

integrating reproducibility into the undergraduate statistics curriculum

mine çetinkaya-rundel

weaving reproducibility through the undergraduate statistics curriculum

mine çetinkaya-rundel

#1

two-pronged approach

Convince
researchers to adopt
a reproducible
research workflow

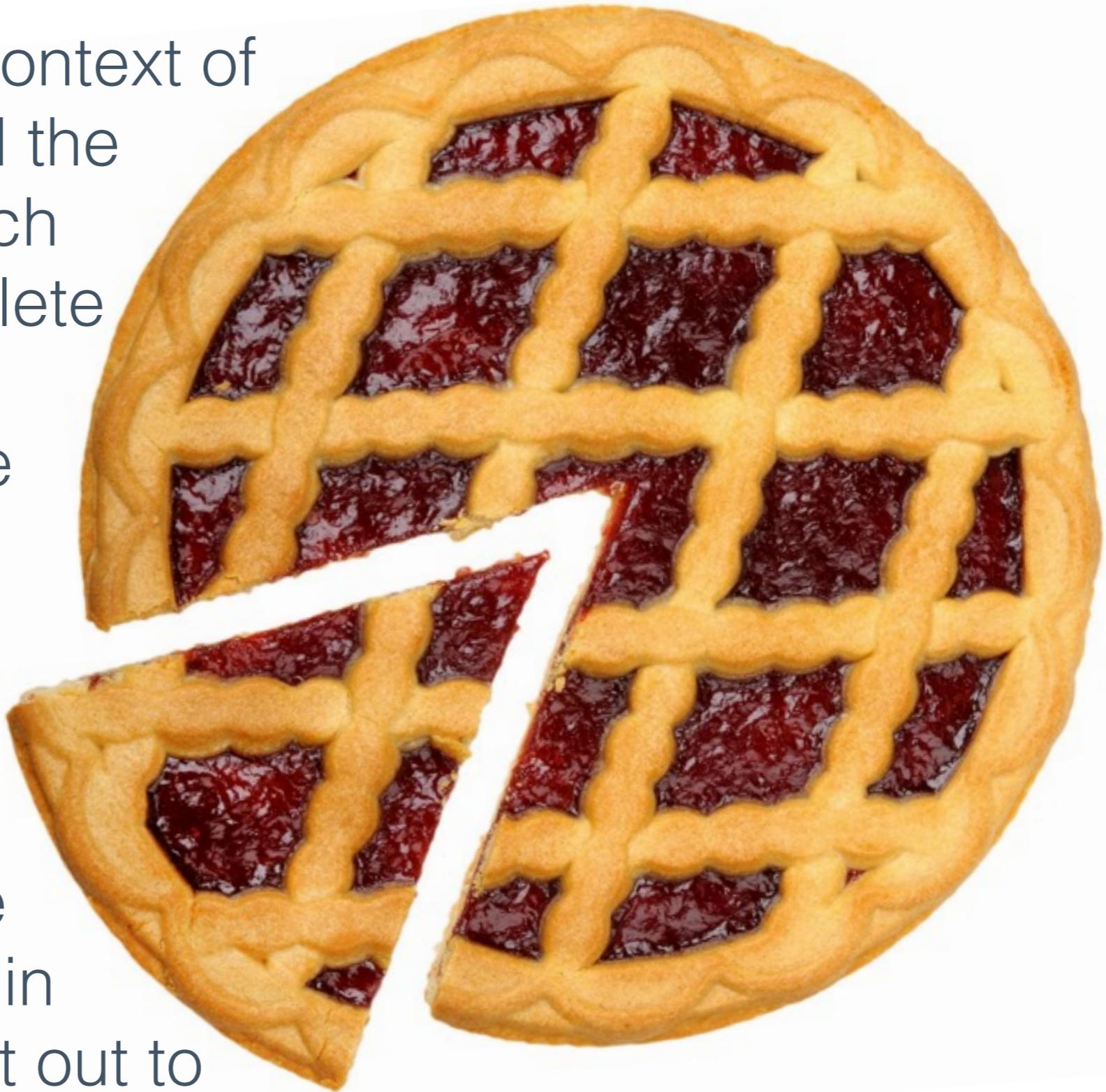
#2

Train new
researchers who
don't have any
other workflow



reproducibility

often comes up in the context of published research and the need to accompany such research with the complete data and analyses, including software/code



statistics

educators who teach data analysis should be instilling best practices in students before they set out to do research

