SMARTSkills Workshop

Reproducible Analyses and Data Management

Cóilín Minto
Marine and Freshwater Research Centre
Galway-Mayo Institute of Technology

October 24th, 2013

Outline

Outline

Research is reproducible if it can be reproduced by others

Baggerly and Berry (2011)

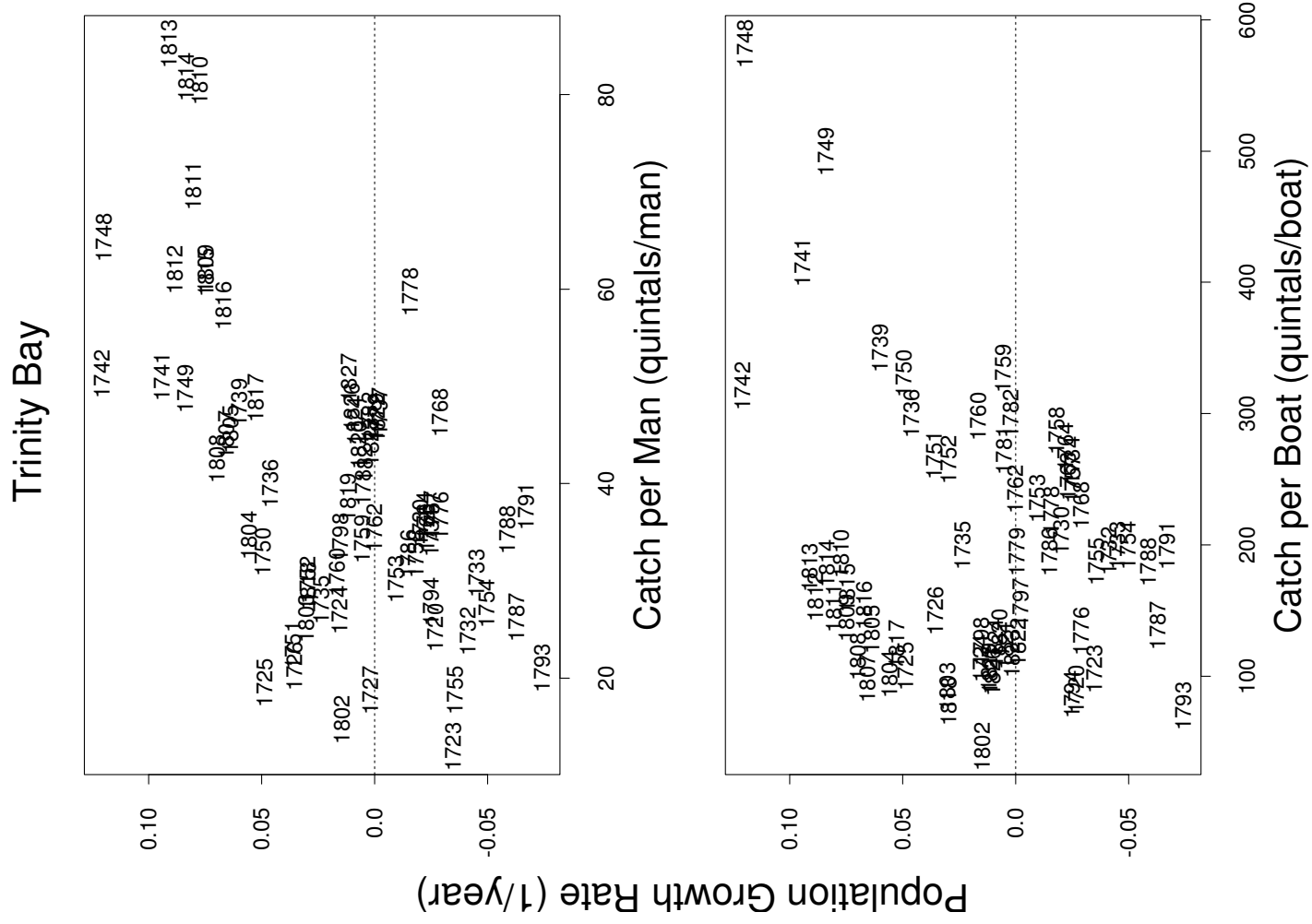
Image source: http://www.therooms.ca

Emigration to Newfoundland

Testing Ecological Models:
The Influence of Catch Rates on Settlement of Fishermen in Newfoundland, 1710-1833[1]

Ransom A. Myers

Emigration to Newfoundland



Myers (2001)

Emigration to Newfoundland

Analysis was:

- Conducted in 2000
- Run on a Sun server with documented (`READMEs`) folders containing:
  - Data
  - Text
  - Analysis code (S-Plus)
- Archived

