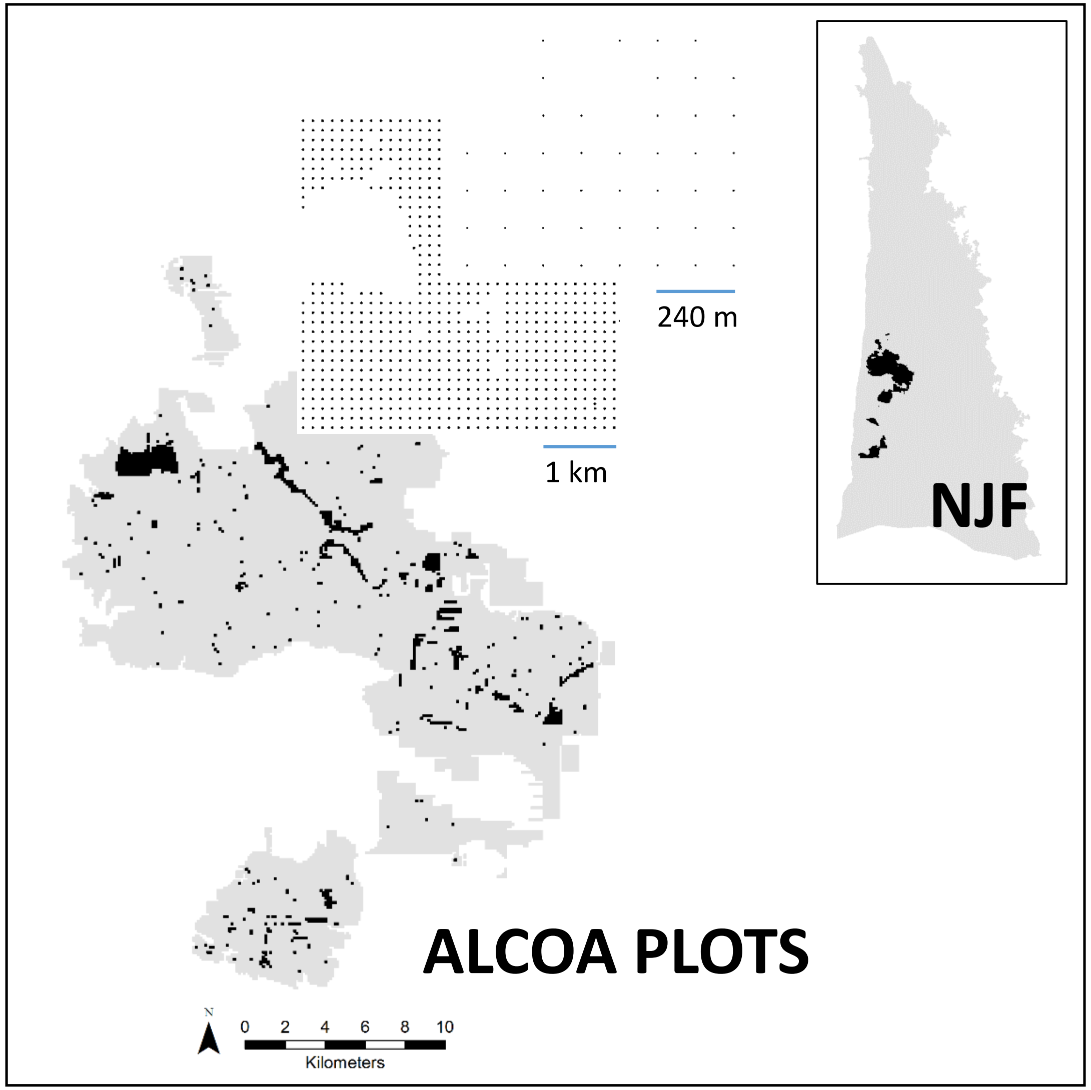
The Data – Geographic Context





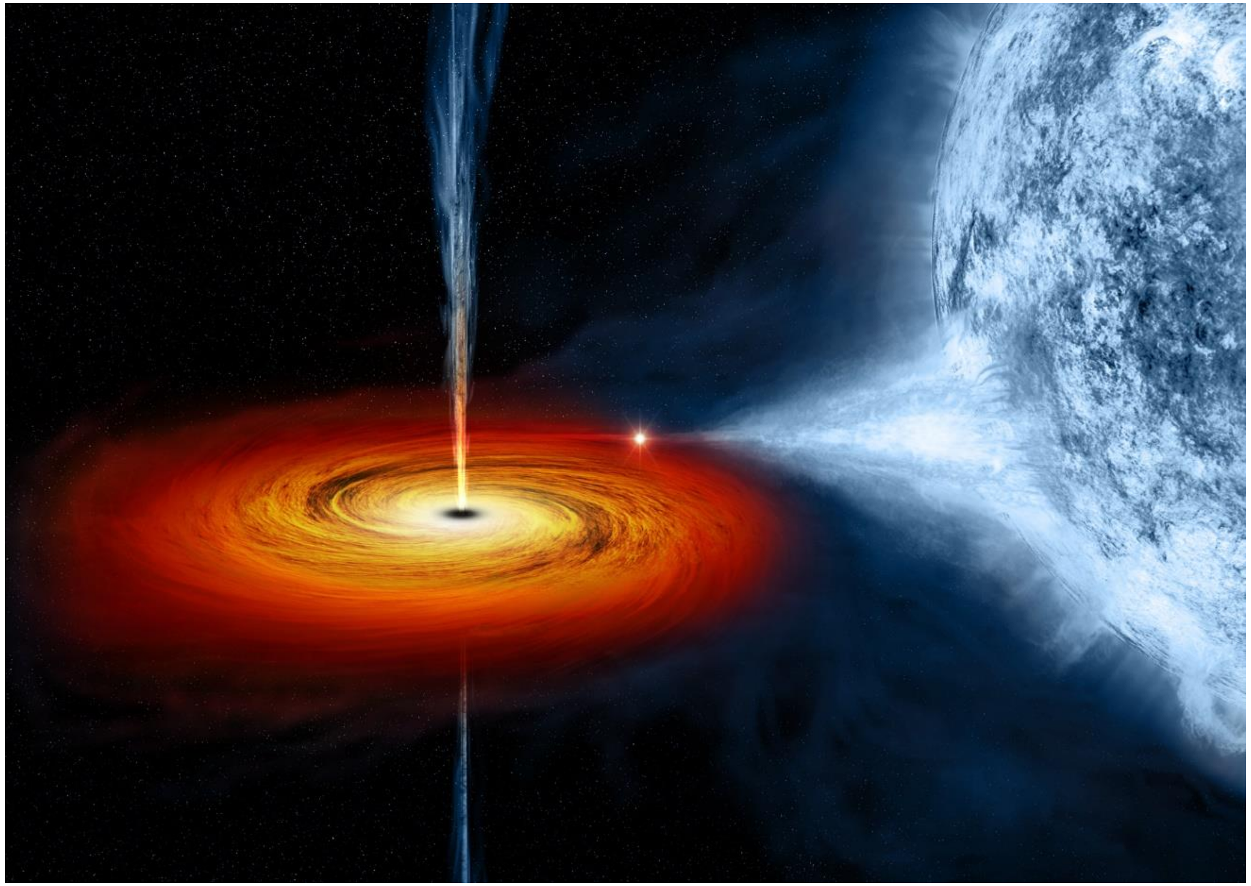
Goal – Cluster the Dataset

- **30,000 plots, 500 spps**
- **GOAL – Cluster the Dataset**
  - Exploratory analysis
  - Best method to use?
- **Method options**
  - Analysis software
  - Dissimilarity measures
  - Clustering algorithms



Methods options

- **Software options**
  - JUICE / PATN / PAST
  - R / Python
- **Dissimilarity measures**
- **Clustering algorithms**
