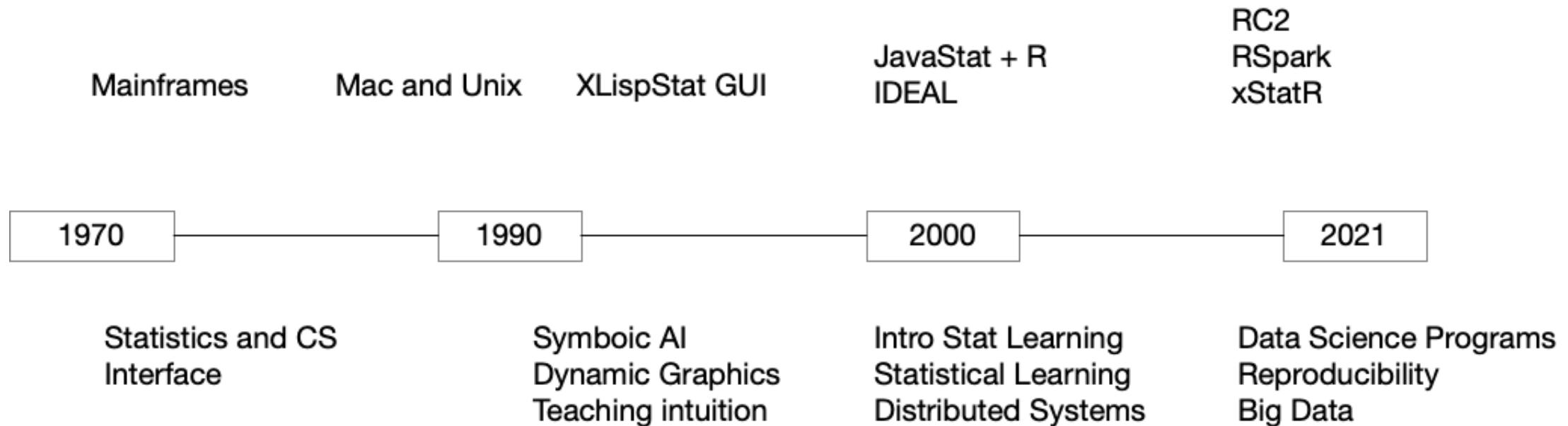


How Usability and Reproducibility in Software Improves Teaching and Research

Soren Harner, RC2AI

Jim's 50 years in statistical computing

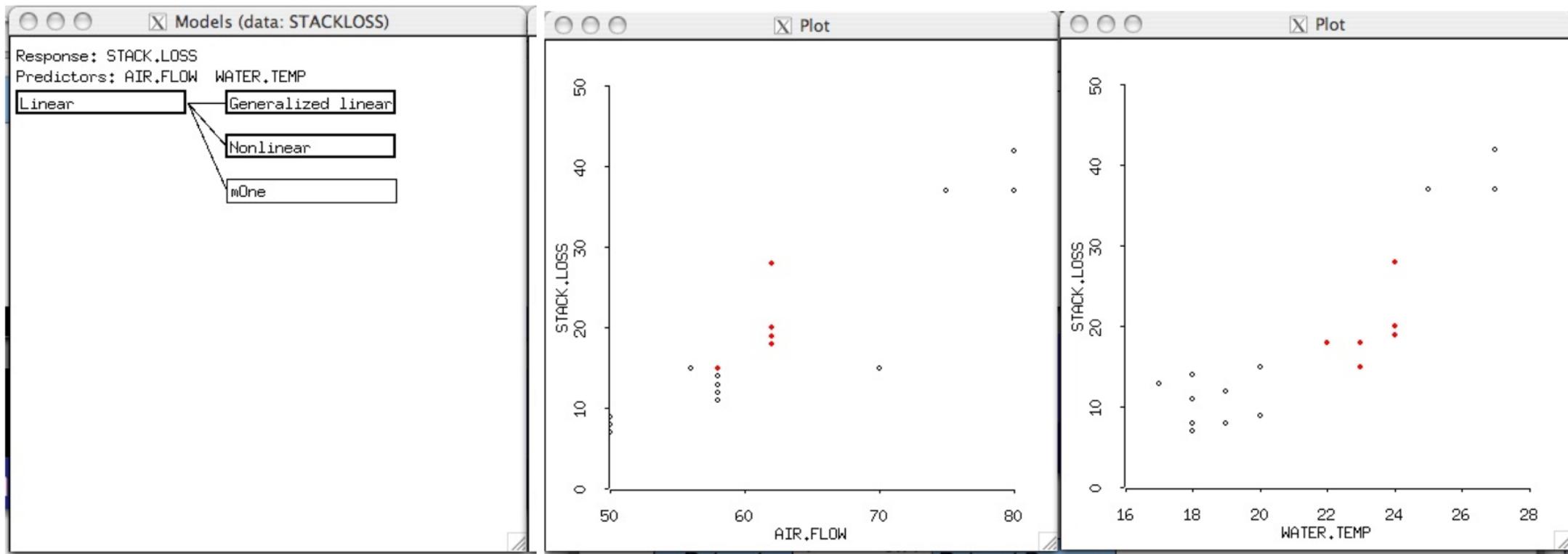


1970s and 1980s Interface of CS and Stat

- Early Interface Symposia
- Punchcards as flashcards
- Homework on line-printer paper
- Memory lane, Mountain View, CA



1990s Interactive Graphics in XLispStat