current

future

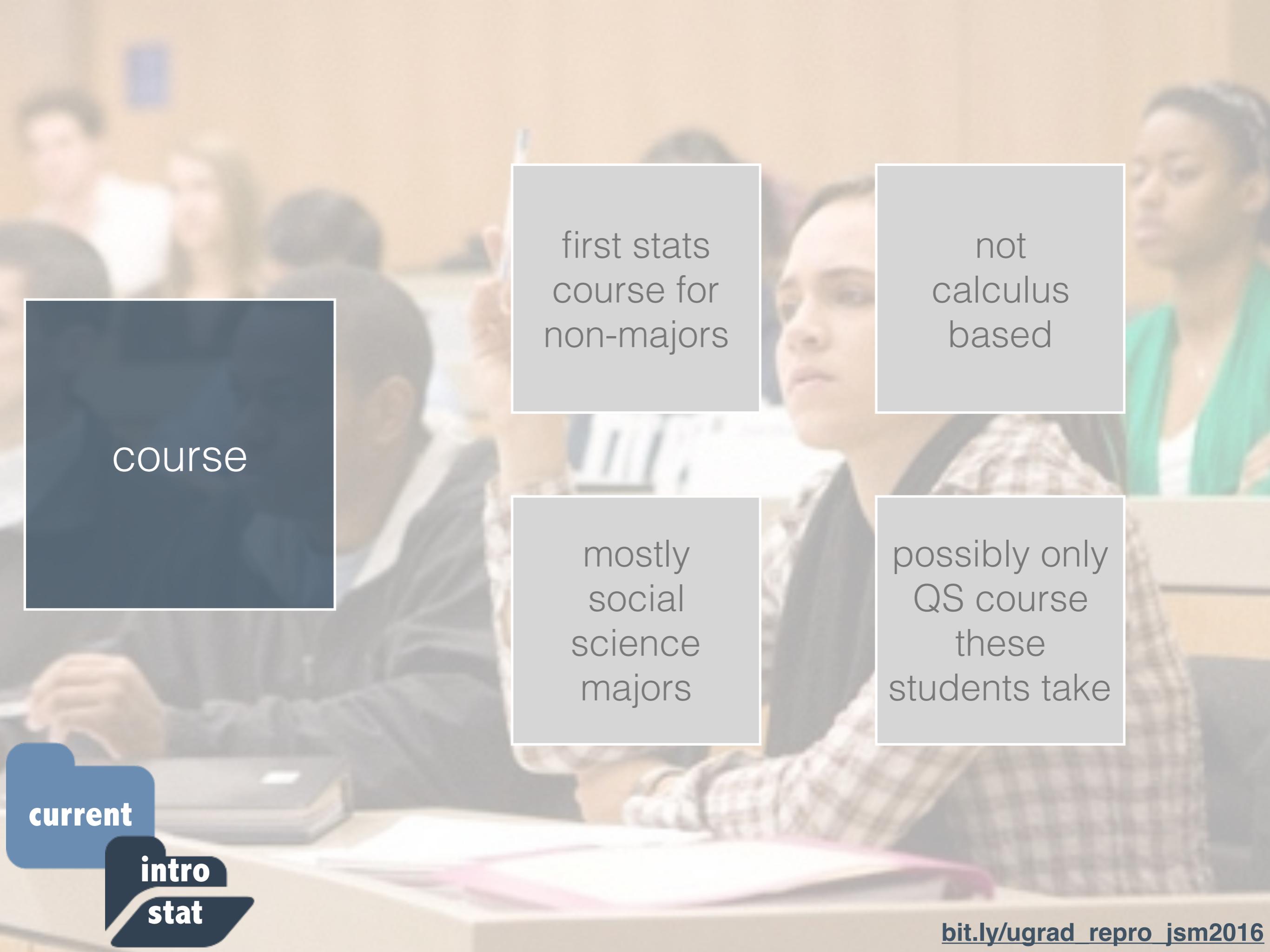
toolkit

**side
effects**

**intro
stat**

**intro
ds**

**stat
comp**



course

first stats
course for
non-majors

not
calculus
based

mostly
social
science
majors

possibly only
QS course
these
students take

current

intro
stat

A background image showing a collection of clear plastic petri dishes containing white bacterial cultures, arranged in a grid pattern.