Emigration to Newfoundland

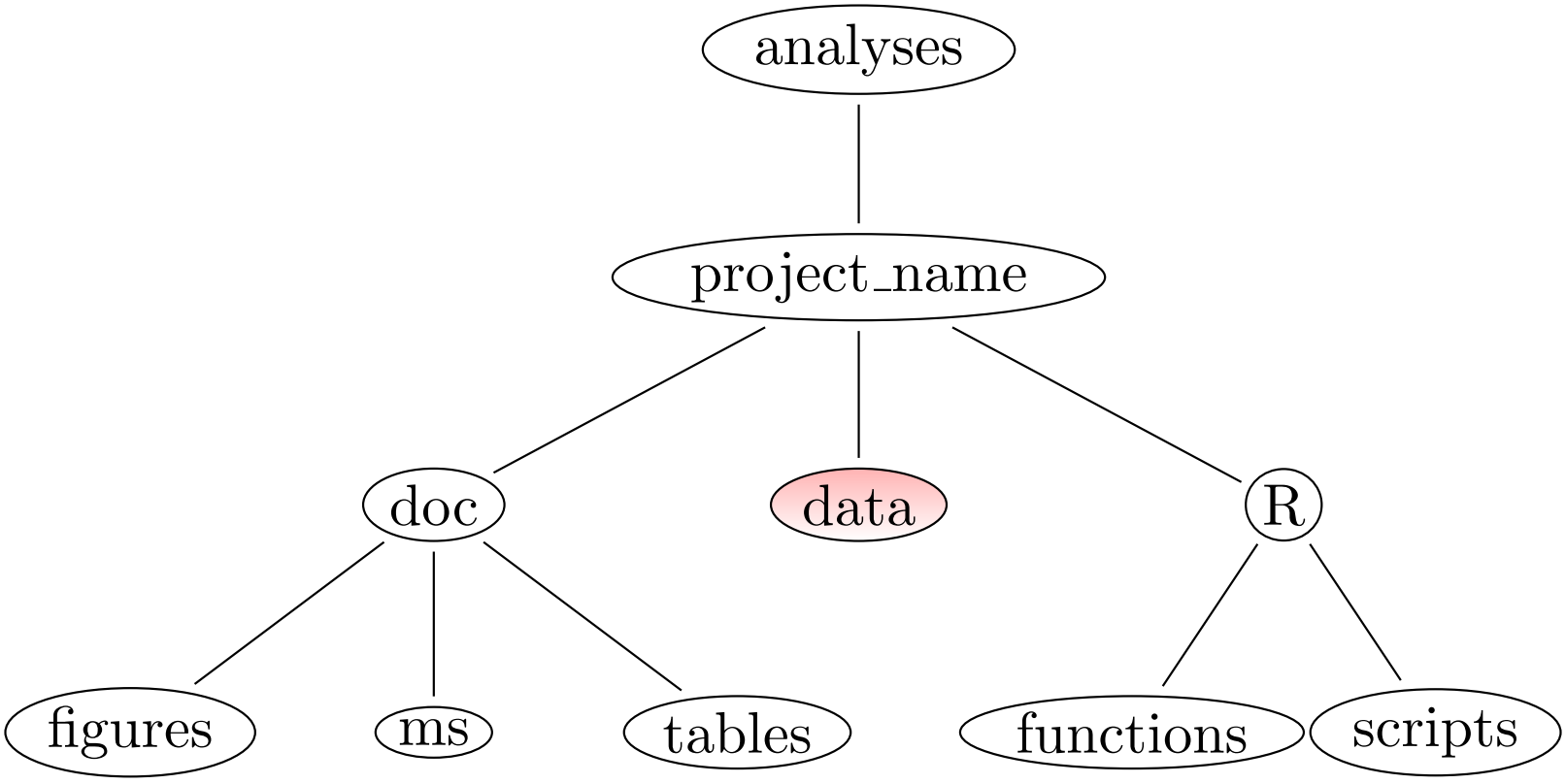
Year 2009:
Contacted by a Norwegian researcher wishing to re-run the analysis but the sole author (RAM) had very unfortunately passed away in 2007

In many cases, this would signal the end of the line and we go back to collating the data over-again or forget about it.
But in this case, three steps:

```
$ ssh server
$ cd relevant_folder
$ make
```

recovered the complete analysis, figure and table preparation and dynamicaly linked to a fresh write-up