2000s IDEAL and JavaStat

- IDEAL: On-line, adaptive learning in HTML and Java Applets
- GUI with interactive graphics for building intuition
- Used for 13+ years at WVU
- JavaStat: Client-server architecture with Java and R

JavaStat: a Java/R-based statistical computing environment

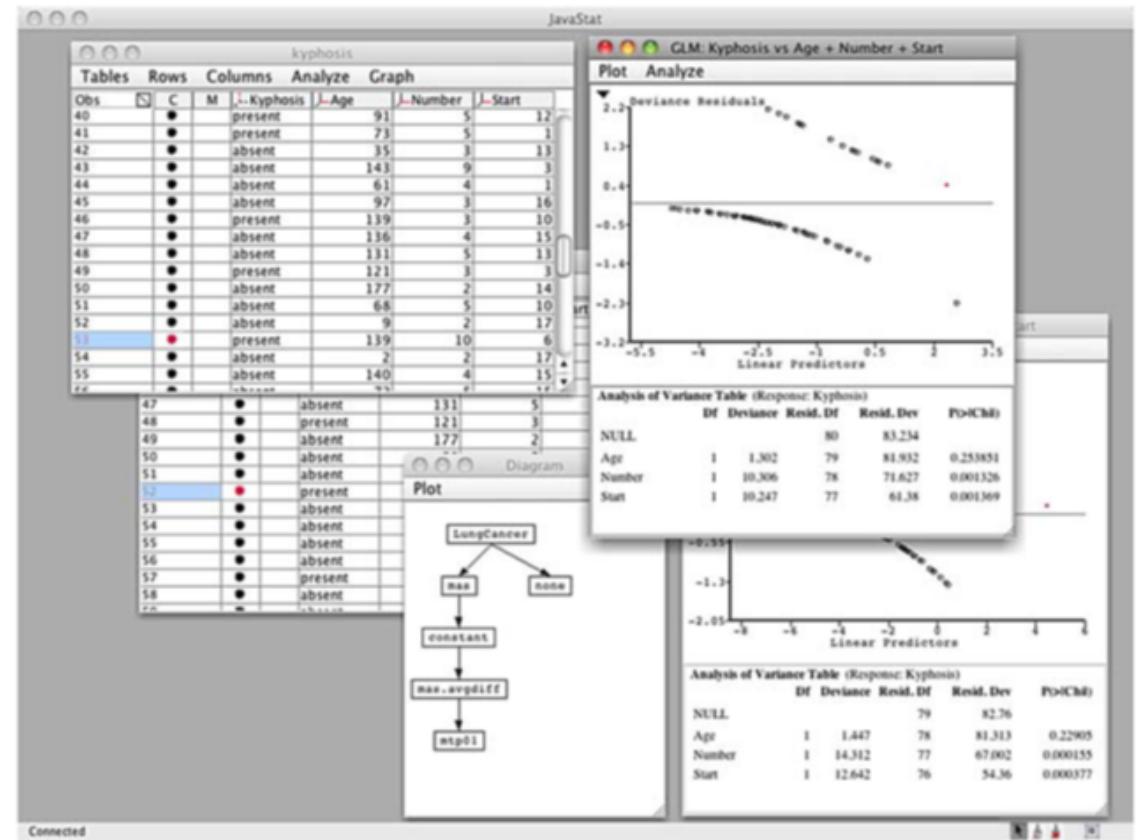


Fig. 3 JavaStat Java application

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 [xStatR](#)

