

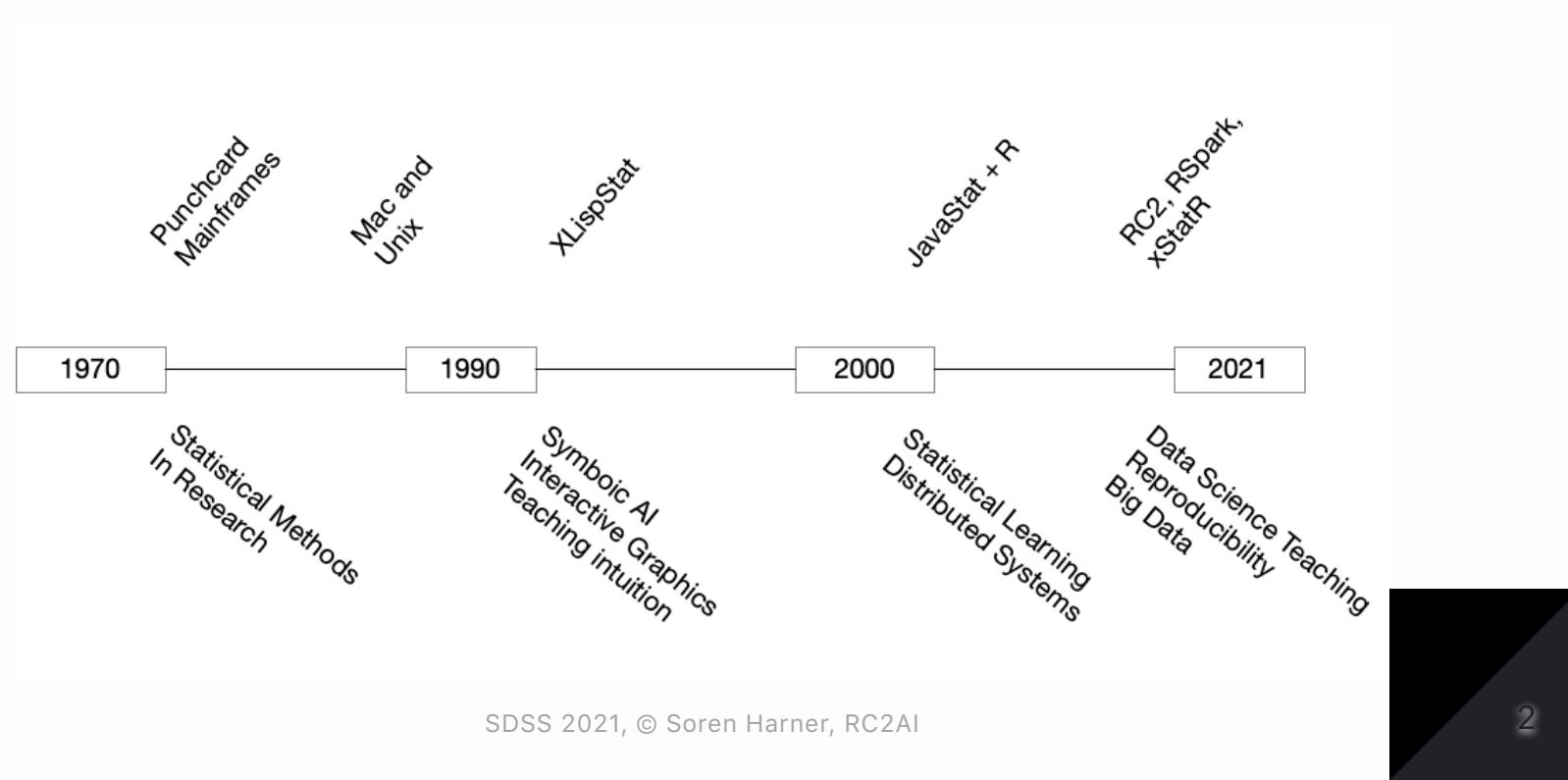
# **How Usability and Reproducibility in Software Improves Teaching and Research**