reproducibility

literate
programming



current

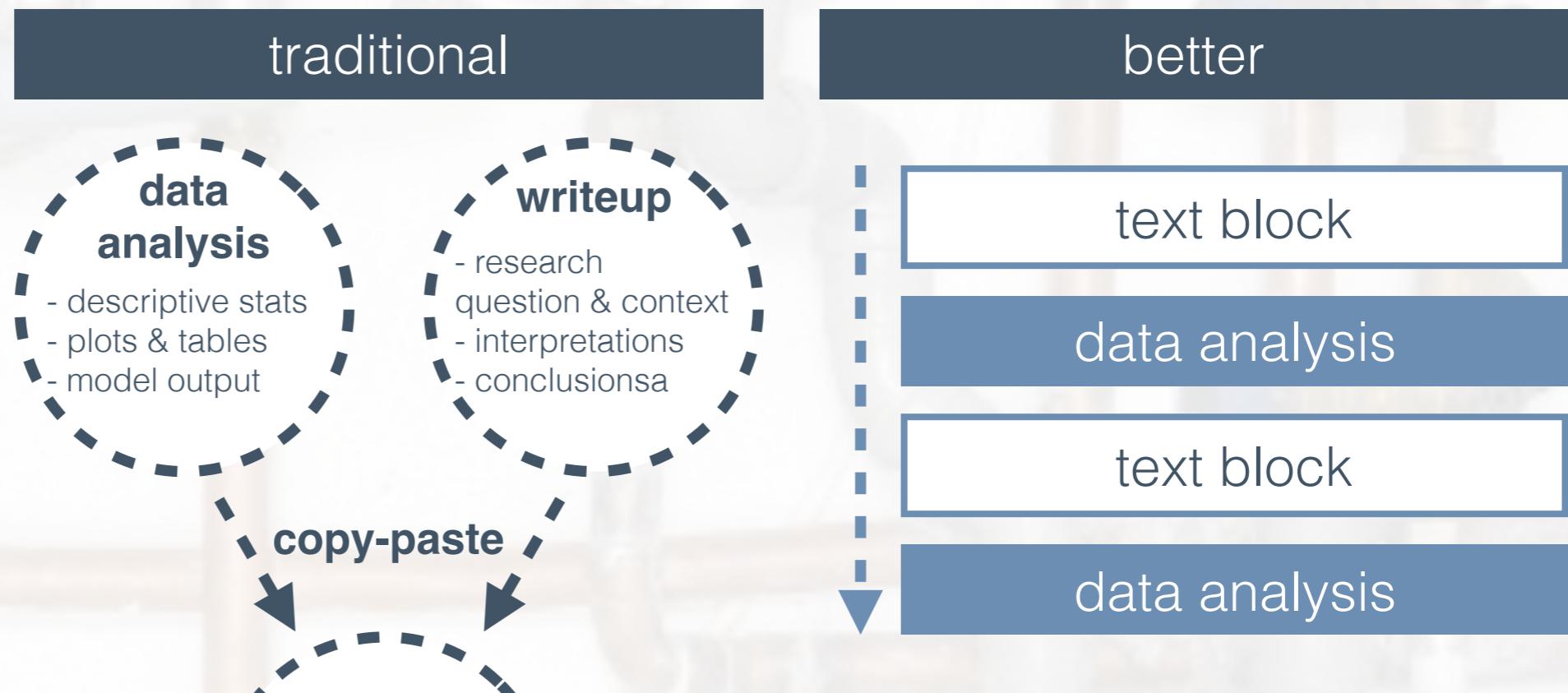
intro
stat

[bit.ly/ugrad repro jsm2016](http://bit.ly/ugrad_repro_jsm2016)