Rc^2 Reproducibility for Teaching

Workspace: Rethinking

Source Files

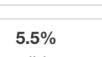
- Rethinking.Rmd
- firstScript.R
- Other
- Rethinking.html

Building the Regression Model

```
```{r}
library(rethinking)
data(Howell1)
d <- Howell1
precis(d)
````
```

R Console

data.frame
4 x 5

| | mean | sd | 5.5% | 94.5% | histogram |
|--------|--------|-------|-------|--------|---|
| | <dbl> | <dbl> | <dbl> | <dbl> | <chr> |
| 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)
````
```

dnorm(x, 178, 20)

0.000 0.010 0.020

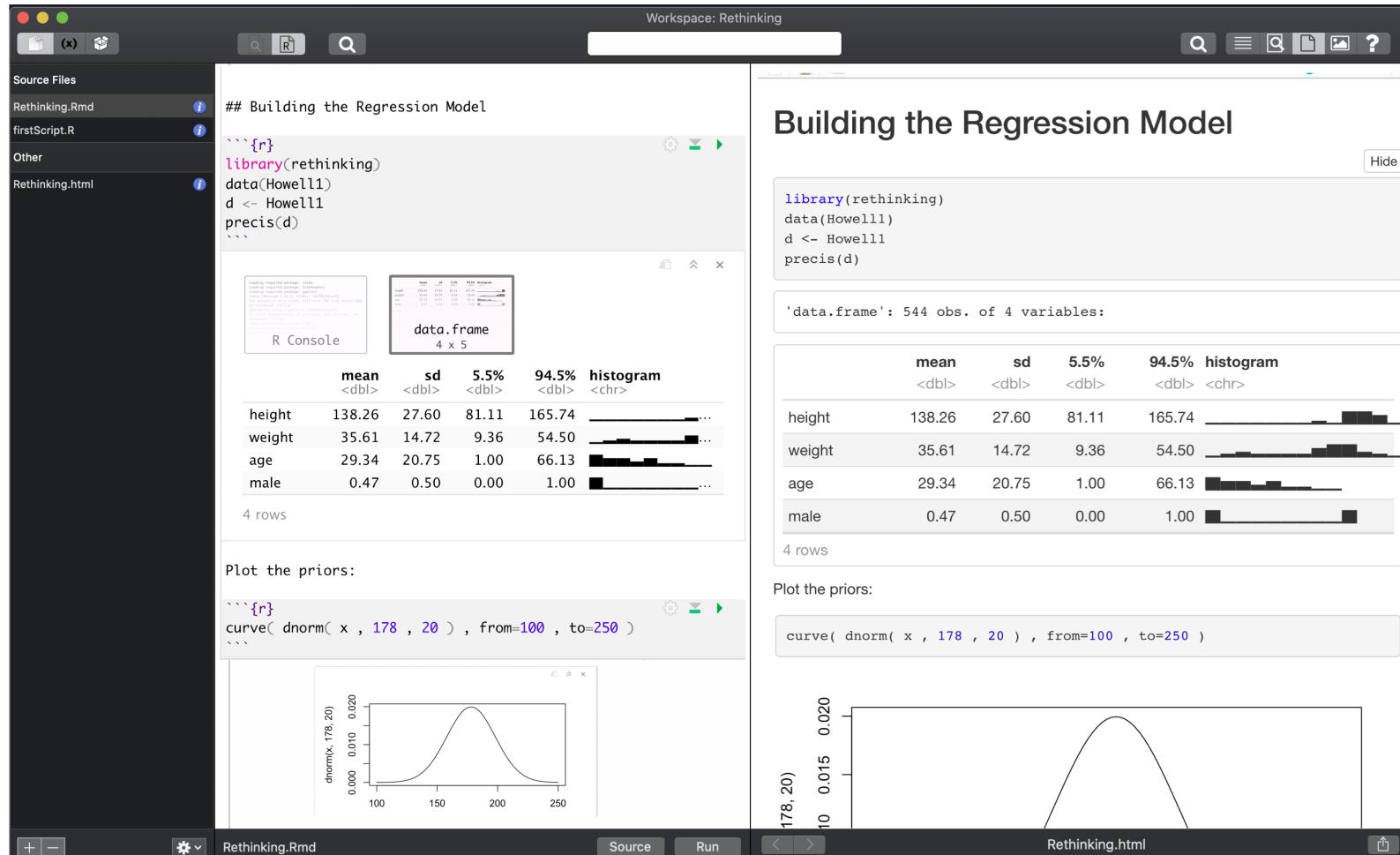
100 150 200 250

178, 20