  - Hierarchical vs non-hierarchical
  - Divisive vs agglomerative
  - Polythetic vs monothetic



An artist's drawing of black hole Cygnus X-1.

Source: https://www.nasa.gov/sites/default/files/cygx1_ill.jpg

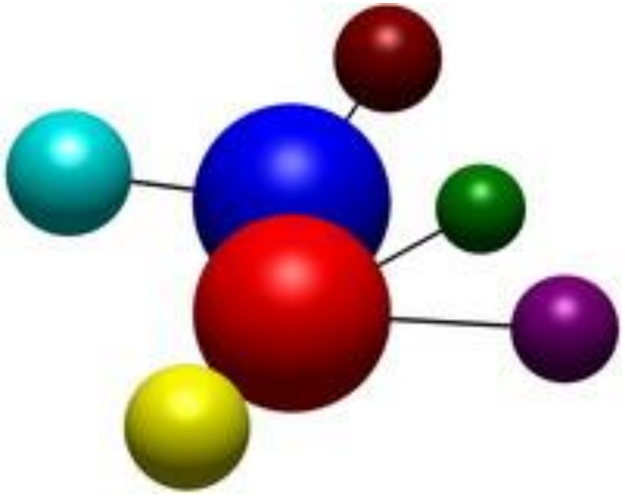
Methods options – Software

- Pros
  - Handle large datasets
  - Reproducible code
  - Well-written packages for vegetation analysis, developed & supported by international community
  - Free
- Cons
  - Have to learn R