Getting the structure right



Fastidious data management is paramount for reproducibility

Image source: moods of norway

Outline

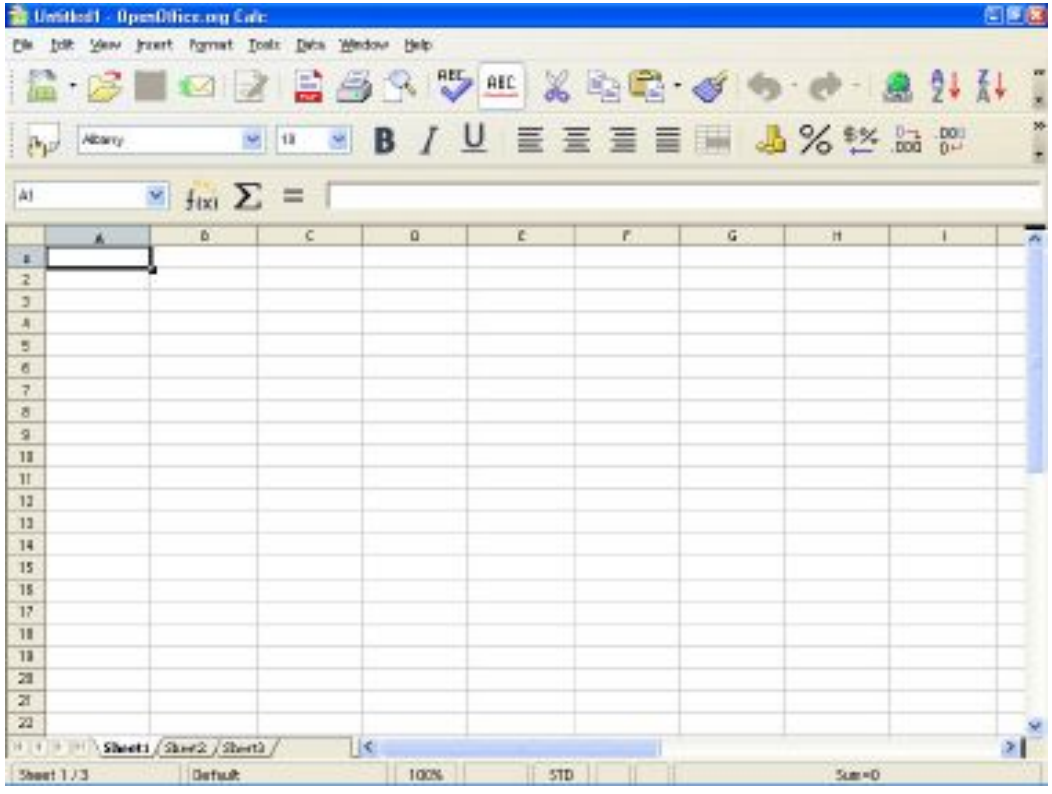
What's data?

Ultimately, a stored array of electrical charges but I like to think of data as the map and mode of transport that gets you from the start of a research project or program to the final product

Image source: http://www.deviantart.com

What's data?

It's not just a spreadsheet!



Data encompasses

- **Metadata** on what the work was about (who, what, where, when and why?)
- **Records** Measurements, dates, treatments, etc.
- **Code** Data extraction and analysis
- **Results** (value-added collections of records) Figures, tables, calculations
- **Reporting** Documents, mark-up

Losing our way

In science we often lose our map and mode of transport via:

- Damage to files or storage device
  `Error: cannot open ...`
- Purported storage device ageing or becoming redundant
  "That was three laptops ago"
- Software change
  House of punch-cards
- Personnel change
  "They left with the laptop"
- Bounce to the next project

Why do some scientists treat data poorly?

Among other reasons:

- Incentive potentially lacking in highly competitive publishing arena
- Focus on the publication as self-contained product of the business
- Data husbandry viewed as diminishing returns
- Shoulders of giants mis-interpreted
- Illusion of ownership

Why these reasons don't cut the mustard now

Among other reasons:

- Large collaborative initiatives consisting of many sub-projects necessitate data management
- Journal publishing ethics changing and valuing data husbandry, e.g.,
  - Debes PV, Fraser DJ, McBride MC, Hutchings JA (2013) Multigenerational hybridisation and its consequences for maternal effects in Atlantic salmon. Heredity 111: 238-247. doi:10.1038/hdy.2013.43
  - Debes PV, McBride MC, Fraser DJ, Hutchings JA (2013) Data from: Multigenerational hybridisation and its consequences for maternal effects in Atlantic salmon. Dryad Digital Repository. doi:10.5061/dryad.9cs2v
- Granting bodies requesting data management planning

Data management

ONCE COLLECTED AND ELECTRONICALLY ENTERED
DON'T TOUCH THE DATA!

Tempting as it might be to fire up a spreadsheet and start creating worksheets and *pasting specially*, this will only lead to data woes