Rethinking.Rmd

Source Run

Rethinking.html



The image shows a comparison between RStudio (left) and a web browser (right) displaying the same R code and its results. The RStudio interface includes a sidebar with source files, a code editor with R code, and a console window showing histograms for variables height, weight, age, and male. The web browser shows the same R code and the resulting HTML output, which includes the histograms and a normal distribution plot for the prior curve.

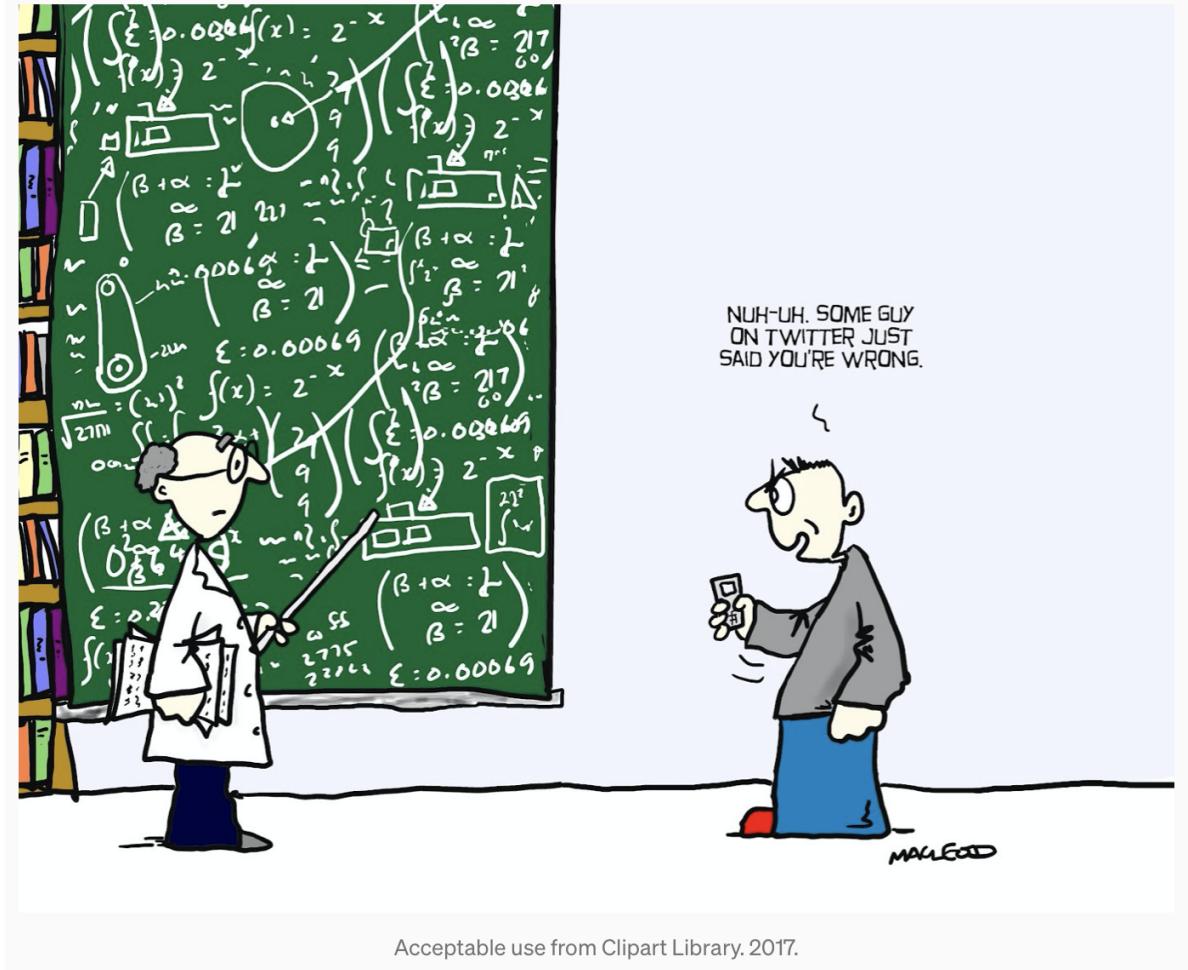


Reproducibility

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

Why does it matter?