**Soren Harner, RC2AI**

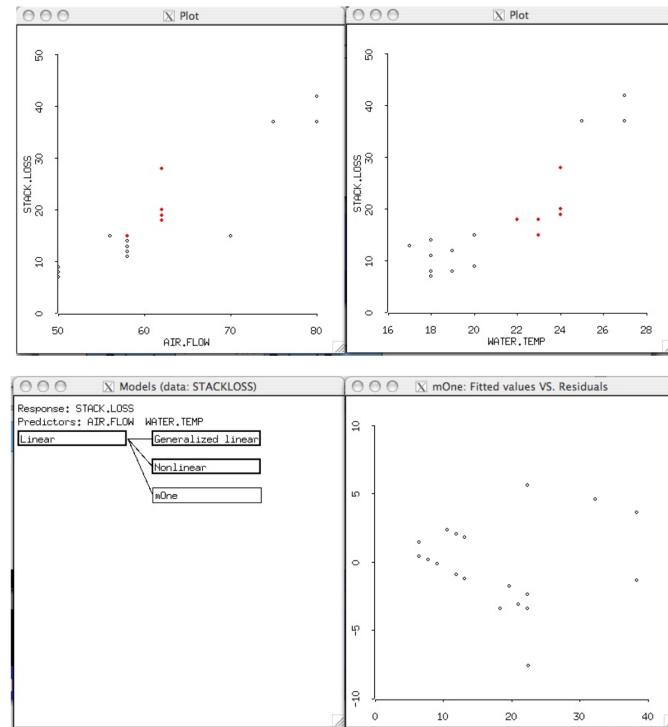
SDSS 2021, © Soren Harner, RC2AI

1

# Jim's 50 years in statistical computing



# 1990s Interactive Graphics in XLispStat



SDSS 2021, © Soren Harner, RC2AI



# 2000s JavaStat: GUI with R Backend

JavaStat: a Java/R-based statistical computing environment

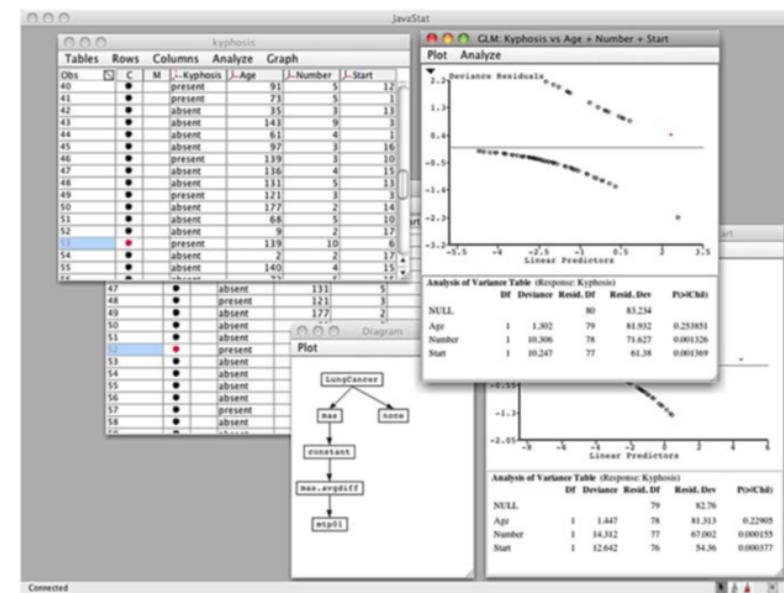


Fig. 3 JavaStat Java application

SDSS 2021, © Soren Harner, RC2AI

4

# After 2015: RSpark and Reproducibility

- Teach data science courses and seminars
- Introduce SQL, DataFrames, MapReduce, Streaming
- Emphasis on reproducibility with Docker and Git
- Built on R, Spark, Apache Arrow, and Postgres
- Building on [rocker](#), R on Docker
- Brought back XLispStat with [xStartR](#)