To avoid wondering whether

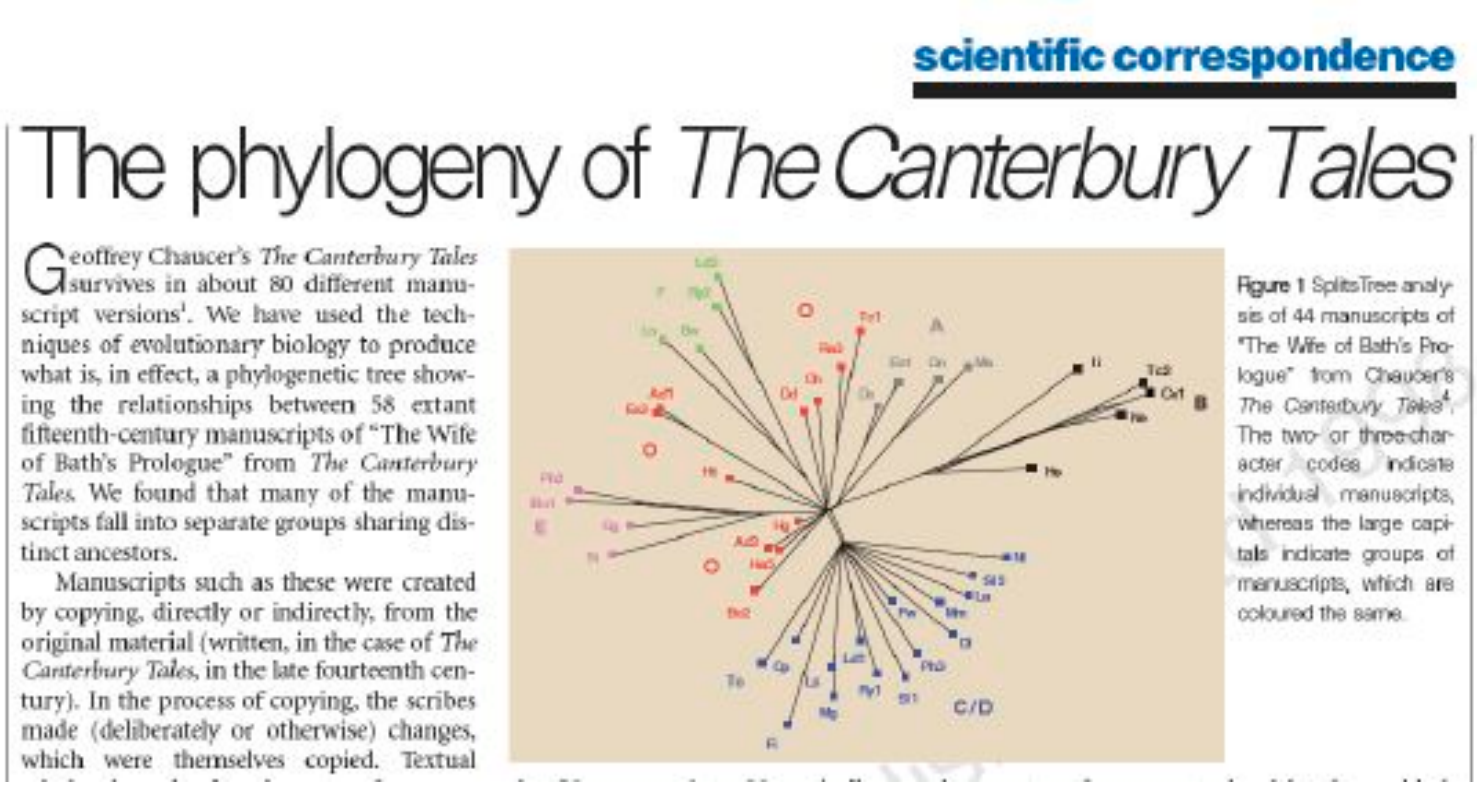
`data_new.xls`

or

`data_updated.xls`

is the relevant copy, leave the data in the data folder or repository alone

Data management: Spreadsheet Tales



scientific correspondence

The phylogeny of *The Canterbury Tales*

Geoffrey Chaucer's *The Canterbury Tales* survives in about 80 different manuscript versions[1]. We have used the techniques of evolutionary biology to produce what is, in effect, a phylogenetic tree showing the relationships between 58 extant fifteenth-century manuscripts of "The Wife of Bath's Prologue" from *The Canterbury Tales*. We found that many of the manuscripts fall into separate groups sharing distinct ancestors.

Manuscripts such as these were created by copying, directly or indirectly, from the original material (written, in the case of *The Canterbury Tales*, in the late fourteenth century). In the process of copying, the scribes made (deliberately or otherwise) changes, which were themselves copied. Textual

Figure 1 SplitsTree analysis of 44 manuscripts of "The Wife of Bath's Prologue" from Chaucer's *The Canterbury Tales*[4]. The two- or three-character codes indicate individual manuscripts, whereas the large capitals indicate groups of manuscripts, which are coloured the same.

"In the process of copying, the scribes made (deliberately or otherwise) changes, which were themselves copied."

Barbrook *et al.* (1998). *Nature* (394) p.839.

Data management: solution

All data manipulations should be done programmatically

- Read raw data in analytical software
- Subset, remove, adjust via code
- Leaves a reproducible trail and
- Leaves the original (hard-won) data intact
- Pipe results dynamically into your document (e.g., `Sweave`, `knitr`)

A contention

My over-arching contention with the status quo is that an individual's laptop or PC is not an acceptable research environment, as it:

- Risks complete data loss
- Fosters the "Chaucer" effect (more later)
- Is anti-collaborative
- Is license hungry and therefore costly
- Is less powerful, slower

A back-to-the-future solution

Need to return to a common research environment - the server

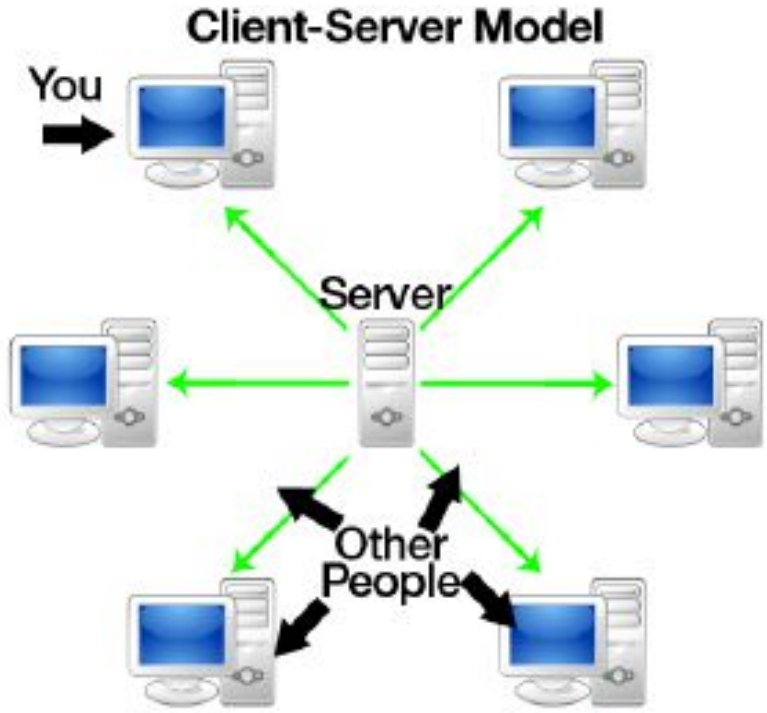


Image source: http://my.opera.com

A back-to-the-future solution

In as much as we have the focal point of the wet-lab to process specimen samples, we should have a central place for data storage and processing, as it:

- Keeps single copies of data centrally
- Has a longer life than the project
- Has a longer life than the researchers (??)
- Gives everyone equal access to high-performance architecture (no need a new laptop, just use laptop for)
- Managed centrally

A back-to-the-future solution

Many institutes have servers but rarely used as a common research environment outside of the physical sciences

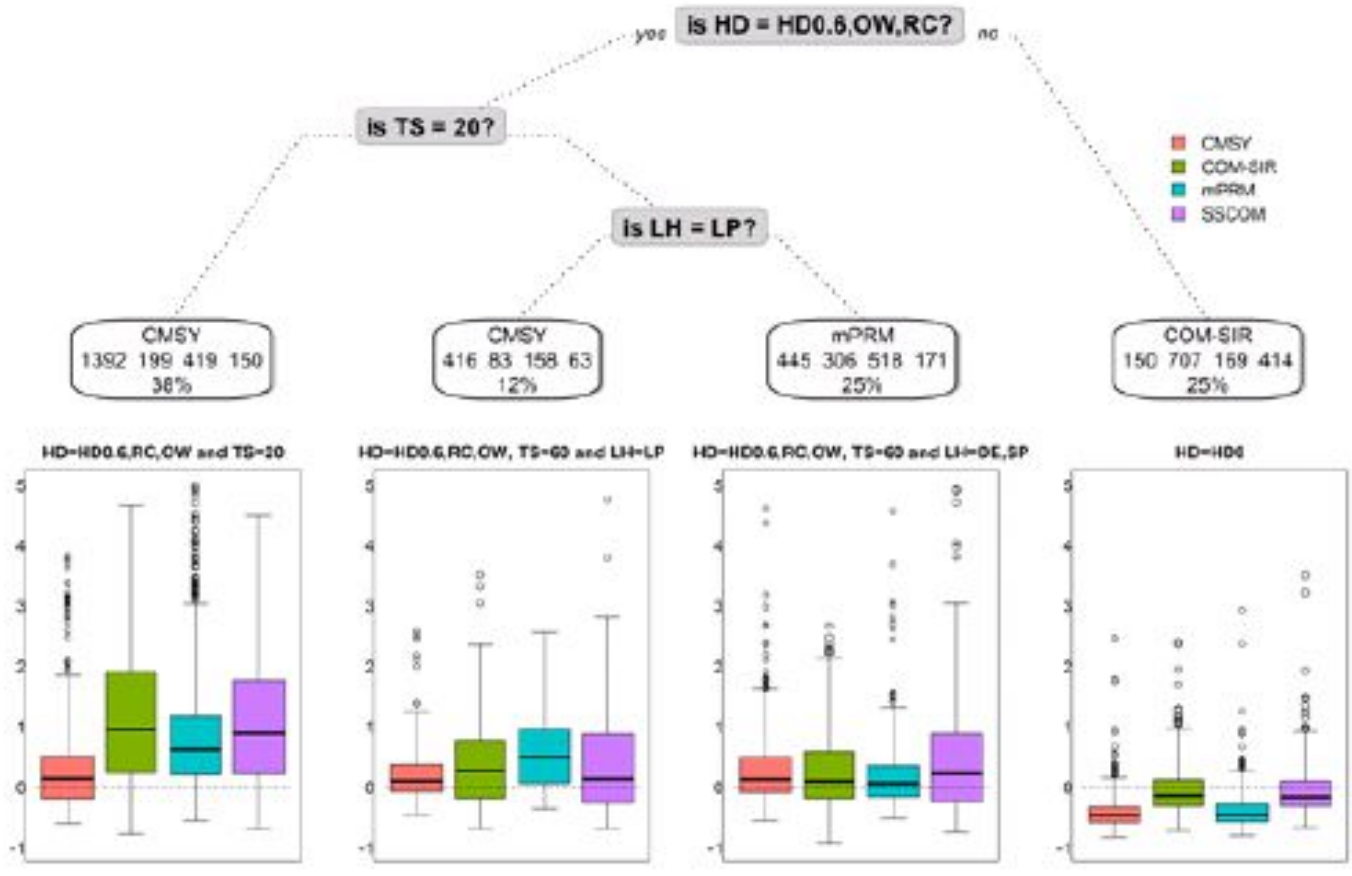
But the coming of age of high-performance computing now necessitates that we make the move back

Example: data-poor stock status

- FAO and Conservation International project to globally assess status of "data-poor" stocks
- Two research teams from 8 different countries
- Had to work in a central environment - `hexagon.bccs.uib.no` in Bergen, Norway.