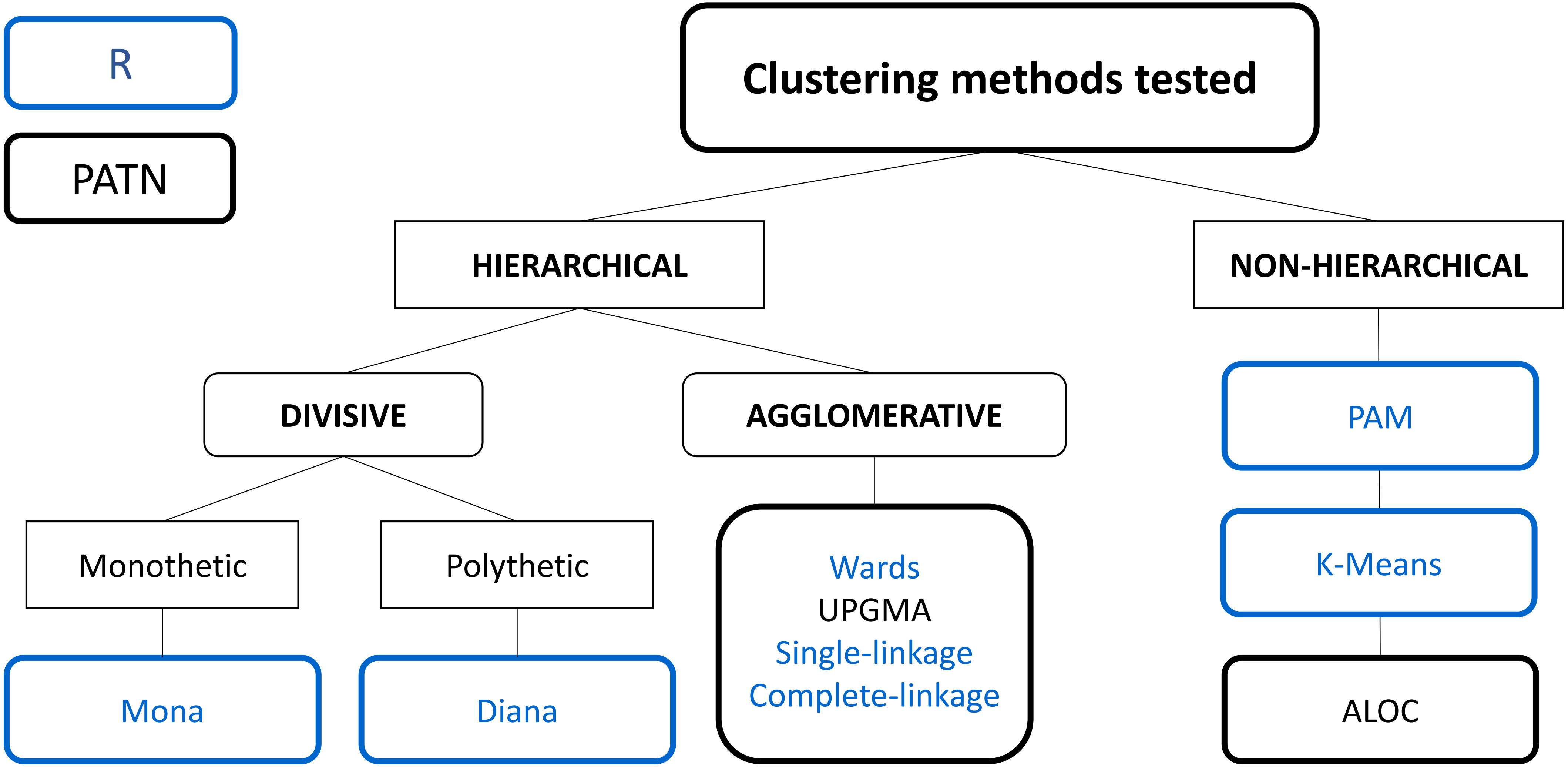
RStudio®

Methods options – Software

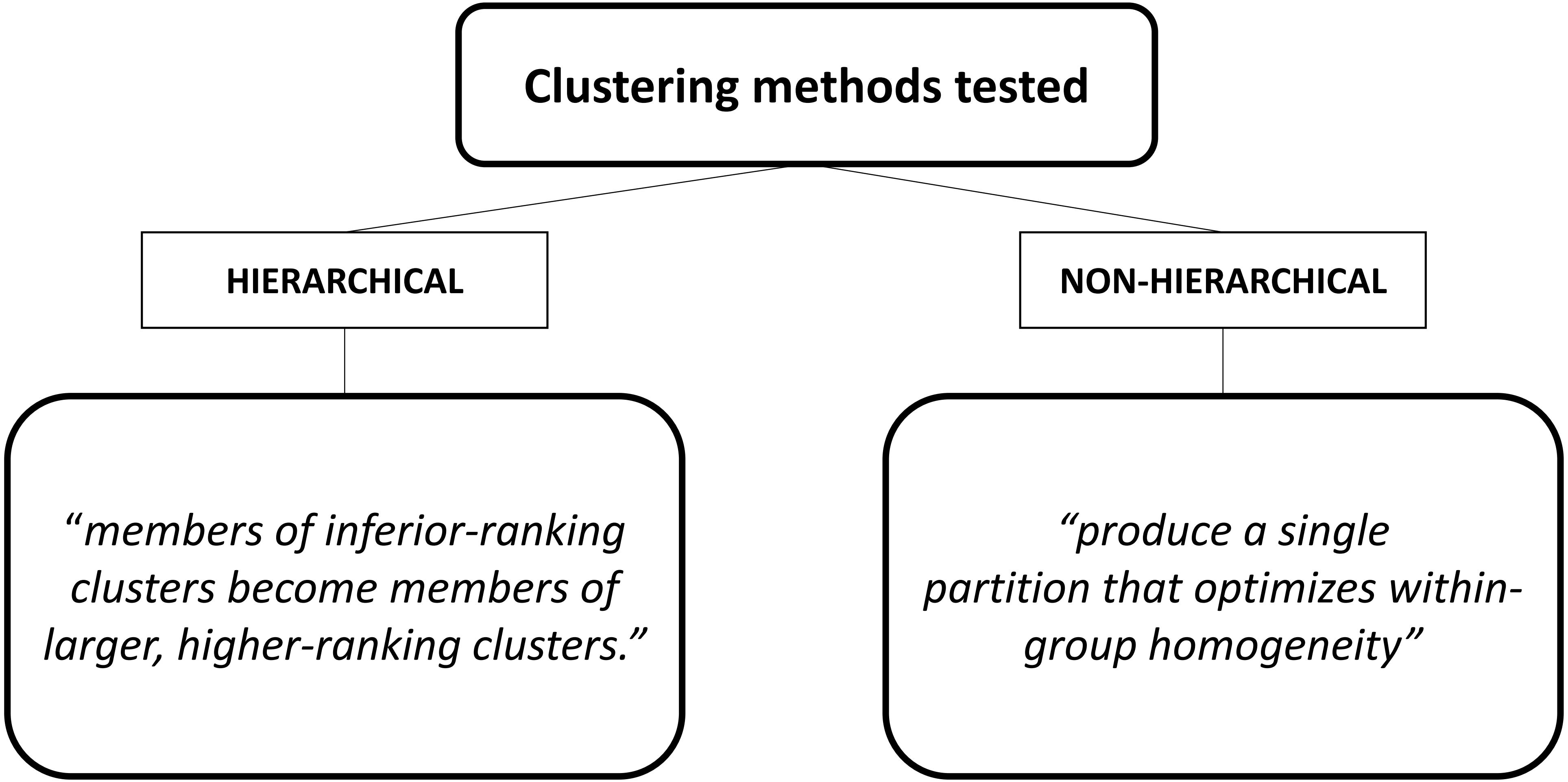
- **PATN**
  - Pros
    - Don’t need to learn R
  - Cons
    - Less-flexibility in range of analyses available
    - Less help
    - Handle large datasets?



Methods options – Clustering algorithms



Methods options – Clustering algorithms



Clustering methods tested

HIERARCHICAL

“members of inferior-ranking clusters become members of larger, higher-ranking clusters.”