SDSS 2021, © Soren Harner, RC2AI

5

# $Rc^2$ Reproducibility for Everyone

The screenshot shows two side-by-side RStudio sessions. The left session is 'Rethinking.Rmd' and the right is 'Rethinking.html'. Both sessions show the same code and output.

```

## Building the Regression Model
````{r}
library(rethinking)
data(Howell1)
d <- Howell1
precis(d)
````



	mean	sd	5.5%	94.5%	histogram
height	138.26	27.60	81.11	165.74	
weight	35.61	14.72	9.36	54.50	
age	29.34	20.75	1.00	66.13	
male	0.47	0.50	0.00	1.00	



4 rows



Plot the priors:



```

````{r}
curve(dnorm( x , 178 , 20 ) , from=100 , to=250 )
````



```



The right session ('Rethinking.html') shows the same results, indicating reproducibility.


```

SDSS 2021, © Soren Harner, RC2AI





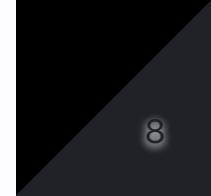
SDSS 2021, © Søren Harner, RC2AI

7

# Reproducibility

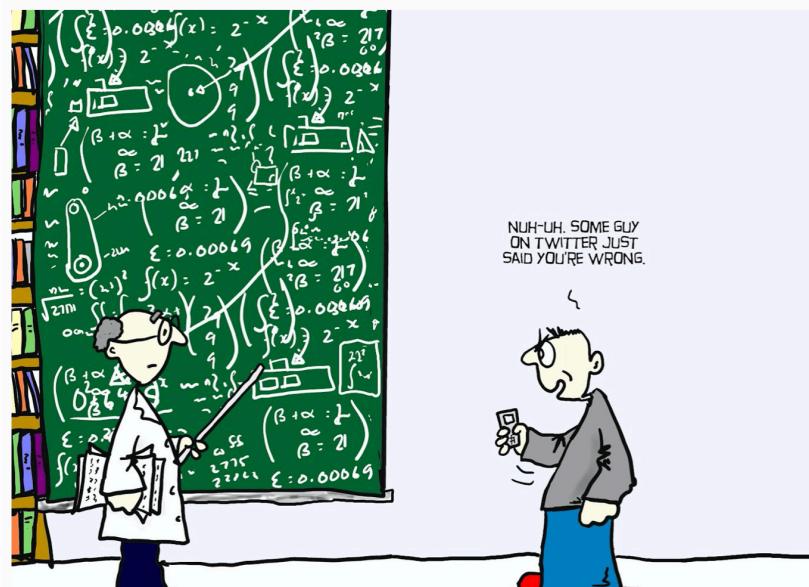
Given the same raw data, can you follow the steps and understand the assumptions of how the authors arrived at their conclusion?

SDSS 2021, © Soren Harner, RC2AI

8

# Why does it matter?

- Building on others' work
- Transparency over authority



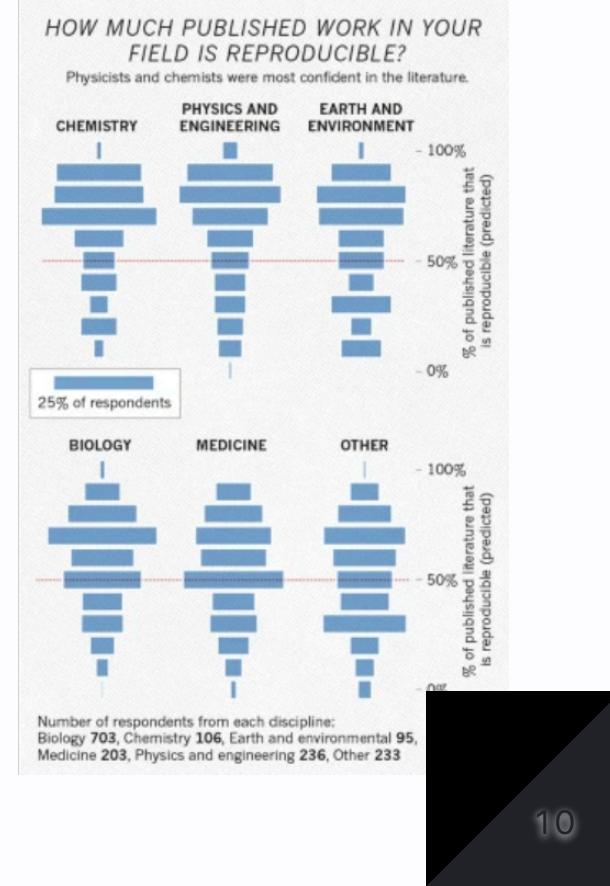
Acceptable use from Clipart Library. 2017.