Example: data-poor stock status

- 576 (scenarios) x 10 (iterations) x 4 (methods) = 23,040 stock assessments
- For agreed convergence level (MCMC,SIR) requires 19.5 CPU years on single processor
- Completed work in walltime of 7.5 days on Hexagon cluster



Outline

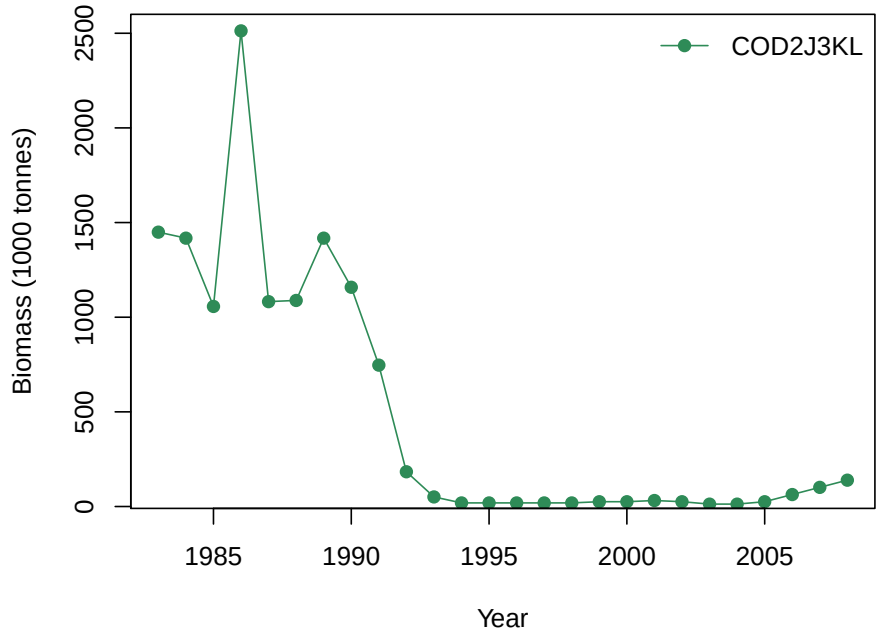
Original database

Ransom Myers' Stock Recruitment Database

- Approximately 640 stocks.
- Used in many publications on fish population dynamics, e.g.
  - Relationship between recruitment and spawning stock size
  - Density dependence
  - Depensation (Allee effects)
  - Productivity rates across taxa
  - Patterns of depletion and recovery
- Housed in flat text files
- Archived (not updated anymore) version available from: `http://www.mscs.dal.ca/~myers/welcome.html`

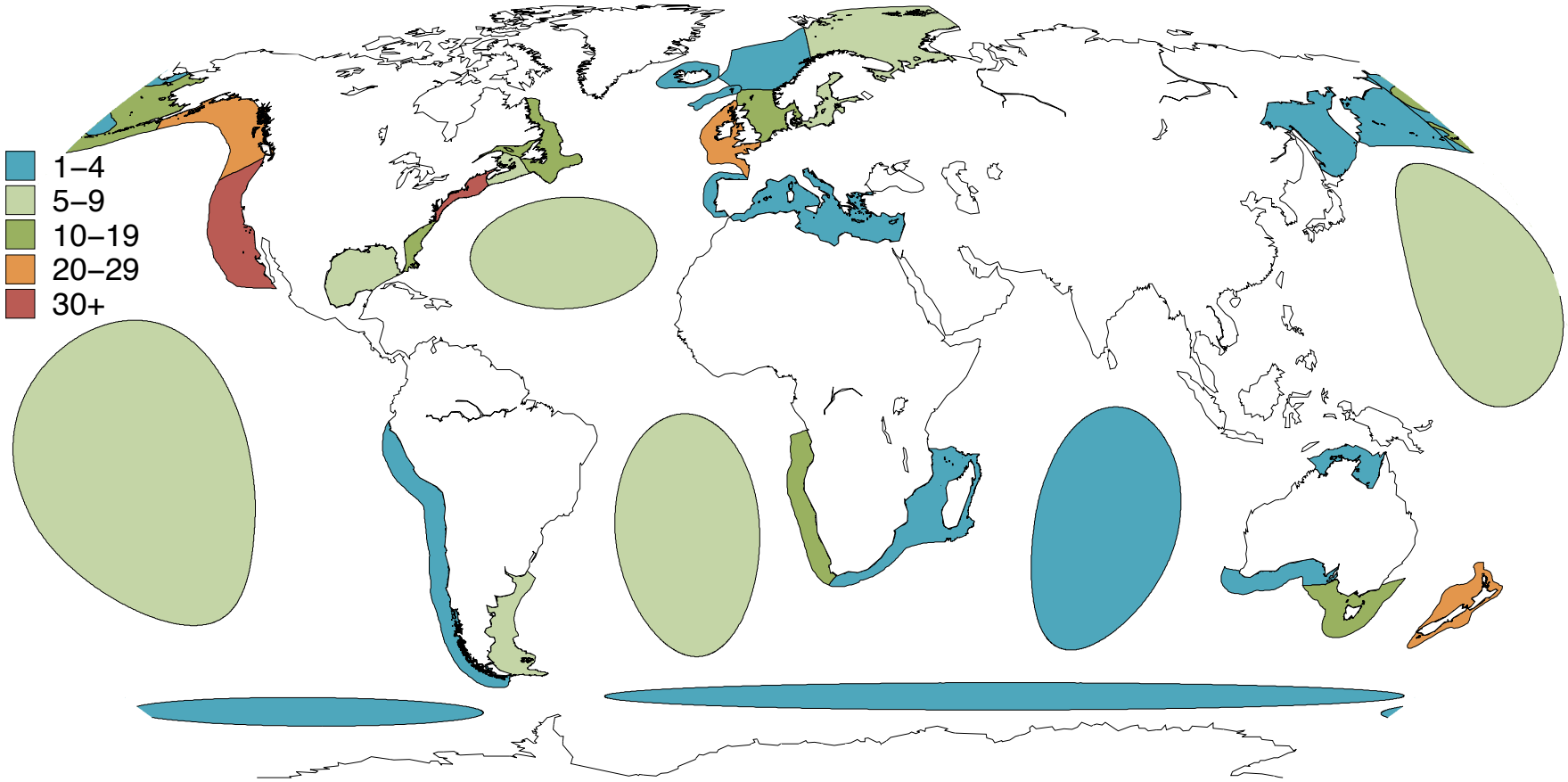
Why an updated database?