workflow

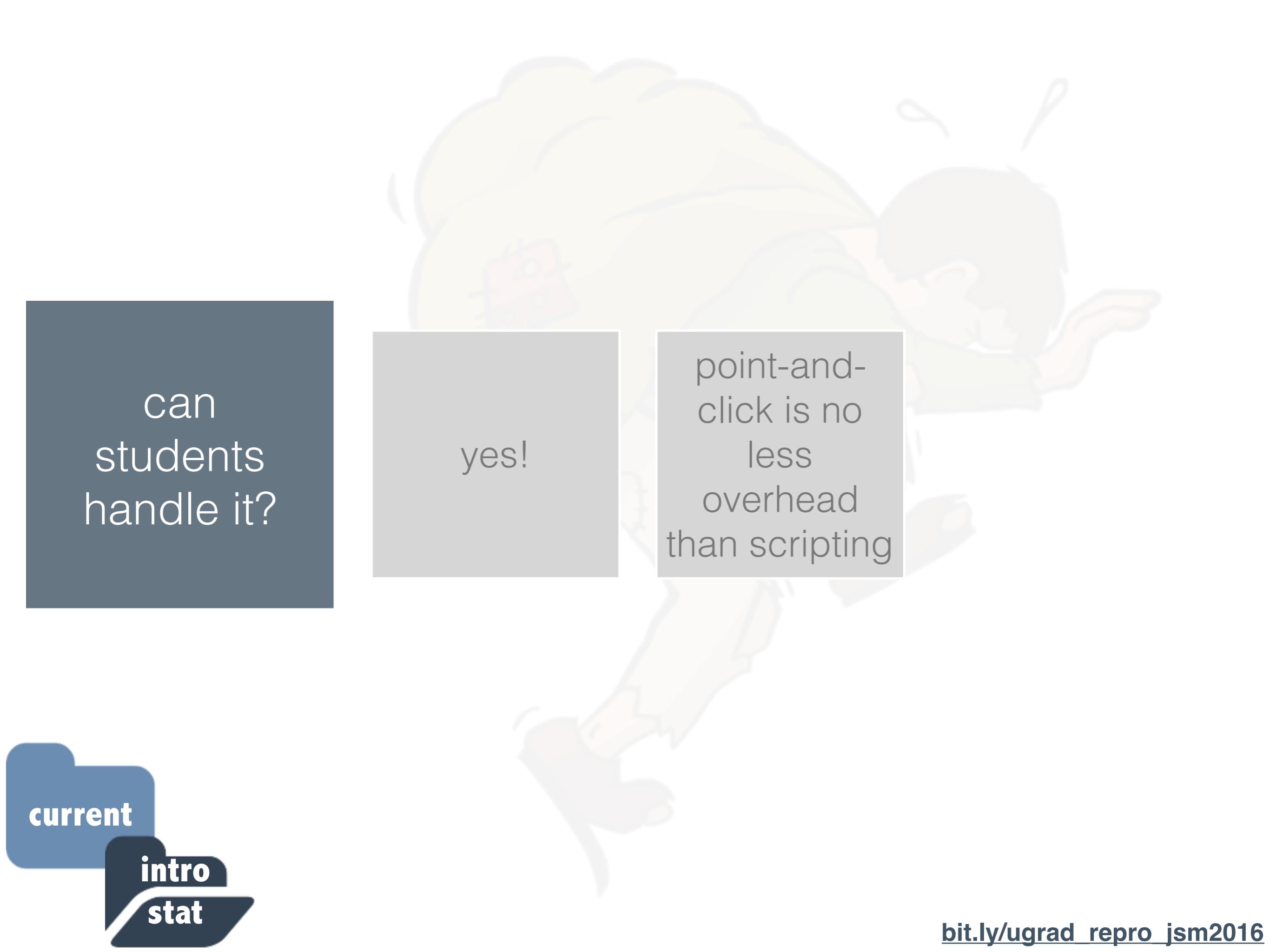
current

intro
stat



→ CMS → ✓

bit.ly/ugrad_repro_jsm2016



can
students
handle it?

yes!