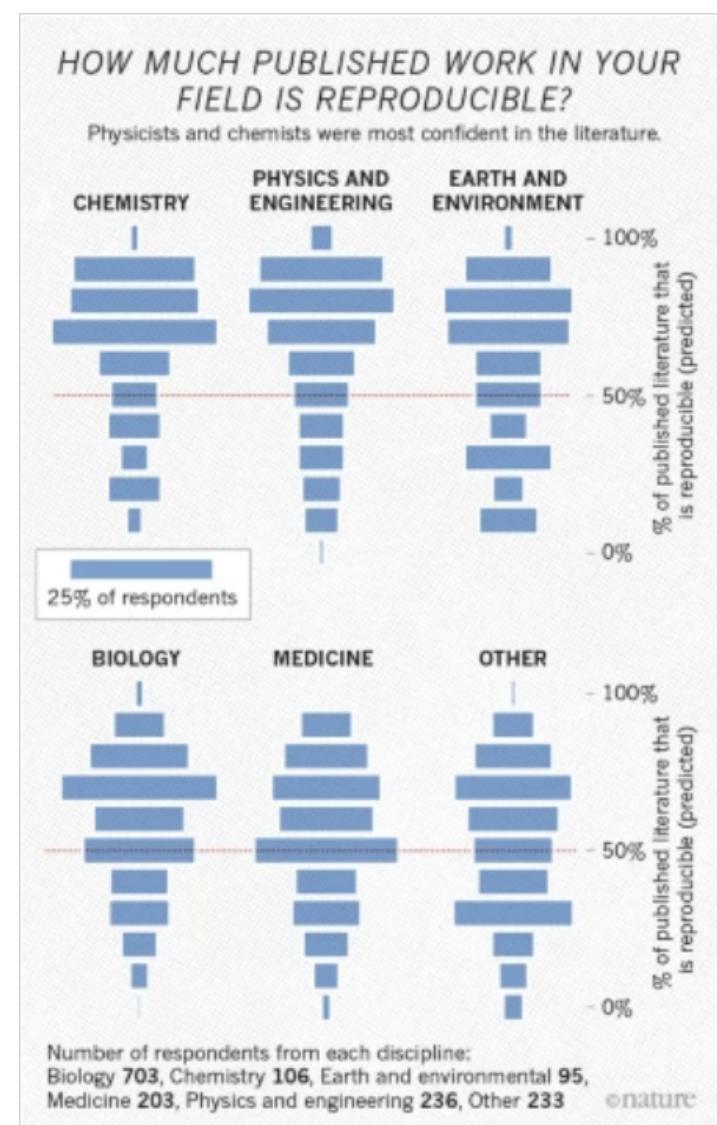
- Building on others' work
- Transparency over authority



Acceptable use from Clipart Library. 2017.