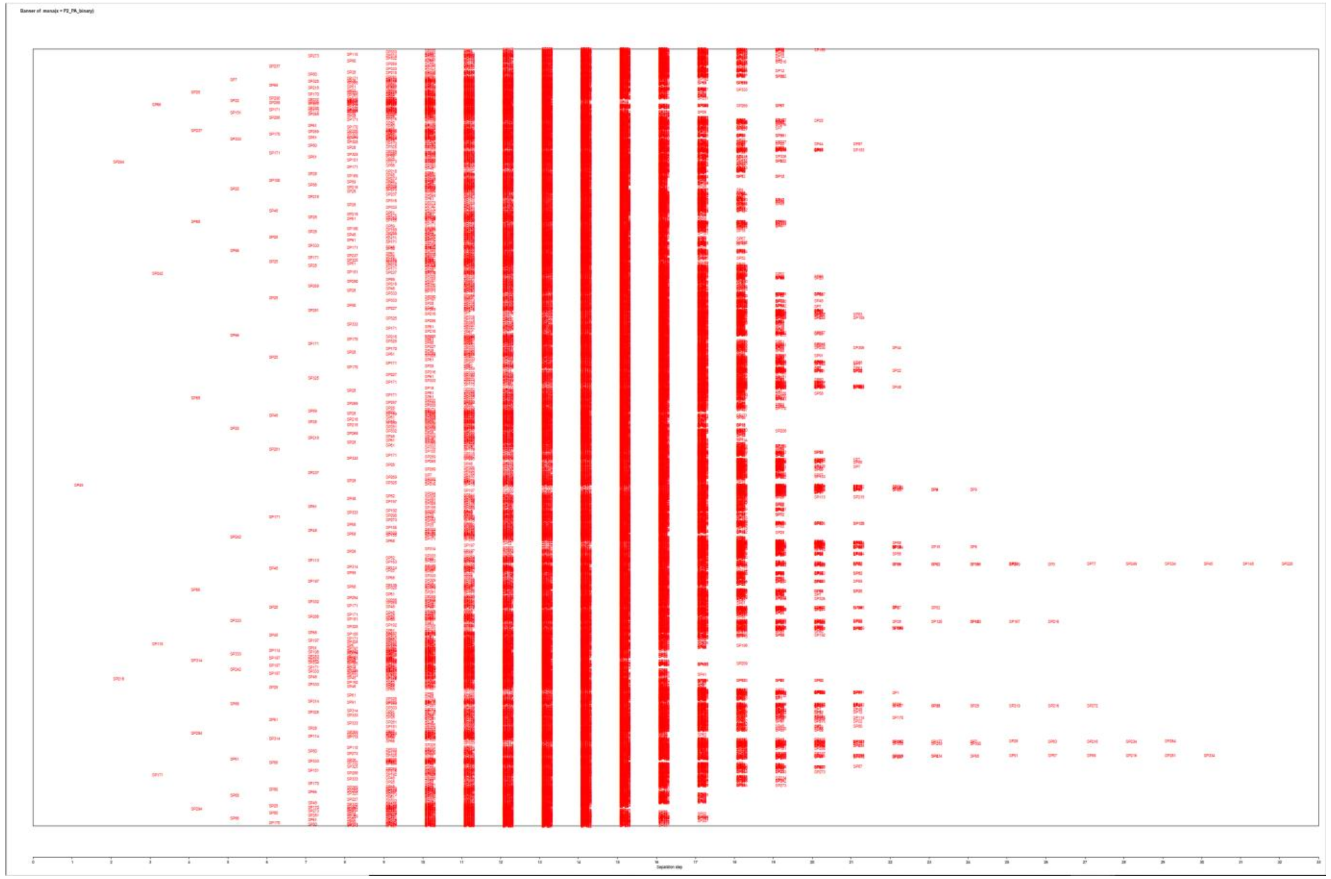
NON-HIERARCHICAL

“produce a single partition that optimizes within-group homogeneity”

Source: Chapter 8, Numerical Ecology (Legendre & Legendre, 2012), pp 349.

Clustering methods - DIVISIVE

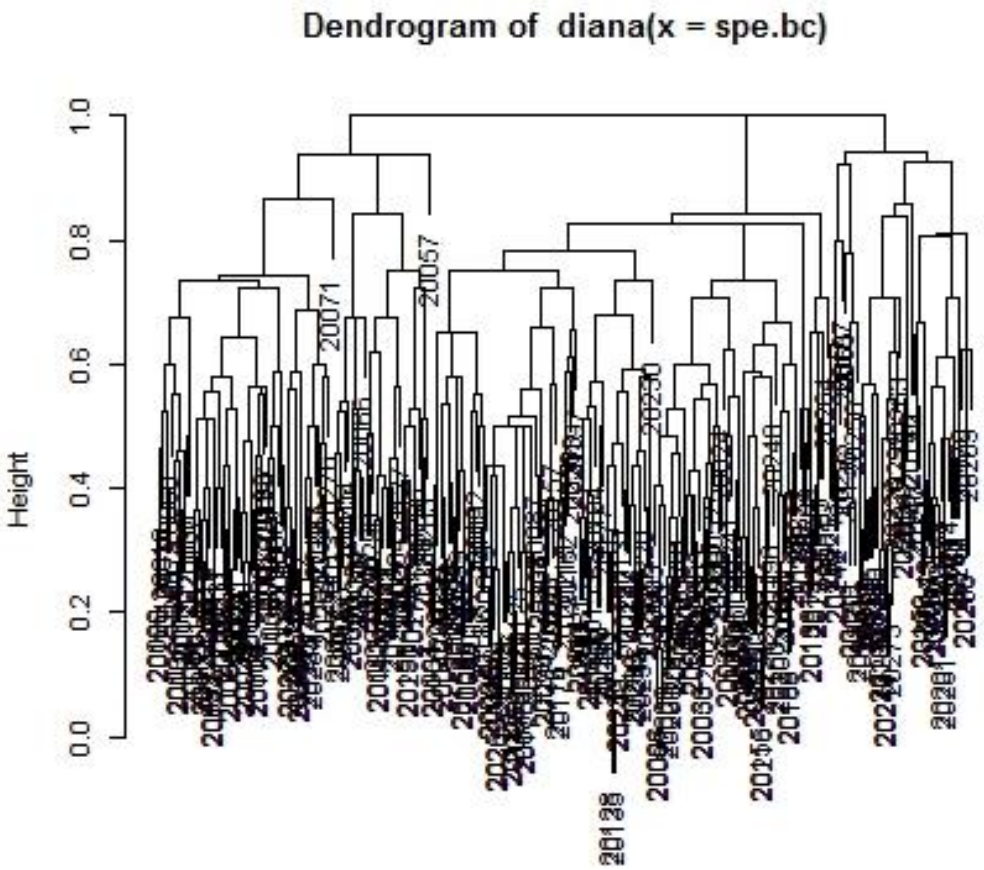
- **Divisive** starts with full dataset and splits groups (“top down”)
- **Mona** splits groups by species.
  - Pro – computational lighter
  - Con – clusters based on single-species
  - Interpreting groups tricky



MONA “banner” plot

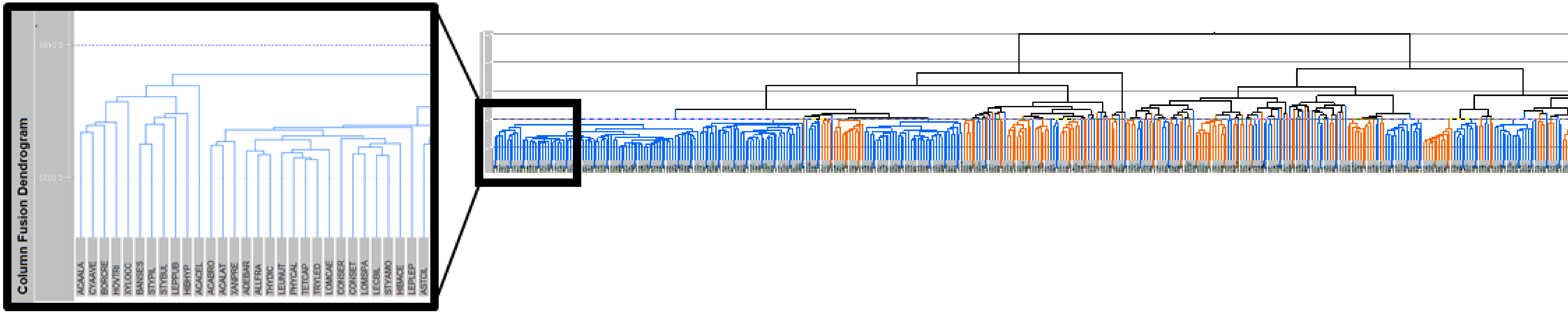
Clustering methods - DIVISIVE

- **Divisive** starts with full dataset and splits groups (“top down”)
- **Diana** splits groups by plots.
  - Pro – clusters based on plots (**more ecologically meaningful**)
  - Con – difficult to interpret >100 plots.



