

Aiming for impact as a Research Software Engineer in Biology

Edward Wallace, @ewjwallace

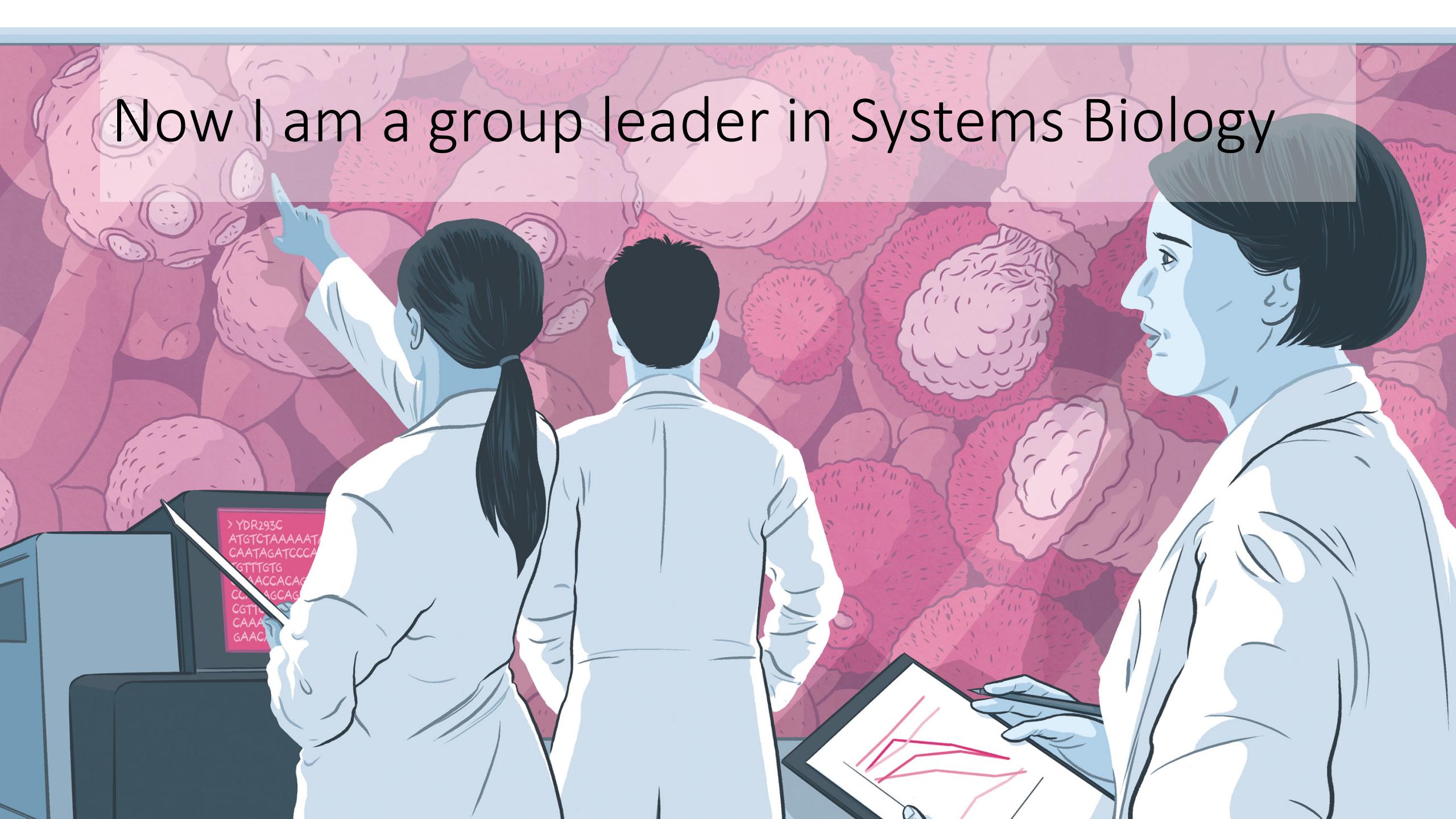
Community of Edinburgh Research Software Engineers, 20 Feb 2020



THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences

Am I a research software engineer?