- Many stocks 15 years out of date
- New data often at:
  - Low population levels
  - Reduced fishing intensities

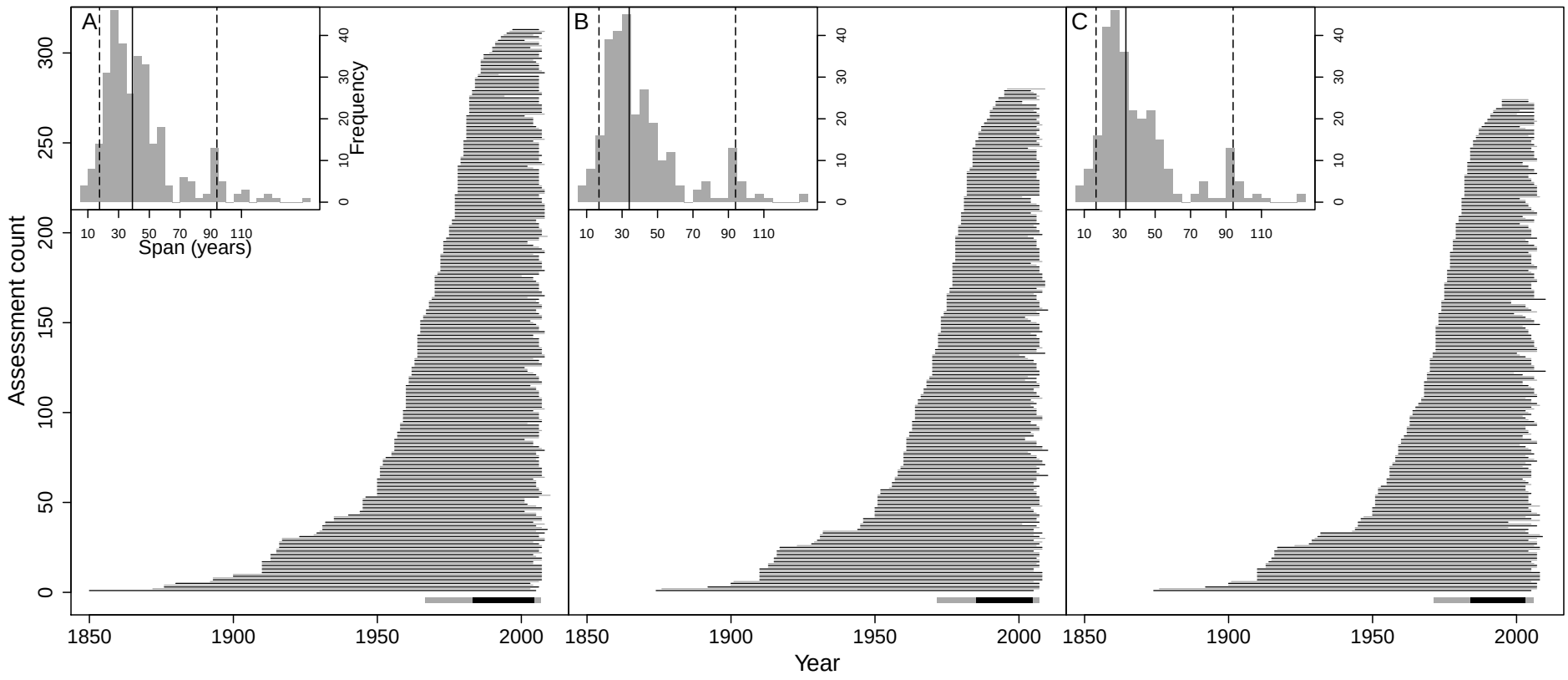


- Interest in:
  - Effects of exploitation on trends in abundance across taxa from many ecosystems
  - Efficacy of harvest policies
  - Recovery trajectories post fishing mortality reductions
- Relational database to support reproducible analyses