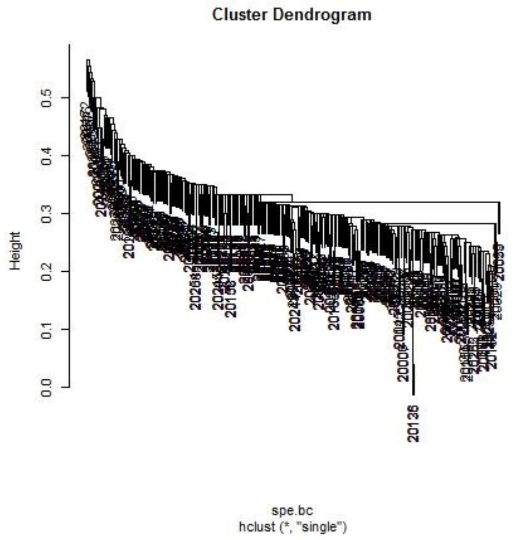
DIANA dendrogram (300 plot sub-sample)

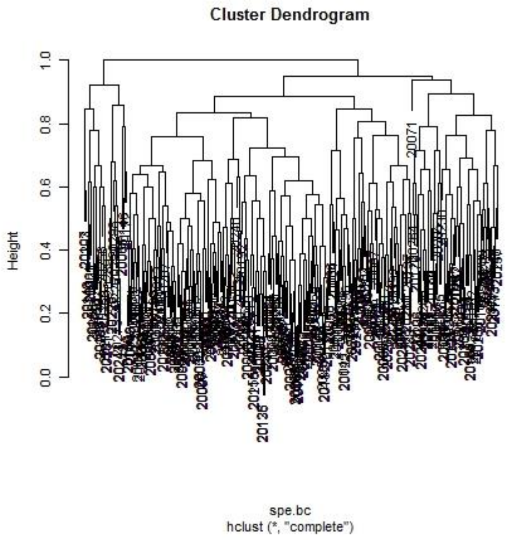
Clustering methods - AGGLOMERATIVE

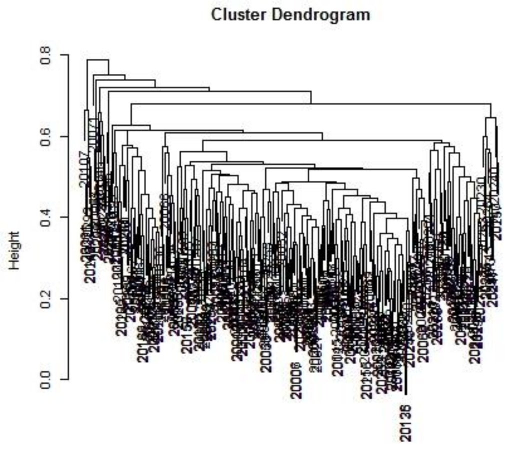


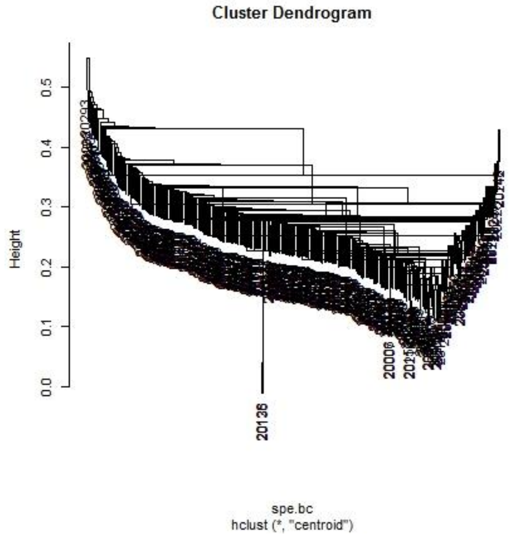
- **Agglomerative** starts w individual plots & builds groups (“bottom-up”)
- **Pros:** can see overall relationships
- **Cons:** interpreting figures when >100 plots