point-and-
click is no
less
overhead
than scripting

current

intro
stat

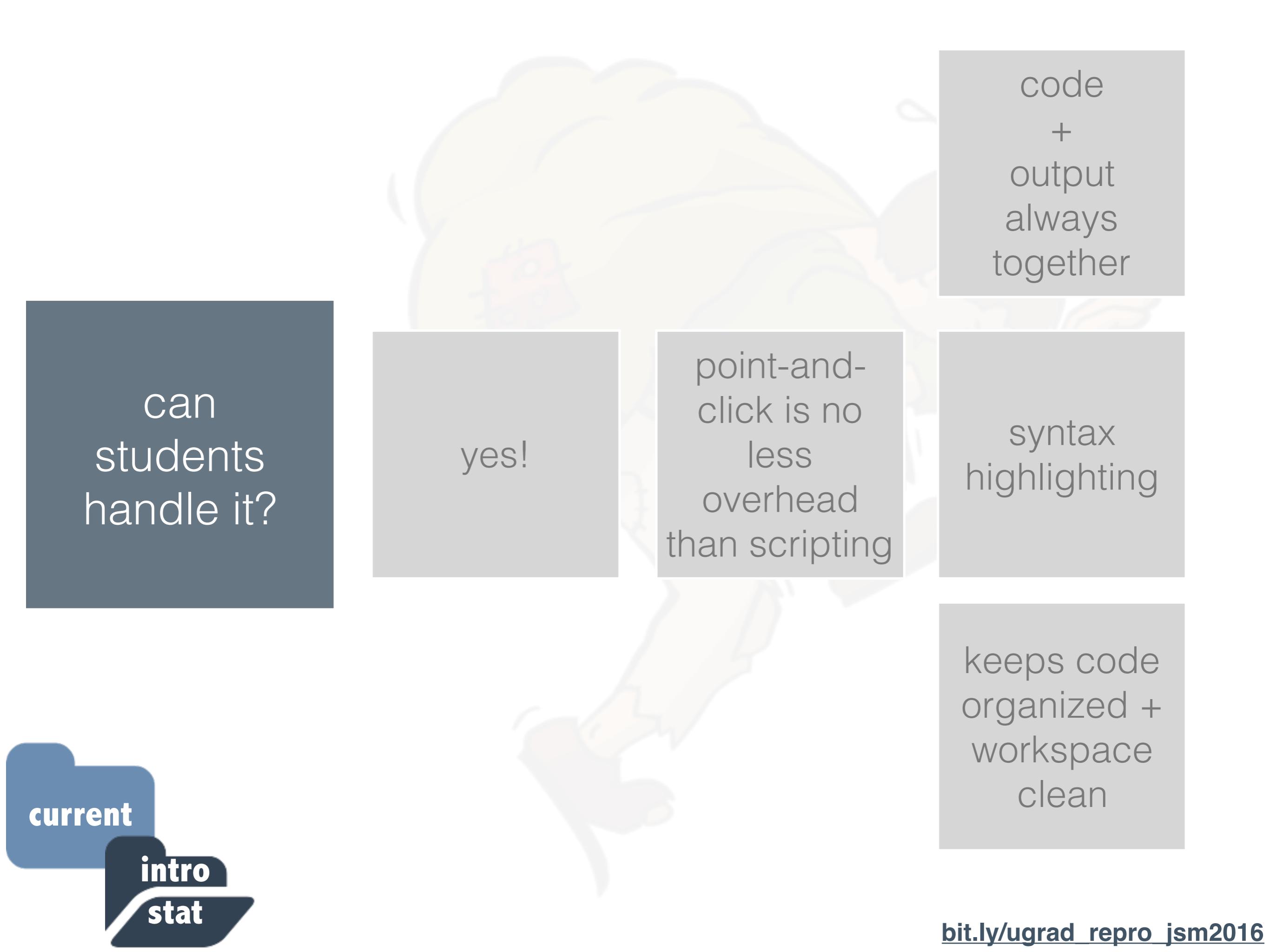
bit.ly/ugrad_repro_jsm2016

III. Adding Proportions to Summary Table

For categorical variables, you should see the counts of each possible outcome of that variable in the **Summary Table**. To see the breakdown of proportions or percentages, follow these steps:



- Click on the **Summary Table** to highlight it, click on the “**Summary**” drop-down menu and select “**Add Formula**”. In general, whenever you click and select a *Fathom* object (such as a **Table**, **Graph**, or **Summary**) the menu at the top of the screen will change to give you options for working on that object.
- In the formula editor that pops up, type “*rowproportion*” (without the quotes) to see the row proportions or “*columnproportion*” to see the column proportions. Be sure to spell the names of the formulas correctly or else *Fathom* will give you an error. (If you spell the names correctly, they should change to a purplish color in your editor.)
- You will see that each cell in the **Summary Table** now includes numbers for multiple statistics. To see which numbers correspond with which statistics, simply look at the bottom of your summary table to see the order of the statistics or formulas within each cell.
- To delete (or change) a particular statistic from the table, you can double click on its name at the bottom of the **Summary Table**. In the formula editor, press delete (or make your changes) and then click “**OK**”.



can
students
handle it?