- Mathematics PhD, University of Chicago 2005-10
 - *stochsimcode*, MATLAB for stochastic simulations of neural networks, *PLoS Computational Biology*
- Systems Biology postdoc, Harvard 2010-13
 - *codonFits*, bad R package for evolution of protein-coding sequences, *Molecular Biology and Evolution*
- Biochemistry postdoc, U. Chicago 2013-15
 - R code for analysing/visualising protein aggregation, *Cell & Dryad*
- Informatics / Cell Biology fellow, Edinburgh 2016-17
 - R code for analysing RNA splicing data, *RNA*



Now I am a group leader in Systems Biology

Now I am a group leader in Systems Biology

We have software projects in the lab

- *riboviz*, bioinformatics pipeline for processing data measuring protein translation (ribosome profiling)
 - EPCC collaboration with Dr. Mike Jackson
- *tidyqpcr*, R package for tidy quantitative PCR analysis
 - eLife Open Innovation Leaders programme
- Routinely trying to analyse all our data in reproducible ways!
 - Mostly in R with Rmarkdown
 - Some python
 - Shell and others as needed

Am I *still* a research software engineer?

- I have less time to code than I used to
 - I go to meetings, run my lab, write papers & grants, teach
- Everyone in my research group needs to code
 - Even wet-lab biologists need to wrangle and plot their data
- Actually, all research biologists need to learn how to code
 - Reproducibly, reliably, efficiently
 - How are they going to learn?

**How can I promote good practices in research software,
when I am writing less code myself?**

Some personal reasons to care about impact

- I'm a research fellow in biology, funded by Wellcome & RS
 - Wellcome officially supports open science
- My work relies on re-using other people's data (past me!)
- Biology is a data-intensive discipline
- Better Software, Better Research - <https://www.software.ac.uk/>
- Training & impact helps get grants funded
- I've learned from others' free training materials
- It just upsets me to see bad data analysis



So what am I doing about it?

- Still working to improve my own skills and to be more efficient with scarce coding time
- Sharing code in better packages
- Helping the people I work with to improve their skills
- Working with the Carpentries to train research computing skills
 - Community-led teaching with open-source materials
 - Edinburgh carpentries is UK's biggest chapter <https://edcarp.github.io/>
- Working strategically to improve research computing training
 - School of Biological Sciences computing survey

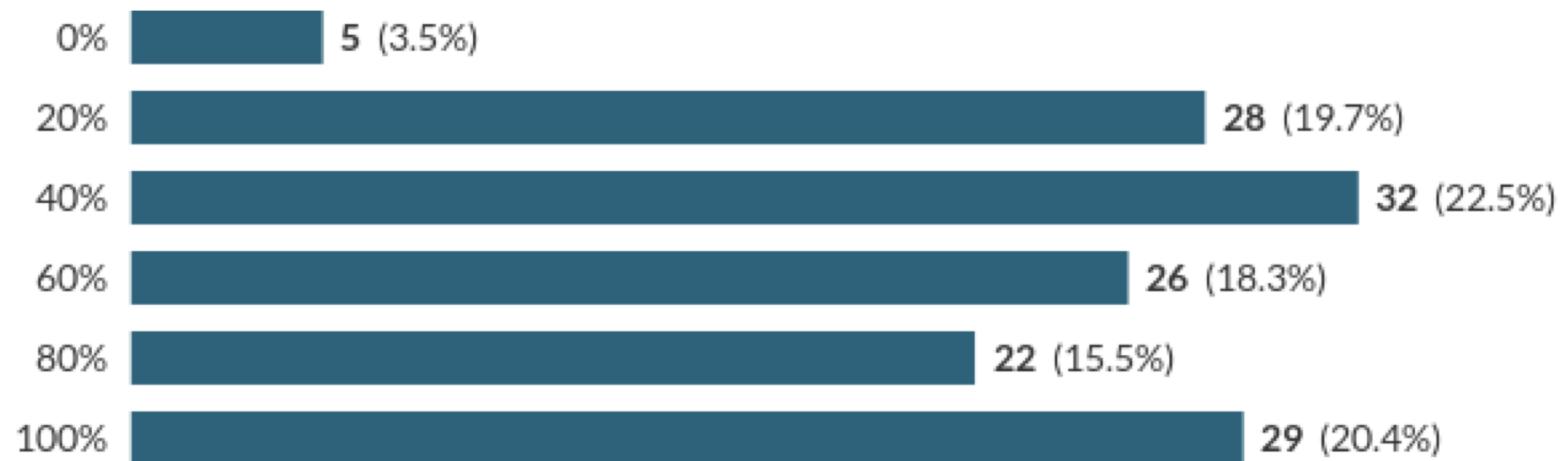
Finding out what biologists need: the SBS research computing survey