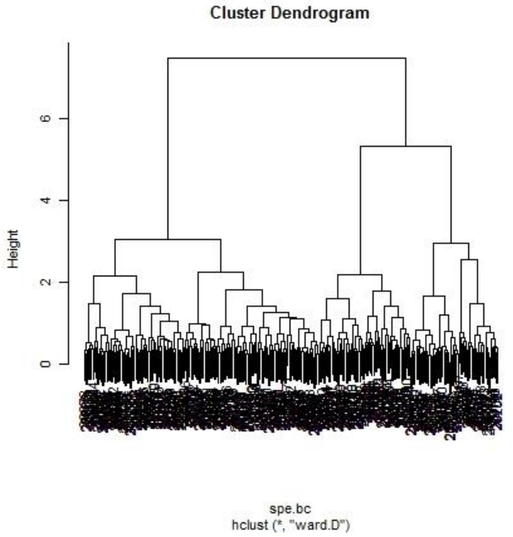
Clustering methods - AGGLOMERATIVE











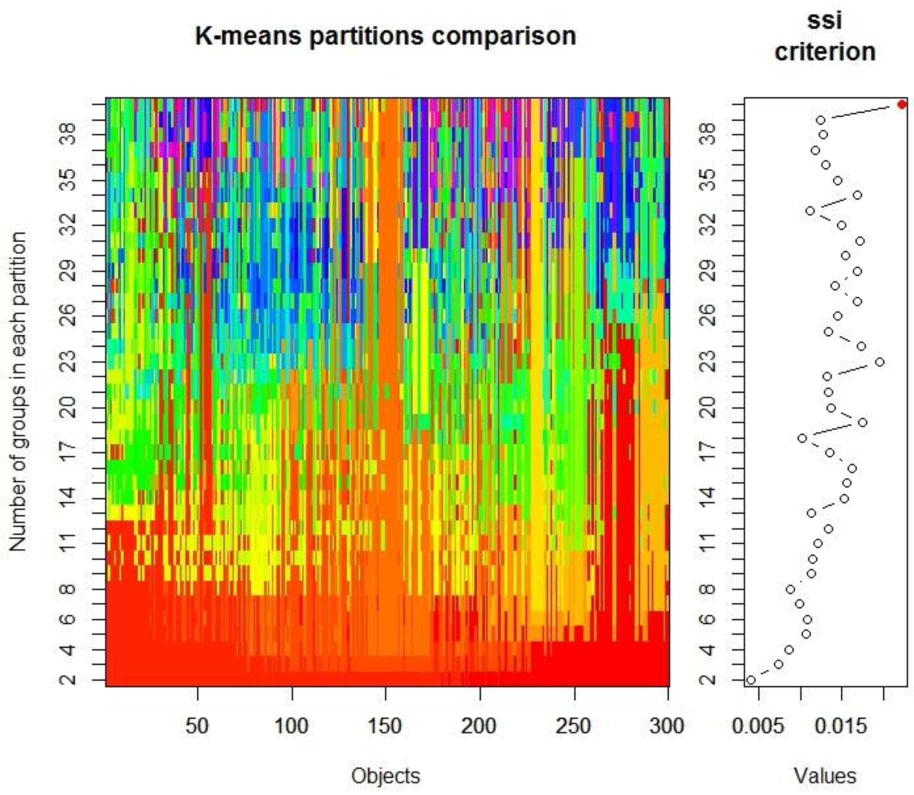
No matter which method you use...

>100 plots = issues.

Clustering methods – NON-HIERARCHICAL

"Plots within each cluster are more similar to one another than to plots in other clusters."

- PAM and K-means
  - Pro – fast
  - Cons – the user defines the number of groups
  - Is not appropriate for raw species abundance data with lots of zeros



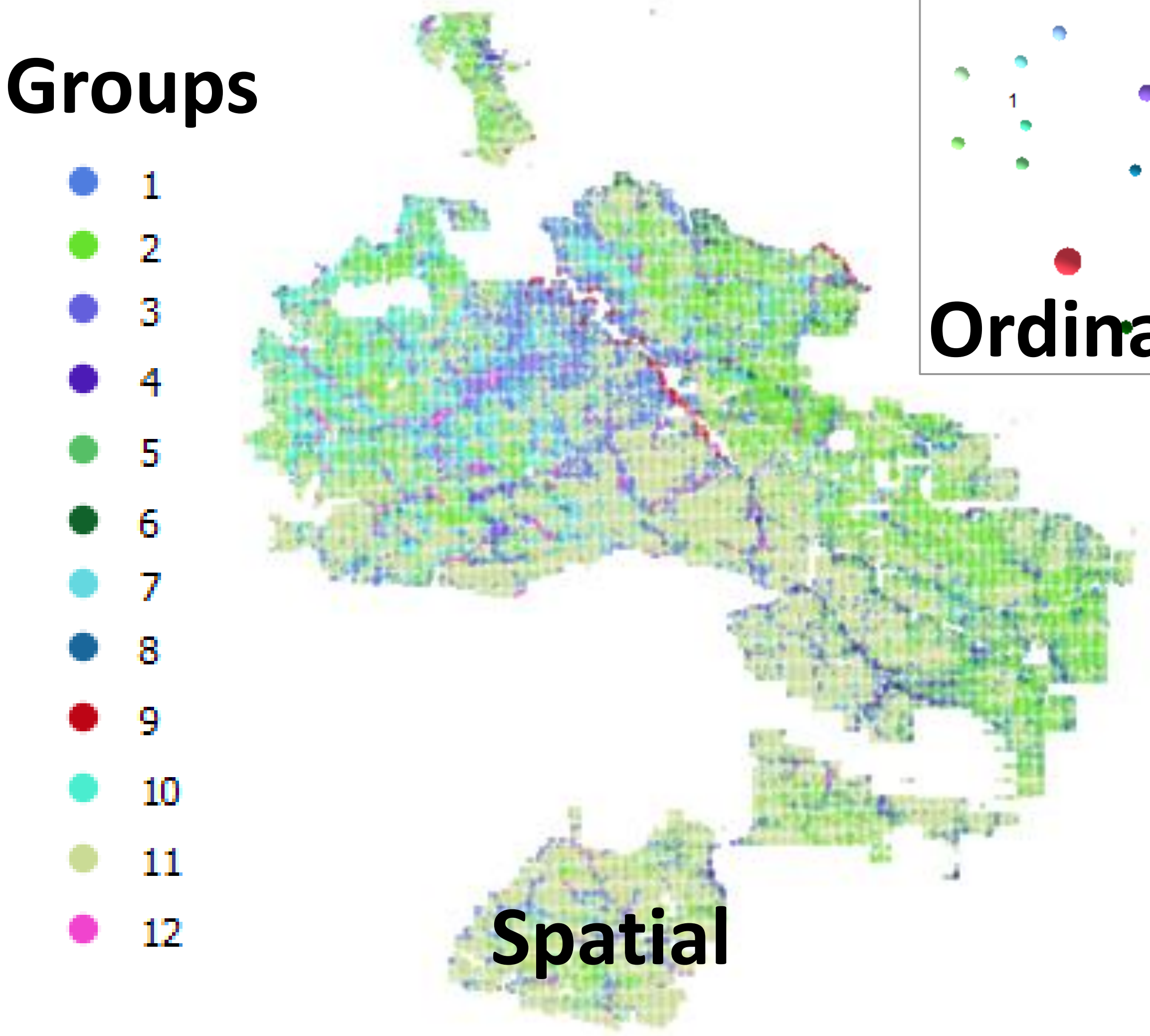
K-means (300 plot sub-sample)

Clustering results – NON-HIERARCHICAL

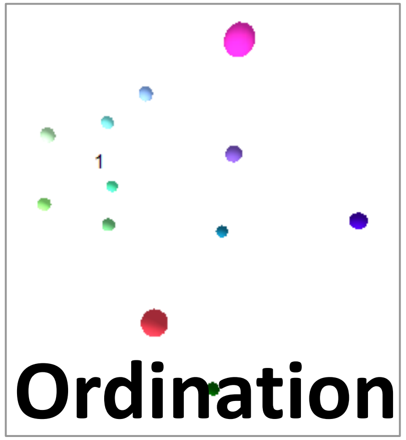
ALOC in PATN

(best option)