SDSS 2021, © Soren Harner, RC2AI

9

# Reproducibility in Research

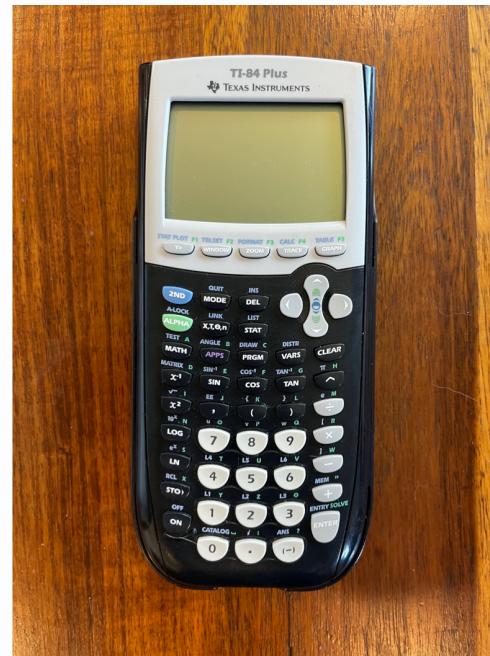
- Nature 2016 Survey 52% say there is a crisis
- Nature 2021 Survey confidence in science leads to vaccination



SDSS 2021, © Soren Harner, RC2AI

# Teach reproducibility

- Data used in all fields
- Computational sciences
- Larger models and big data



In 2021, really?



SDSS 2021, © Soren Harner, RC2AI

# Why is it hard?

- Methods and habits
- Accessible tools and data
- Version everything
- Sharing and verify identity

Zooter.xlsx  
Operations Analytics MOOC

Maximize $150R + 160N$ subject to $4R + 5N \leq 5610$ (frame manufacturing hours) $1.5R + 2.0N \leq 2200$ (wheel and deck manufacturing hours) $1.0R + 0.8N \leq 1200$ (QA and packaging hours) $R, N = \text{integer}$ $R, N \geq 0$			
Profit Contribution (\$/unit)	Razor	Navajo	Total Profit (\$)
	150	160	198000 [1]
Units to Make	840	450	
Resource requirements			
Frame Manufacturing	4	5 [E14]	5610 [2]
Wheels and Deck Assembly	1.5	2	2160
QA and Packaging	1	0.8	1200
			=<=
			5610
			=<=
			2200
			=<=
			1200

[1] =SUMPRODUCT(C9:D9,C10:D10)  
[2] =SUMPRODUCT(\$C\$10:\$D\$10,C14:D14)

Excel fails on reproducibility



# Making reproducible methods more teachable

## Continuing Jim's Work

SDSS 2021, © Soren Harner, RC2AI

13

# Author reproducibly

- VS Code Extension
  - Versioned markdown
  - Remotely executed code
  - Github, Bitbucket integration
  - R, Python, Julia, etc.
  - Javascript, CSS, vegalite
  - Pull request to publish

The screenshot shows an RStudio interface with the following details:

- Code Editor:** The main pane displays R code for Bayesian updating. It includes sections for generating a grid, calculating posteriors, and drawing samples from the posterior distribution.
- Plot:** A plot titled "posterior" shows the probability density of  $\theta_{MP}$ . The x-axis ranges from 0.0 to 1.0 with ticks at 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0. The y-axis ranges from 0.0000 to 0.0500 with ticks every 0.0025. The curve is bell-shaped, centered around 0.5.
- Text Area:** Below the plot, the text "3M2 Draw 10,000 samples from the grid approximation from above. Then use the samples to calculate the 90% HPDI for p." is displayed.
- Console:** The bottom pane shows the R console output, which includes the command `set.seed(150)`.