Goals:

- Inform research computing training for students, staff, faculty
 - Find out what data & software people use
 - Find out what skills & training they think they need
 - Input to UKRI/BBSRC data-intensive bioscience review
-
- We used <https://www.onlinesurveys.ac.uk/>
 - Designed 1-page survey completable in 5 minutes, April 2019
 - We can share the survey design for **you** to adapt

Who filled in the survey?

- We had **147** responses, about 25% response rate
 - 35 Group Leaders (out of 130)
 - 40 Postdocs, 56 PhD students, 16 RA/other
 - Responses from many institutes & subfields
 - Self-selecting!
- Computing is required as **proportion of success of most projects:**



Many of us do not have formal training

In computing:

No  49 (45.4%)

Yes  38 (35.2%)

Would like to take  21 (19.4%)

Or statistics:

No  38 (34.5%)

Yes  52 (47.3%)

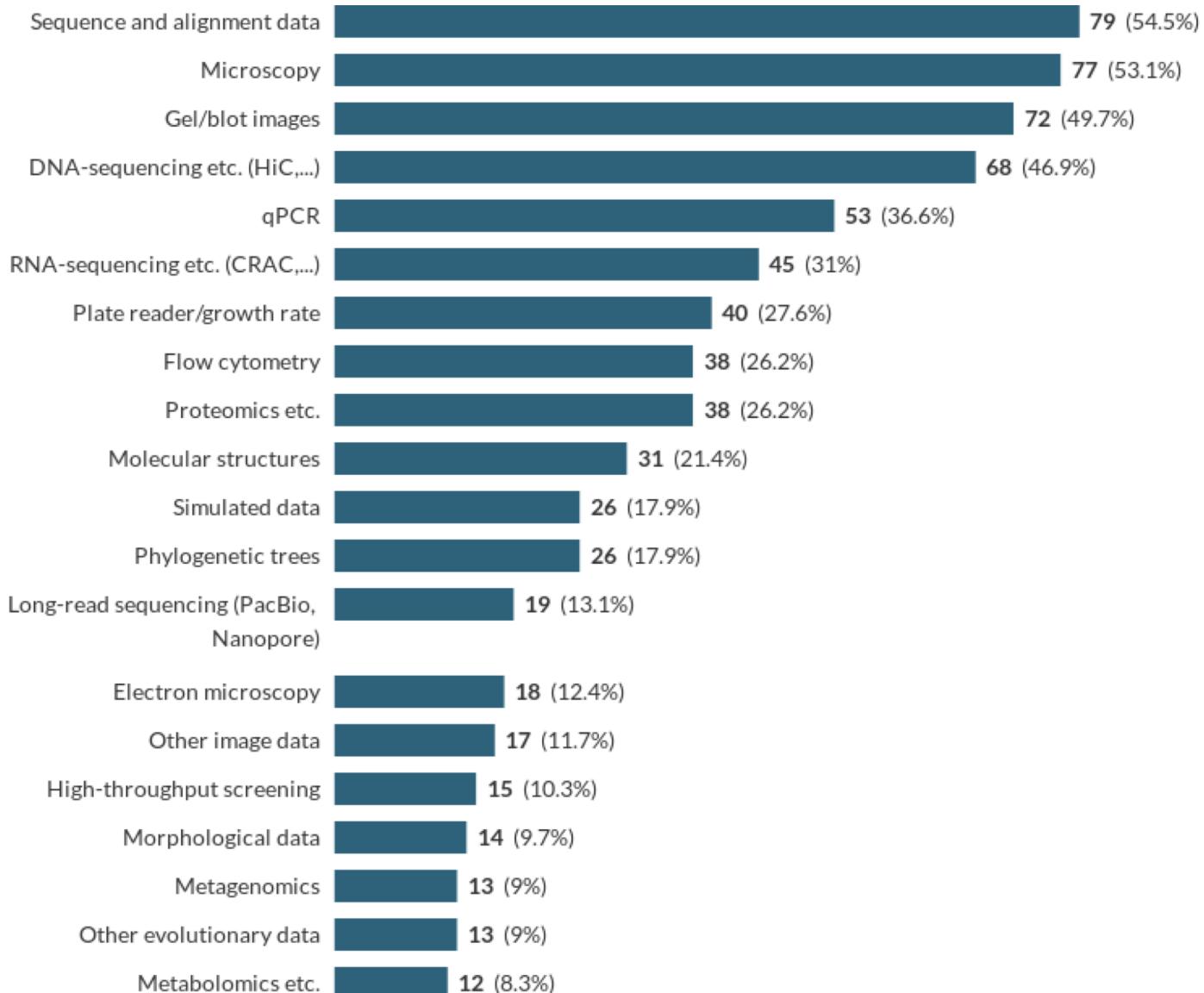
Would like to take  20 (18.2%)