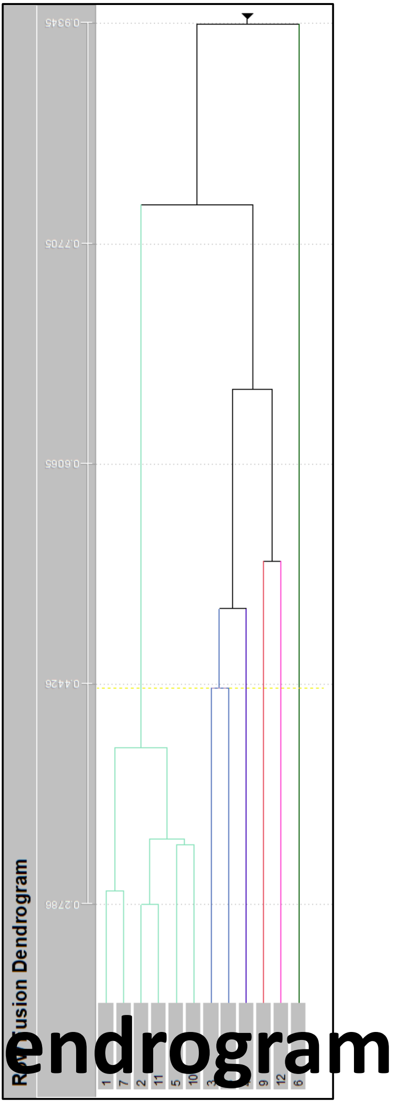
- Fast
- Reliable
- Good visuals



Spatial



Ordination

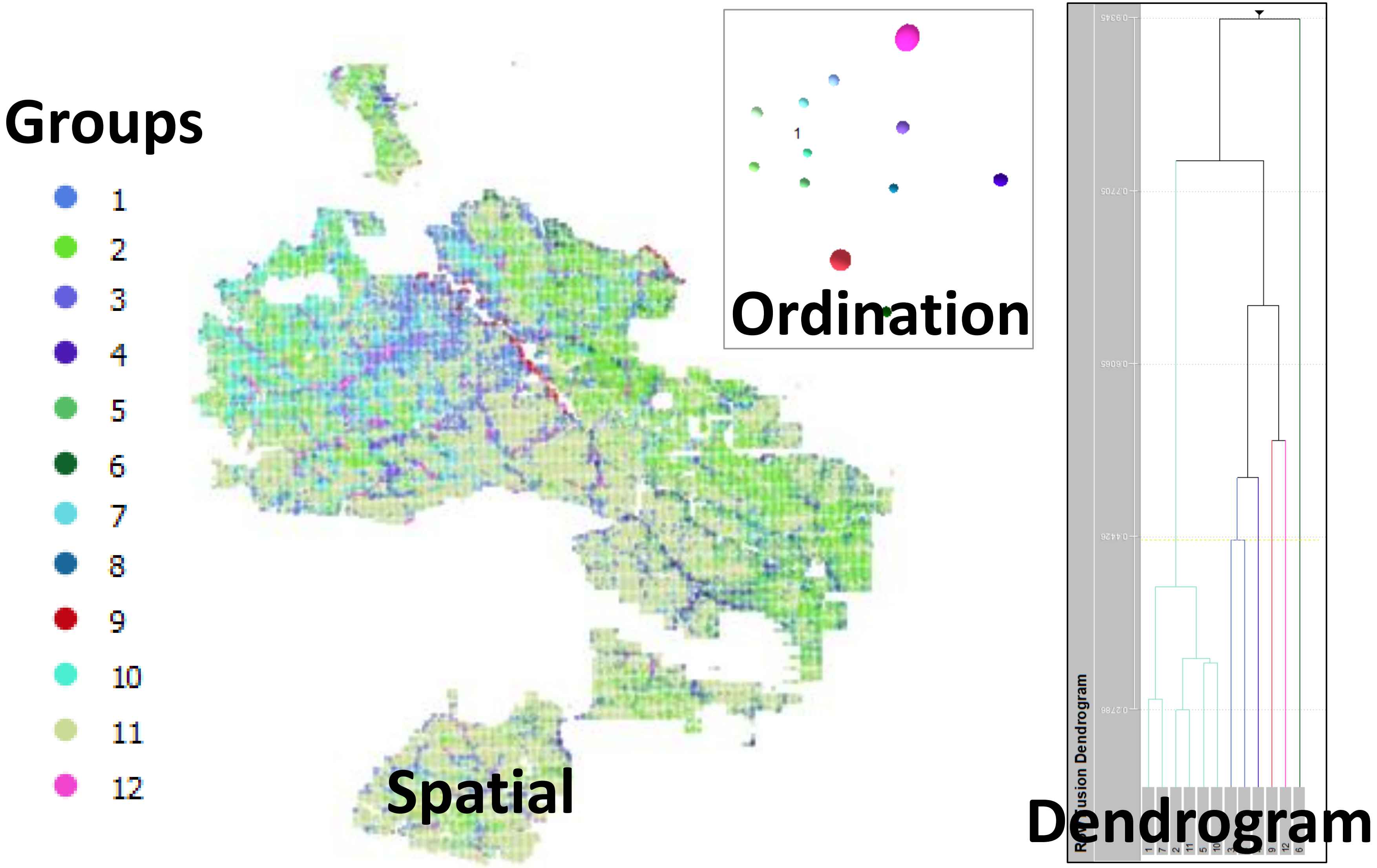


Dendrogram

Clustering results – NON-HIERARCHICAL

ALOC in PATN
(best option)