Geographic coverage



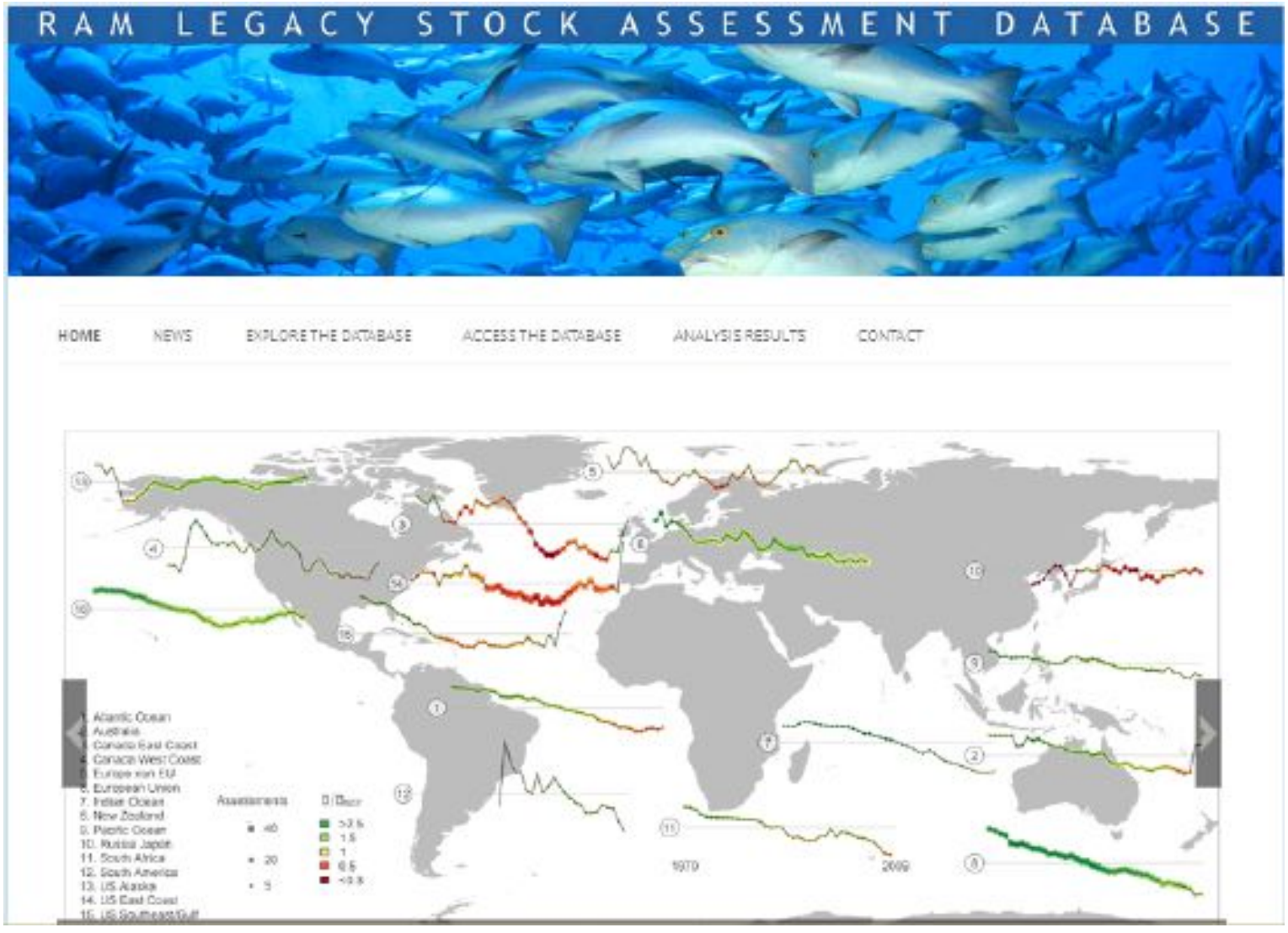
Temporal coverage: orca plots



Taxonomic coverage

356 assessments

Order	N	%
Gadiformes	71	20
Perciformes	66	19
Pleuronectiformes	57	16
Scorpaeniformes	45	13
Clupeiformes	36	10
Invertebrates	42	12



Used in 27 publications since inception in 2009

`http://depts.washington.edu/ramlegac/`

Outline

Summary

- Reproducibility is a central component of science
- To-date our general approach to data has been poor bordering on careless
- Scale of the problems and collaborations now necessitate change for the better
- A laptop/desktop is not a research environment
- Data management increasingly recognized
- Putting in the spade work of data management can reap good rewards

Aknowledgements

Paulha McGrane and John Boyd and organizing committee

Julia Baum
Deirdre Brophy
Olaf Jensen
Ray Hilborn
Rick Officer
Daniel Ricard
Conservation International
University of Washington
FAO
Marine and Freshwater Research Centre, GMIT, Galway
Dalhousie University, Nova Scotia

GMIT
GALWAY-MAYO INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA NA GAILLIMHE-MAIGH EO

DALHOUSIE UNIVERSITY
Inspiring Minds

NDP
National Development Plan 2007 - 2013

European Union
European Regional Development Fund
Investing in your future

Baggerly, KA and Berry, DA (2011). Reproducible Research. Amstat News January 2011.

Myers, RA (2001). Testing ecological models: the influence of catch rates on settlement of fishermen in Newfoundland, 1710-1833. Research in Maritime History, 21, 13-29.