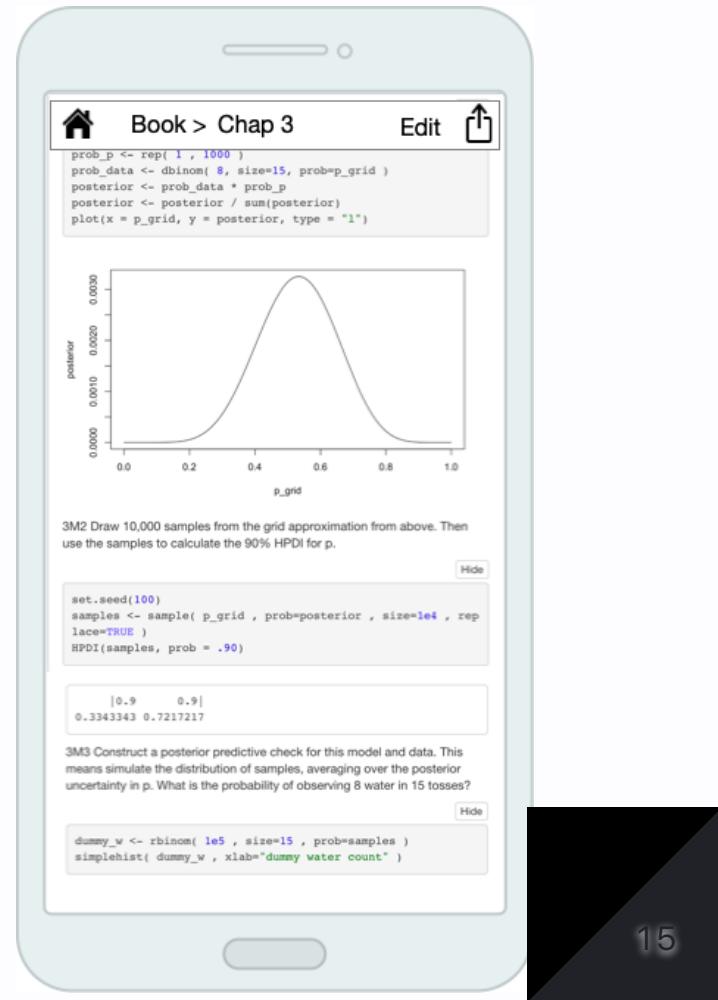
SDSS 2021, © Soren Harner, RC2AI

14

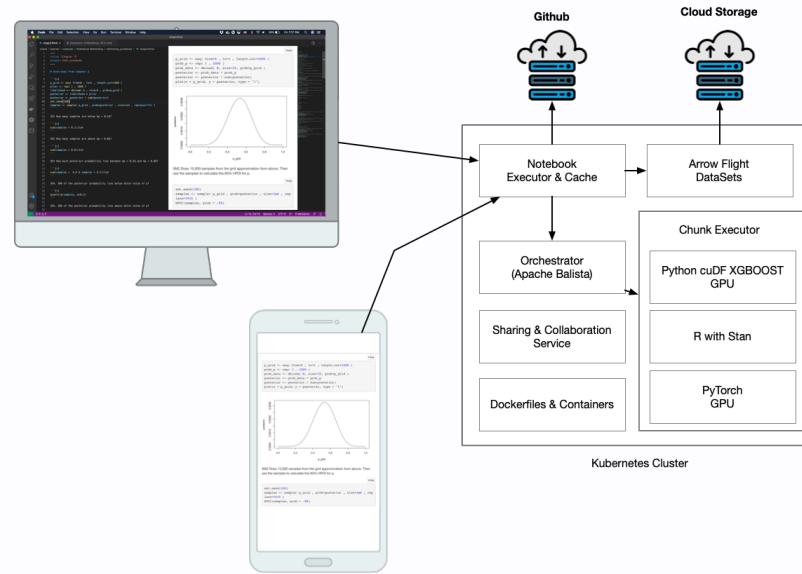
# Read, tinker

- Read, share, re-use interactive notebooks
- Mobile and Web
- Obviate the calculator
- Better than static books and reports
- Verifiable, tamperproof