- Less control
- Cost



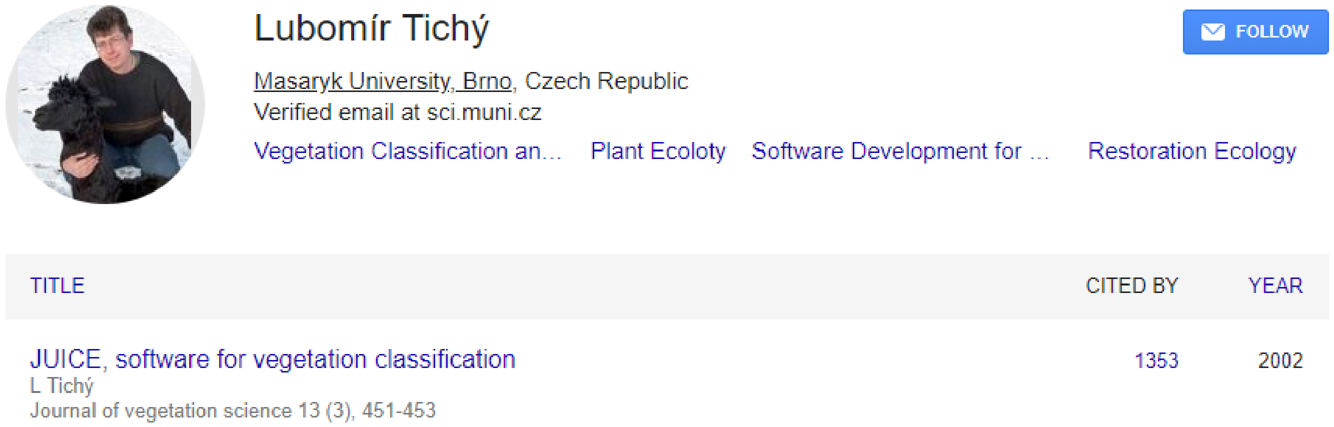
Lesson’s Learnt…

1. For larger datsets – **ALOC in PATN** is an excellent clustering tool.

2. **R** not ideal with “big-data” (data has to be stored in physical memory).

3. Clarify ***Question*** important.

Lesson's Learnt…

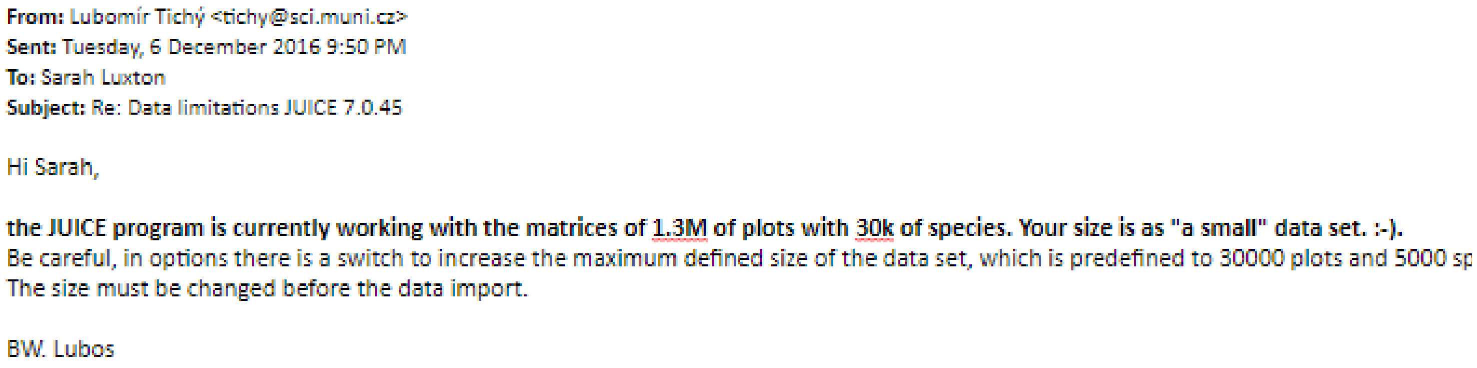


Lubomír Tichý

Masaryk University, Brno, Czech Republic

Verified email at sci.muni.cz

TITLE	CITED BY	YEAR
JUICE, software for vegetation classification L Tichý Journal of vegetation science 13 (3), 451-453	1353	2002



From: Lubomír Tichý <tichy@sci.muni.cz>
Sent: Tuesday, 6 December 2016 9:50 PM
To: Sarah Luxton
Subject: Re: Data limitations JUICE 7.0.45

Hi Sarah,

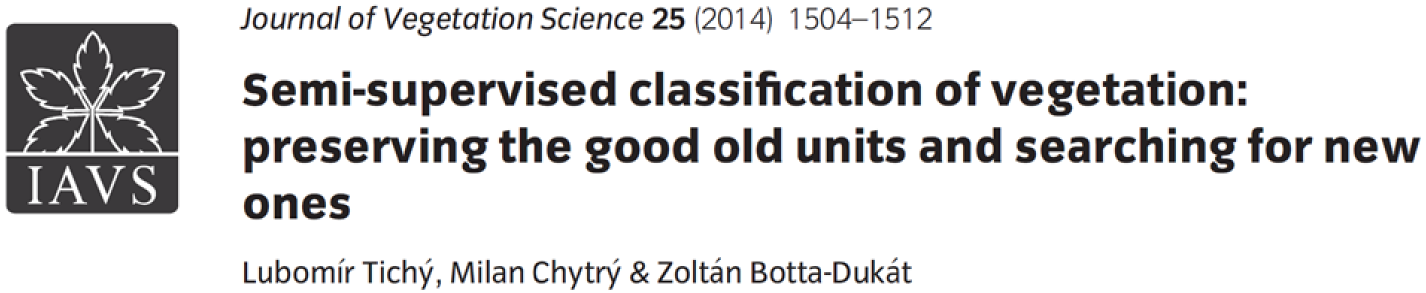
the JUICE program is currently working with the matrices of 1.3M of plots with 30k of species. Your size is as "a small" data set. :-).
Be careful, in options there is a switch to increase the maximum defined size of the data set, which is predefined to 30000 plots and 5000 sp
The size must be changed before the data import.

BW. Lubos

We have the capacity for big analyses, what Q's do we want to ask?

Going forward…

- **Tichy et al. (2014)**
  - Aim to create flexible classifications.
  - Can keep old units, but incorporate new data.
- **IAVS 2018 –** Rethinking biomes… and…**?**



Journal of Vegetation Science **25** (2014) 1504–1512

Semi-supervised classification of vegetation: preserving the good old units and searching for new ones

Lubomír Tichý, Milan Chytrý & Zoltán Botta-Dukát

Workshops

Rethinking biomes – towards a consistent high-level classification of global vegetation

Acknowledgments

- Australian Government 'Research Training Program' Scholarship
- Alcoa of Australia
- The 88+ botanists who collected data over 25 years

ALCOA

Curtin University

AUSTRALIA
Australian Government