What is your biggest frustration in research computing?

A word cloud visualization where the size of each word represents its frequency or importance in the responses. The most prominent words are 'not', 'time', 'software', 'data', 'learn', 'having', 'support', and 'lack'. Other visible words include 'understanding', 'computer training', 'doing', 'how', 'ask', 'skills', 'do', 'R', 'it', 'softwares', 'that', 'using', 'enough', 'don't', 'find', 'when', 'statistical', 'coding', 'bioinformatics', 'python', 'what', 'work', 'need', 'windows', 'no', 'running', 'analysis', 'such', 'knowledge', 'knowing', 'know', 'where', 'use', and 'this'.

not
time
have
being
computer
training
doing
which
how
data
learn
having
support
lack

We use many kinds of biological data



Correspondingly diverse software:
MS Excel, SPSS, R, python,
MATLAB, ImageJ, ImageStudio,
Genome Browsers, Benchling,
Snapgene, Pymol, BLAST, multiple
sequence alignment, FlowJo, ...

What is your biggest need in computing training?

processing
datasets
analysing
more
using
like
tools
imaging
advanced
graphs
running
rnaseq
matlab
bioinformatics
statistics
R
image
seq
analyses
etc
none
mass
rna-seq
how
systems
use
ngs
imagej
linux
eg
software
line
training
with
skills
learning
machine

command
proteomics
documentation
would
nsgs
using
nsgs
rnaseq
sequence
python
data
analysis
programming

What is your biggest need in computing support?

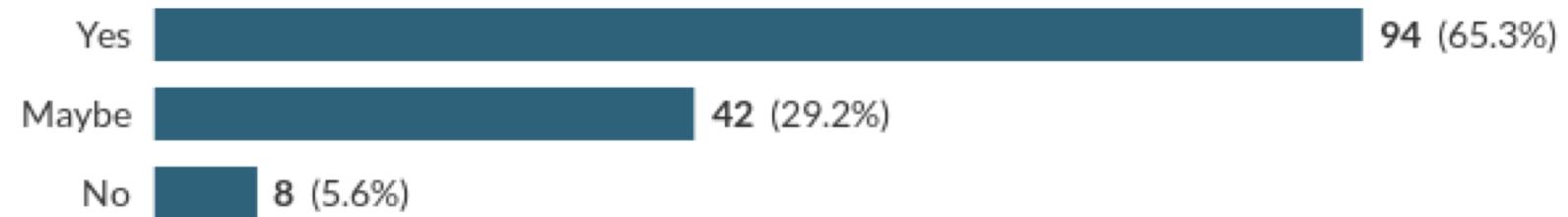
collaboration
access help
online as when resources but
available
like more bioinformatics control
data training
courses always how
free that
just have time photoshop
are learn so i teaching many
R statistics prism version use
different good much need
support python
analysis probably
software

Which statistics methods or topics would you like to learn?