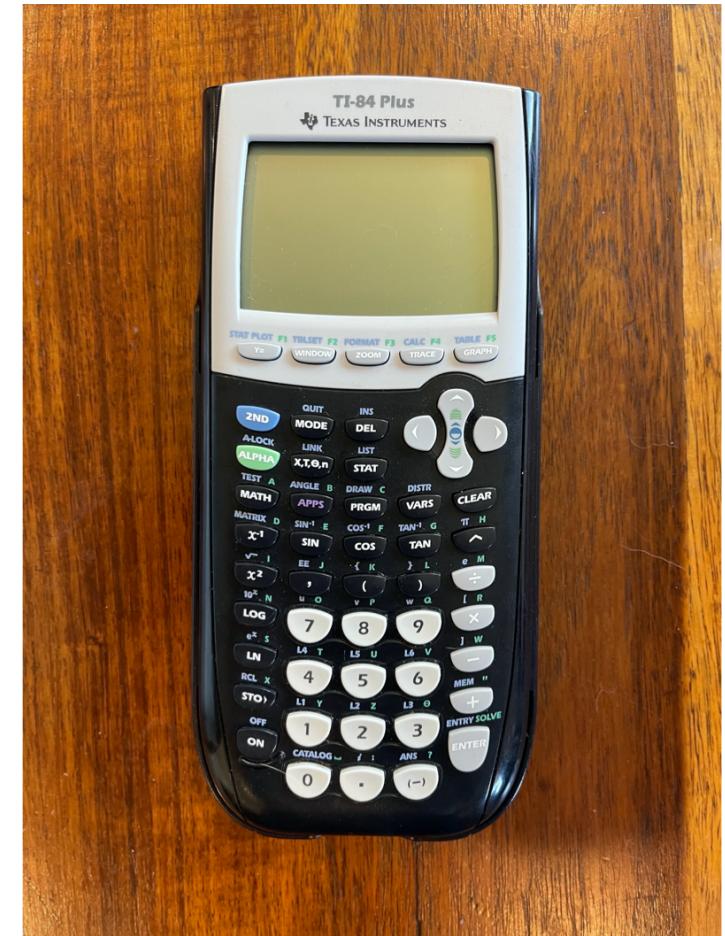
Reproducibility in research

- [Nature 2016 Survey](#) 52% say there is a crisis
- [Nature 2021 Survey](#) confidence in science leads to vaccination



Reproducibility in teaching

- Data plays great role in all fields
- Rise of computational sciences
- Larger models and big data



In 2021, really?

Why is it hard?

- Methods and habits
- "It works on my machine"
- Accessible tools and data
- Everything is versioned
- Sharing and verifying identity

| 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) |
| $R, N = \text{integer}$ | | | |
| $R, N \geq 0$ | | | |

| | Razor | Navajo | Total Profit (\$) |
|-------------------------------|-------|--------|-------------------|
| Profit Contribution (\$/unit) | 150 | 160 | |
| Units to Make | 840 | 450 | 198000[1] |

| | Razor | Navajo | Required (hours) | | Available (hours) |
|--------------------------|-------|--------|------------------|--------|-------------------|
| Frame Manufacturing | 4 | 5 | E14 [2] | \leq | 5610 |
| Wheels and Deck Assembly | 1.5 | 2 | 2160 | \leq | 2200 |
| QA and Packaging | 1 | 0.8 | 1200 | \leq | 1200 |