current

intro
stat

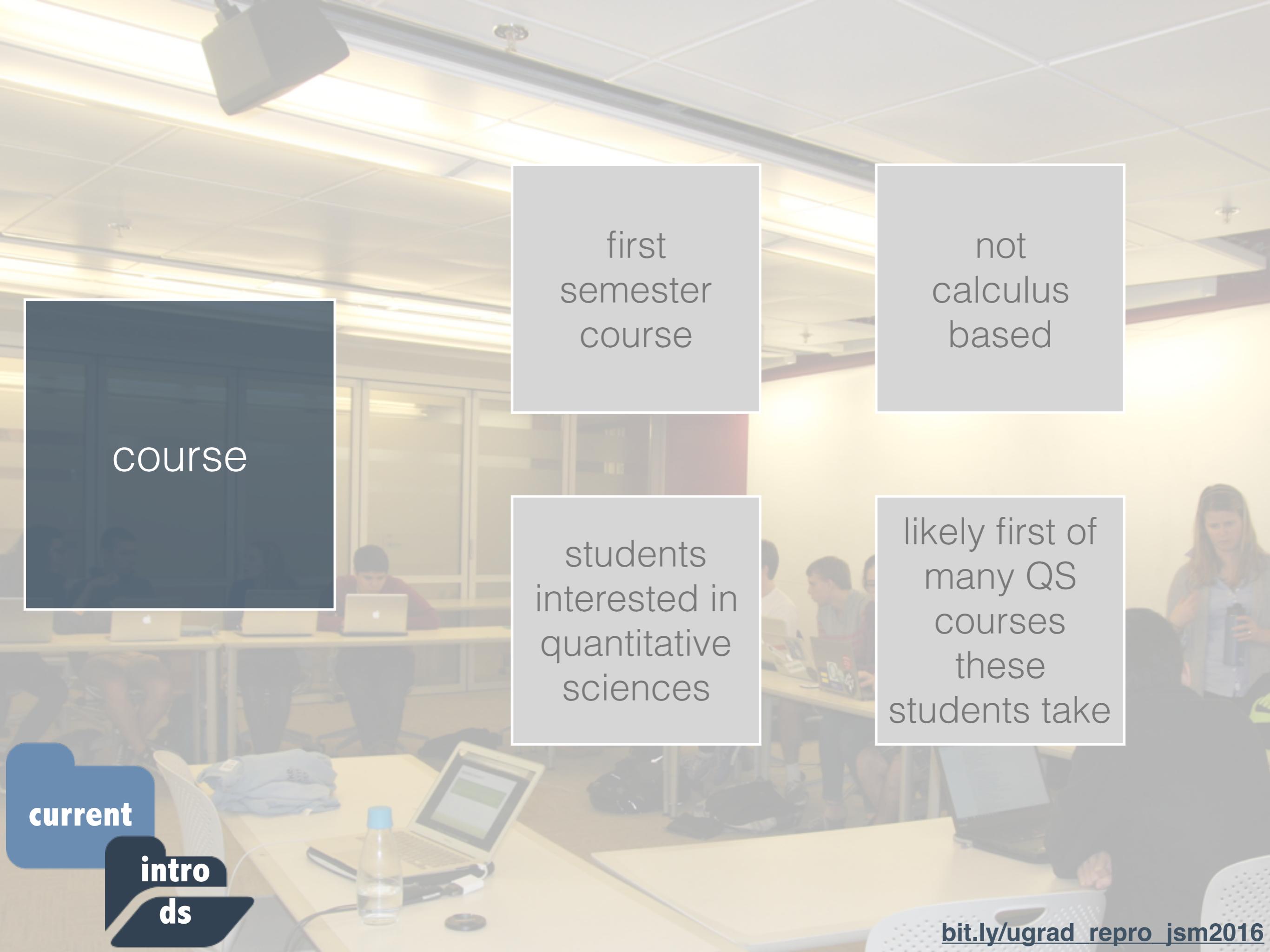
yes!

point-and-
click is no
less
overhead
than scripting

syntax
highlighting

keeps code
organized +
workspace
clean

code
+
output
always
together

The background image shows a classroom environment. Students are seated at wooden desks, working on laptops. The room has a modern feel with large windows and recessed lighting in the ceiling.

course

first
semester
course

not
calculus
based

students
interested in
quantitative
sciences

likely first of
many QS
courses
these
students take

current

intro
ds

bit.ly/ugrad_repro_jsm2016

The background of the slide features a collection of clear plastic petri dishes arranged in a grid pattern. Each dish contains a different type of bacterial culture, showing various growth patterns and colors like white, yellow, and green.