SDSS 2021, © Soren Harner, RC2AI



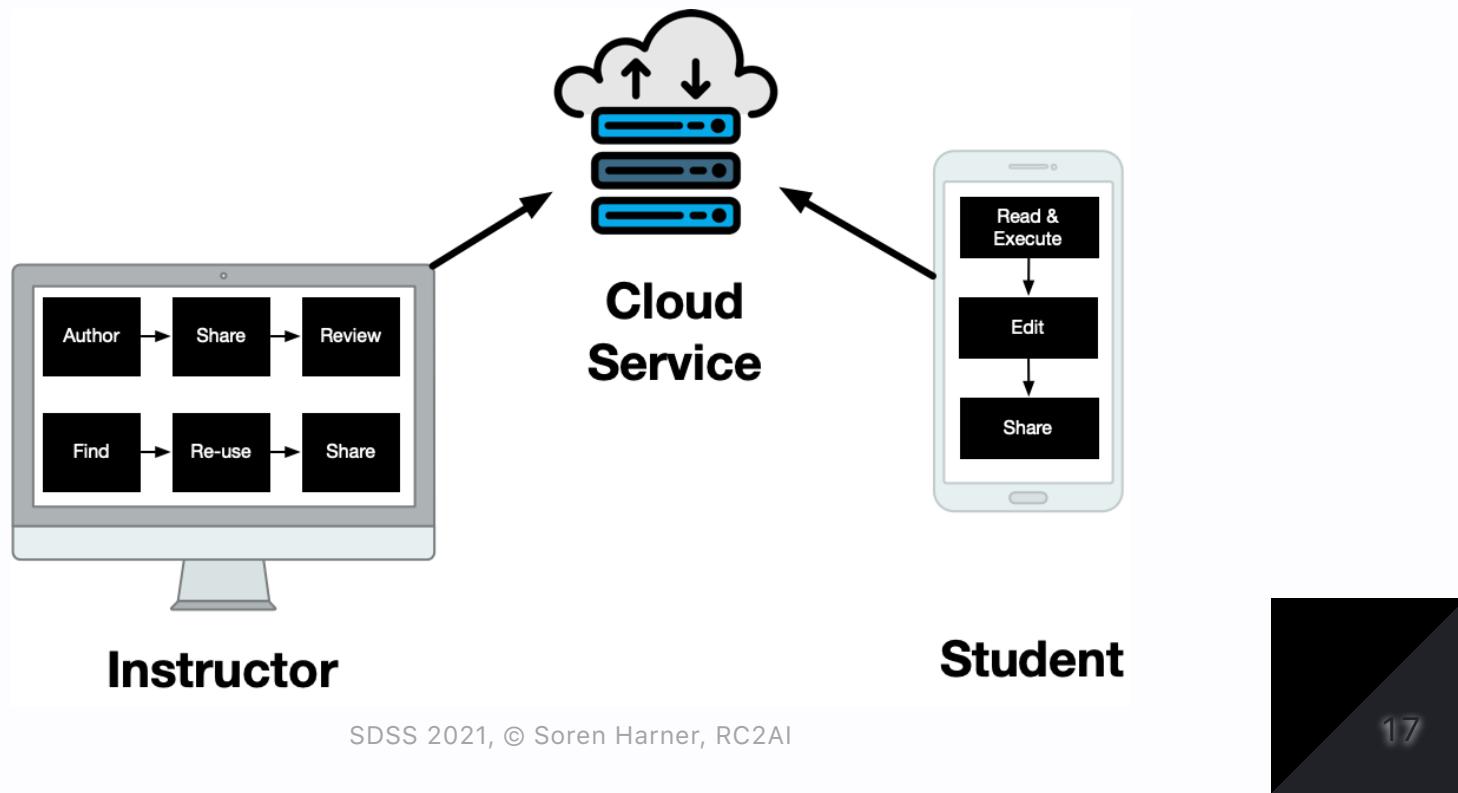
# Accessible, scalable cloud containers



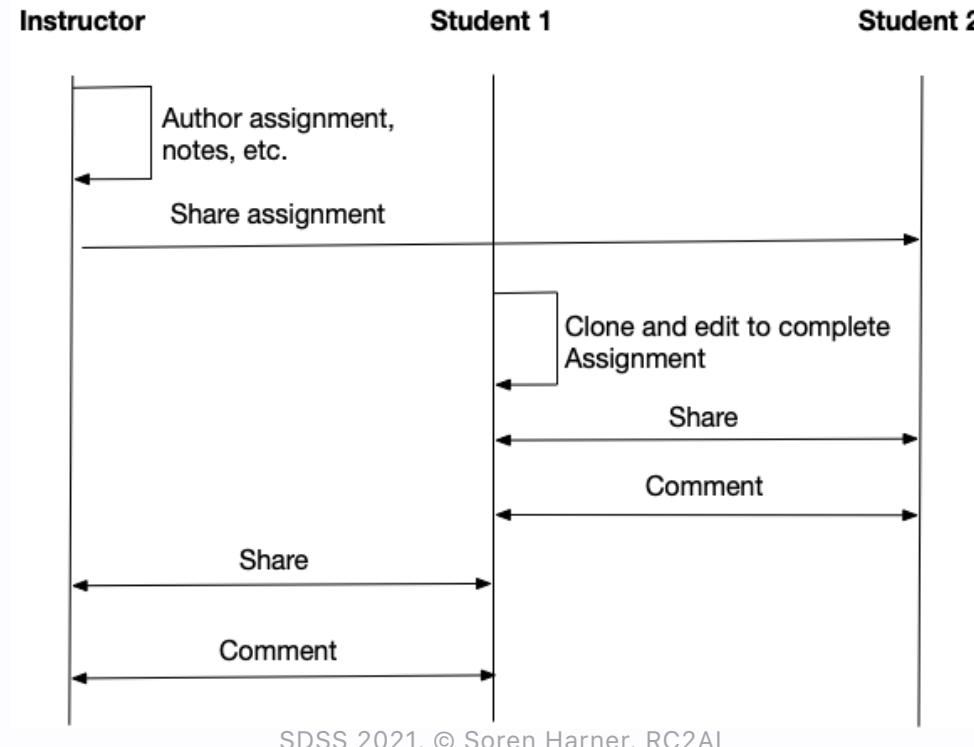
SDSS 2021, © Soren Harner, RC2AI

16

# Integrated collaboration



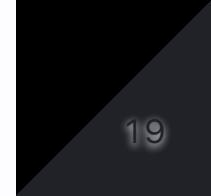
# Use-case: instructor student sharing



# Next steps

- Collect feedback
- Conduct pilots ([sign up](#))
- Continue active development
- Finalize commercial component

SDSS 2021, © Soren Harner, RC2AI

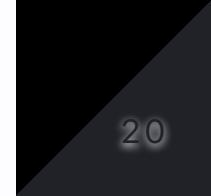


19

# You can get involved

- Join the pilot ([sign up](#))
- Looking for open source collaborators
- Some funding available for developers

SDSS 2021, © Soren Harner, RC2AI



20

# Thank you

[Link to presentation, references, etc.](#)

SDSS 2021, © Soren Harner, RC2AI