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

Excel fails on reproducibility

Continuing Jim's Work

**Reproducibility helps teaching
through sharing while it engenders
skills and habits**



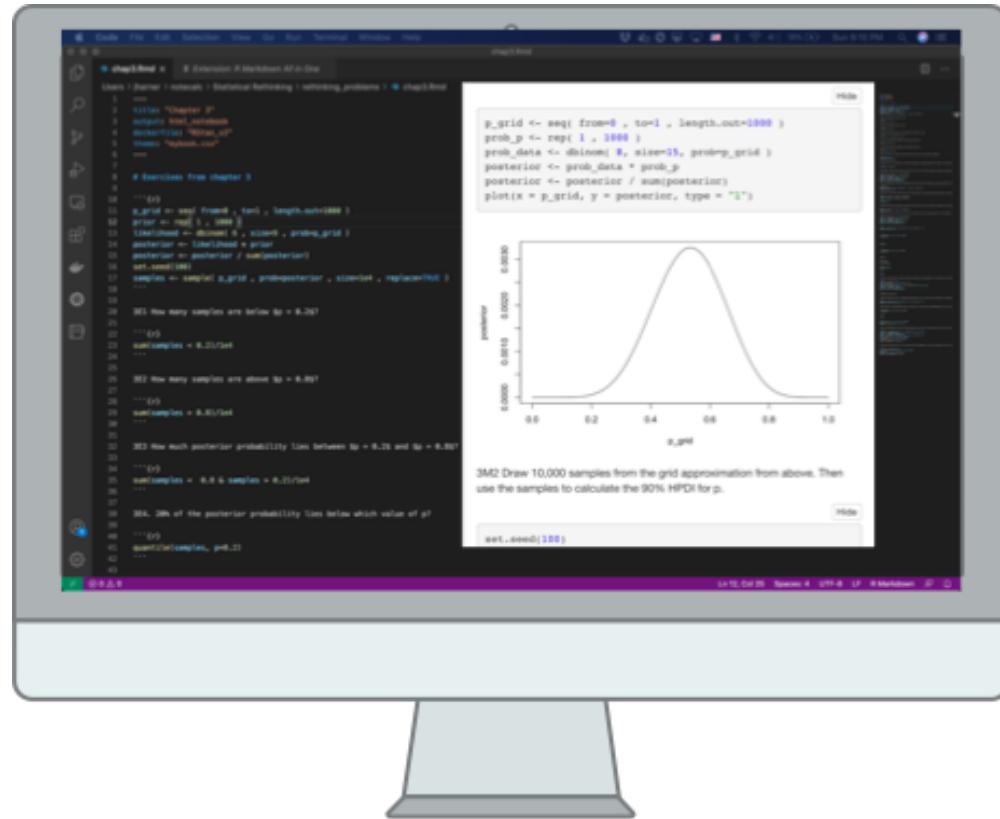
JIM HARNER



SOREN HARNER

Author reproducibly

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



The screenshot shows the R Markdown All-in-One extension in VS Code. The interface includes a sidebar with file navigation, a main editor area with R code, and a preview pane on the right displaying a plot of a posterior distribution. The code in the editor is as follows:

```
p_grid <- seq(from=0, to=1, length.out=1000)
prob_p <- rep(1, 1000)
prior <- dnorm(0, 1, 1000)
likelihoood <- dnorm(0, 1, 1000 * prior)
posterior <- likelihoood * prior
sumposterior <- sum(posterior)
set.seed(100)
samples <- rnorm(p_grid, sumposterior, 1 / sumposterior)
samples <- samples[samples > 0]
samples <- samples[samples < 1]

## How many samples are below 0.5 = 0.251
## 251 samples < 0.5

## How many samples are above 0.5 = 0.251
## 251 samples > 0.5

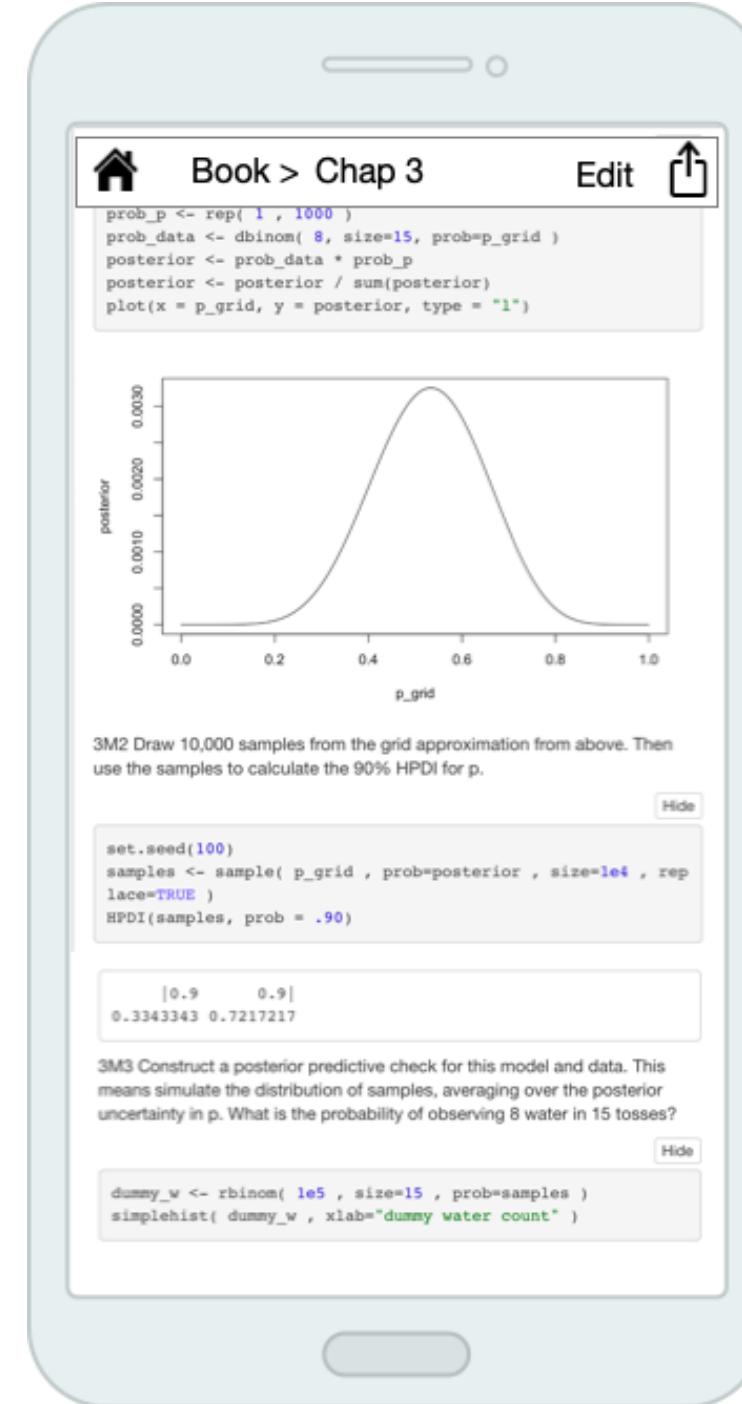
## How much posterior probability lies between 0.5 - 0.51 and 0.5 + 0.51
## 0.02
## 251 samples < 0.51 & samples > 0.49

## 20% of the posterior probability lies below which value of p?
## 0.25
## set.seed(100)
```