reproducibility

literate
programming

version
control

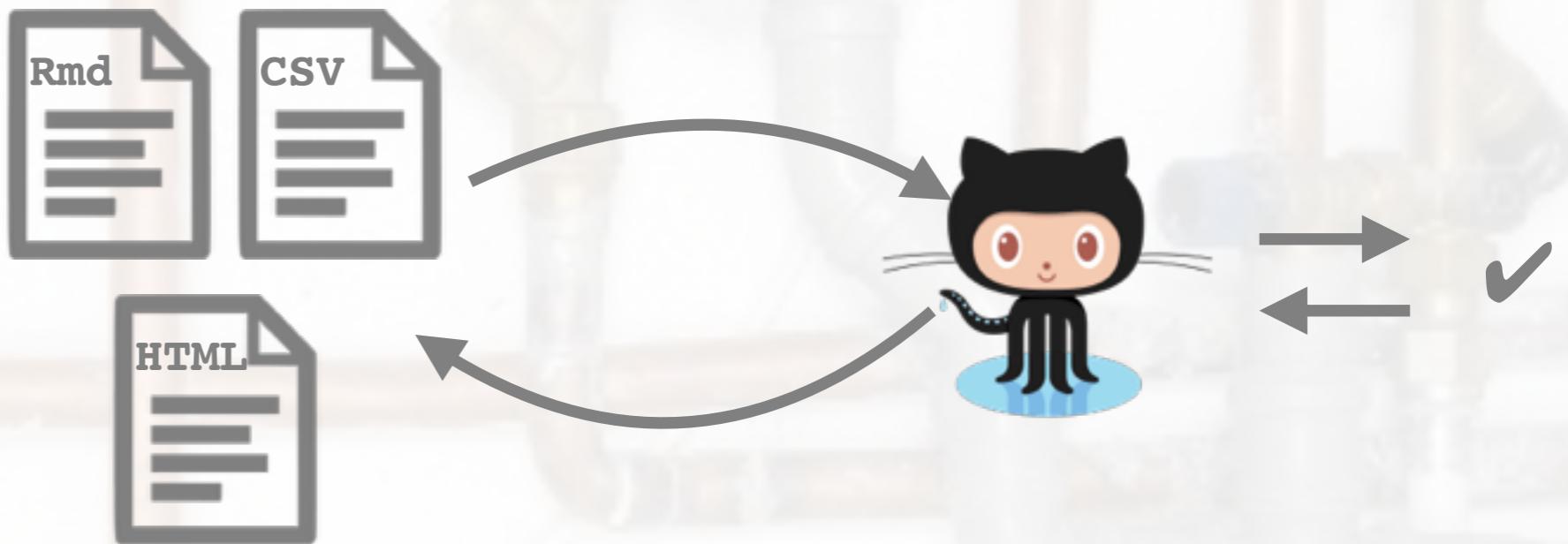
The R logo, consisting of two interlocking 'R's, is positioned next to the word "Studio" in a blue sans-serif font.

R Studio

A blue rounded rectangular icon with a white outline, containing the word "current" in white.currentA dark blue rounded rectangular icon with white wavy lines forming a speech bubble shape, containing the word "intro" in white.introA dark blue rounded rectangular icon with white wavy lines forming a speech bubble shape, containing the letters "ds" in white.ds

[bit.ly/ugrad repro jsm2016](http://bit.ly/ugrad_repro_jsm2016)

workflow



current

intro

ds

bit.ly/ugrad_repro_jsm2016



can
students
handle it?

yes!

but...

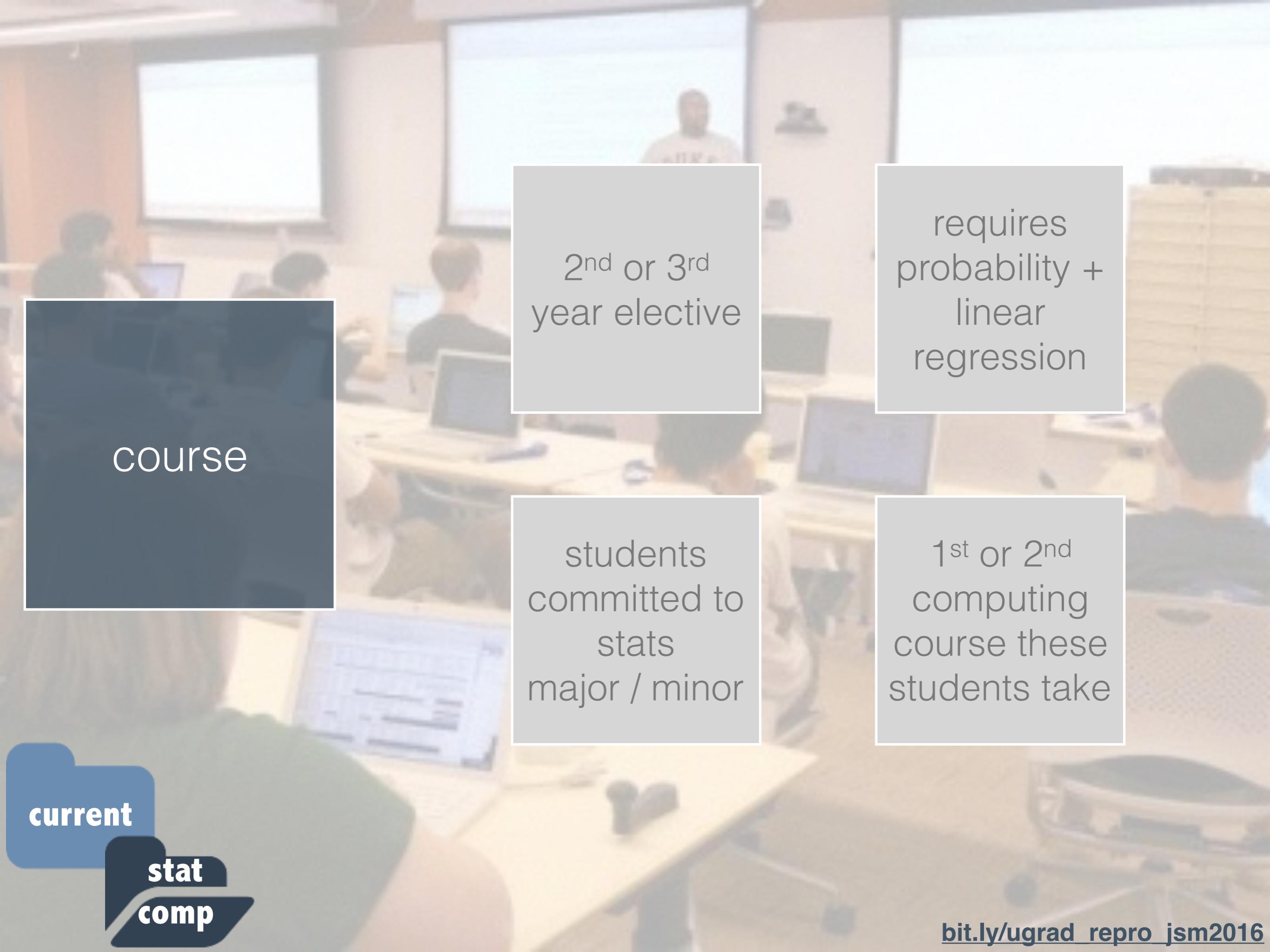
instruction of
workflow
requires time
and care