A word cloud centered around statistics learning topics. The most prominent words are "statistical", "data", "analysis", "how", "statistics", and "tests". Other visible words include "linear", "datasets", "modelling", "multivariate", "glm", "but", "bayesian", "choosing", "learning", "machine", "with", "them", "models", "what", "stat", "it", "mixed", "should", "methods", "more", "R", "would", "when", and "use".

Are you likely to take courses offered within SBS?

In data analysis, bioinformatics, or image analysis:



Would you like more training in statistics?



What did we learn? What should we do?

Summary

- Biologists **rely on quantitative data** and analysis (sequences, microscopy, gels, RNA-seq,...).
- Frustrations center around “**not knowing what to do**” and “**do not know whom I can ask**”.
- Huge demand for research computing training, especially **R**, **python**, and **ImageJ**.
- Also huge demand for **statistical training**, especially regression and Bayesian stats.

What did we learn? What should we do?

Summary

- Biologists **rely on quantitative data** and analysis (sequences, microscopy, gels, RNA-seq,...).
- Frustrations center around “**not knowing what to do**” and “**do not know whom I can ask**”.
- Huge demand for research computing training, especially **R**, **python**, and **ImageJ**.
- Also huge demand for **statistical training**, especially regression and Bayesian stats.

Action recommendations:

- Provide **training in computing** (Edinburgh carpentries) and **image analysis** (imaging network).
- Provide **statistics courses** – how? statistical consultancy unit?
- We need a **strategy** to effectively **connect people with help**.
 - Discussions about a “SBS bioinformatics facility”
- We would like to work cross-department and cross-college
 - Data Driven Innovation, Bayes Centre, EPCC, EdCarp, **YOU?**

Summary and next steps

Summary

- Biologists **rely on quantitative data** and analysis (sequences, microscopy, gels, RNA-seq,...).
- Huge demand for research computing training, especially **R**, **python**, and **ImageJ**.
- Also huge demand for **statistical training**, especially regression and Bayesian stats.
- Frustrations center around “**not knowing what to do**” and “**do not know whom I can ask**”.

Action recommendations:

- Provide **training in computing** (Edinburgh carpentries) and **image analysis** (imaging network).
- Provide **statistics courses** – how? statistical consultancy unit?
- We need a **strategy** to effectively **connect people with help**.
 - Discussions about a “SBS bioinformatics facility”
- We would like to work cross-department and cross-college
 - Data Driven Innovation, Bayes Centre, EPCC, EdCarp, **YOU?**

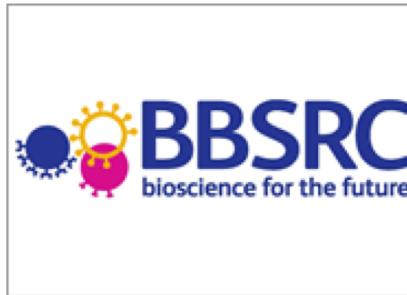
Impact is hard: it takes a community, goodwill, and lots of meetings!

Thank you!

Sign up here: <http://eepurl.com/gI4MsX>



- Edinburgh Carpentries
 - <https://edcarp.github.io/>
 - Giacomo Peru
 - Sean McGeever
 - Jen Daub
 - The whole community!
- The Carpentries
 - <https://carpentries.org/>
- Software Sustainability Institute
 - <https://www.software.ac.uk/>
- SBS Bioinformatics committee
 - Sara Buonomo, Al Ivens



@ewjwallace
<https://ewallace.github.io/>



THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences



2020 EdCarp Programme

SWC workshop at King's Buildings, 4 sessions, 21/01 – 18/01

SWC Workshop at Geosciences, 29-30/01

Data Carpentry Geospatial, TBC

Data Carpentry at Biology, 4 sessions, 7-20/04

Data Carpentry for Genomics, TBC

Data Carpentry for Social Sciences, 4 sessions, 12/02 – 4/03

Data Carpentry for Digital Humanities, 14-15/05

New organising committee

Now engaging:

cross-college, IAD, Bayes centre/DDI, doctoral programs