current

intro

ds

bit.ly/ugrad_repro_jsm2016



course

2nd or 3rd
year elective

requires
probability +
linear
regression

students
committed to
stats
major / minor

1st or 2nd
computing
course these
students take

current

stat

comp



reproducibility

literate
programming

version
control

build
tools

make



R Studio[®]



current

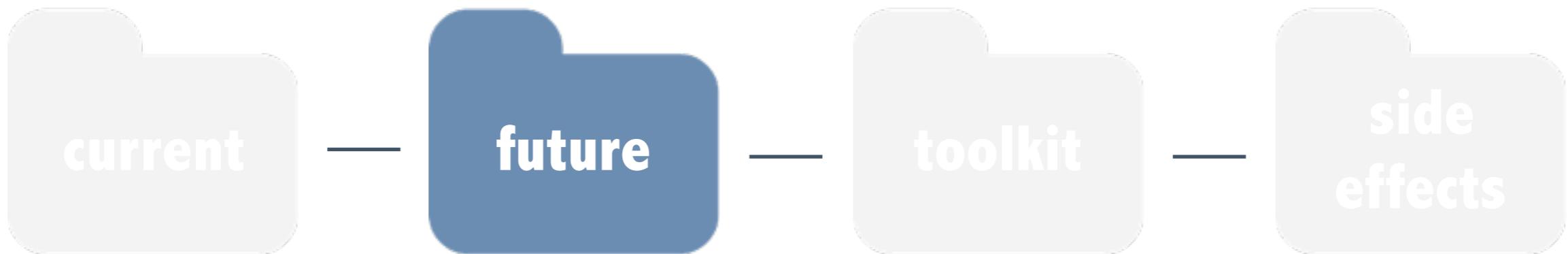
stat

comp

bit.ly/ugrad_repro_jsm2016

details at http://bit.ly/statcomp_jsm2016

grow toolkit along with the complexity of computation



what

capstone
course

senior
thesis /
independent
study

how

need
instructor
buy-in

needs
to be
part of
assessment

easily
adoptable
framework
will help

future

Karl's
steps
2RR

Project
TIER

Reproducible
Science
Curriculum



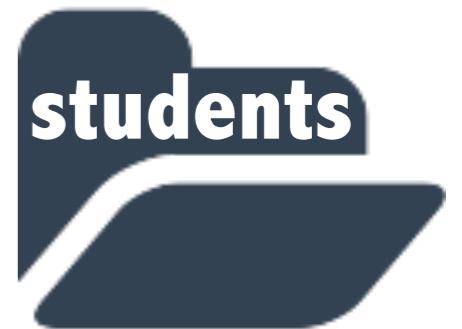
R

other

built-in
seamless
ecosystem
with RStudio

any
scripting
language

more
overhead in
some than
others



for instructors

easy
Q&A

easy
grading

for students

easy
collaboration

self-
promotion

side
effects

bit.ly/ugrad_repro_jsm2016

thank you!

questions?



minebocek



mine@stat.duke.edu



mine-cetinkaya-rundel

bit.ly/ugrad_repro_jsm2016