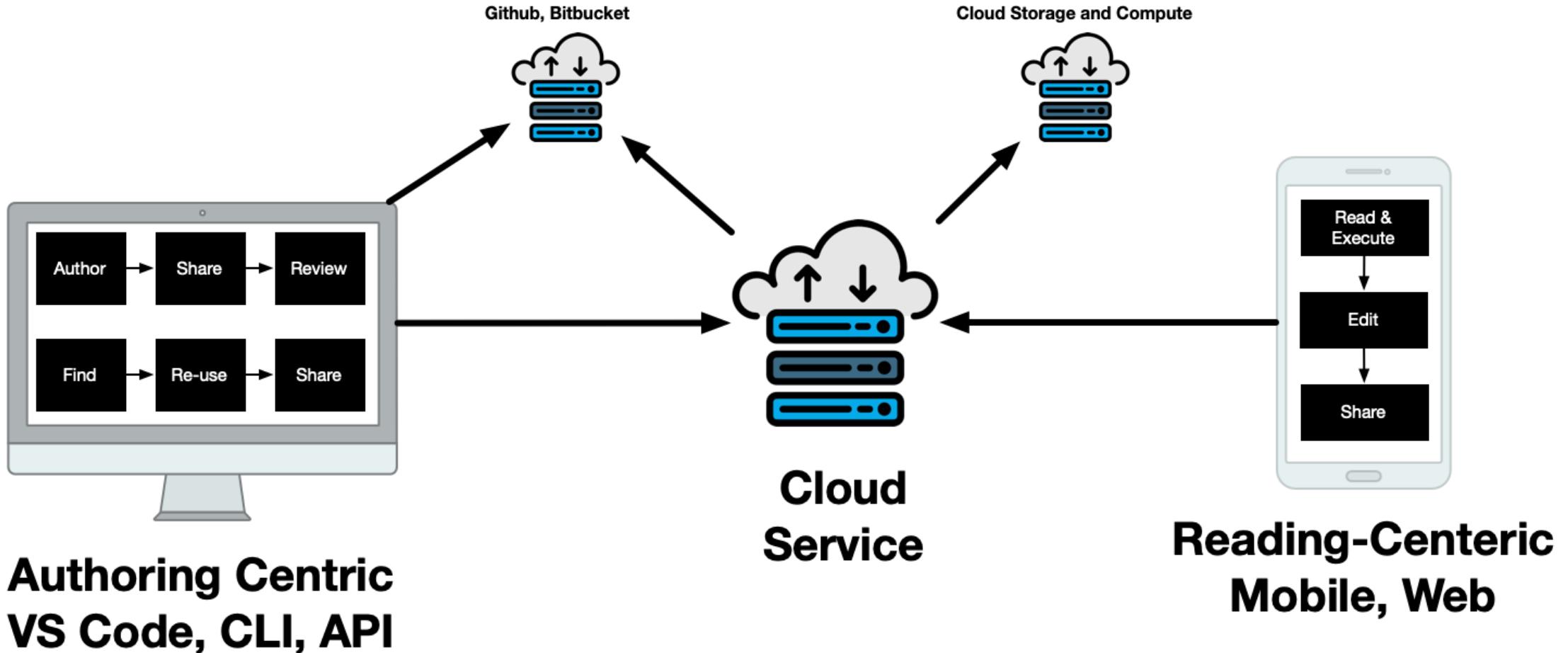
The preview pane shows a bell-shaped curve representing the posterior distribution of p , centered around 0.5. The x-axis is labeled "x_p" and ranges from 0.0 to 1.0. The y-axis is labeled "posterior" and ranges from 0.0000 to 0.2000.

Share, read, tinker

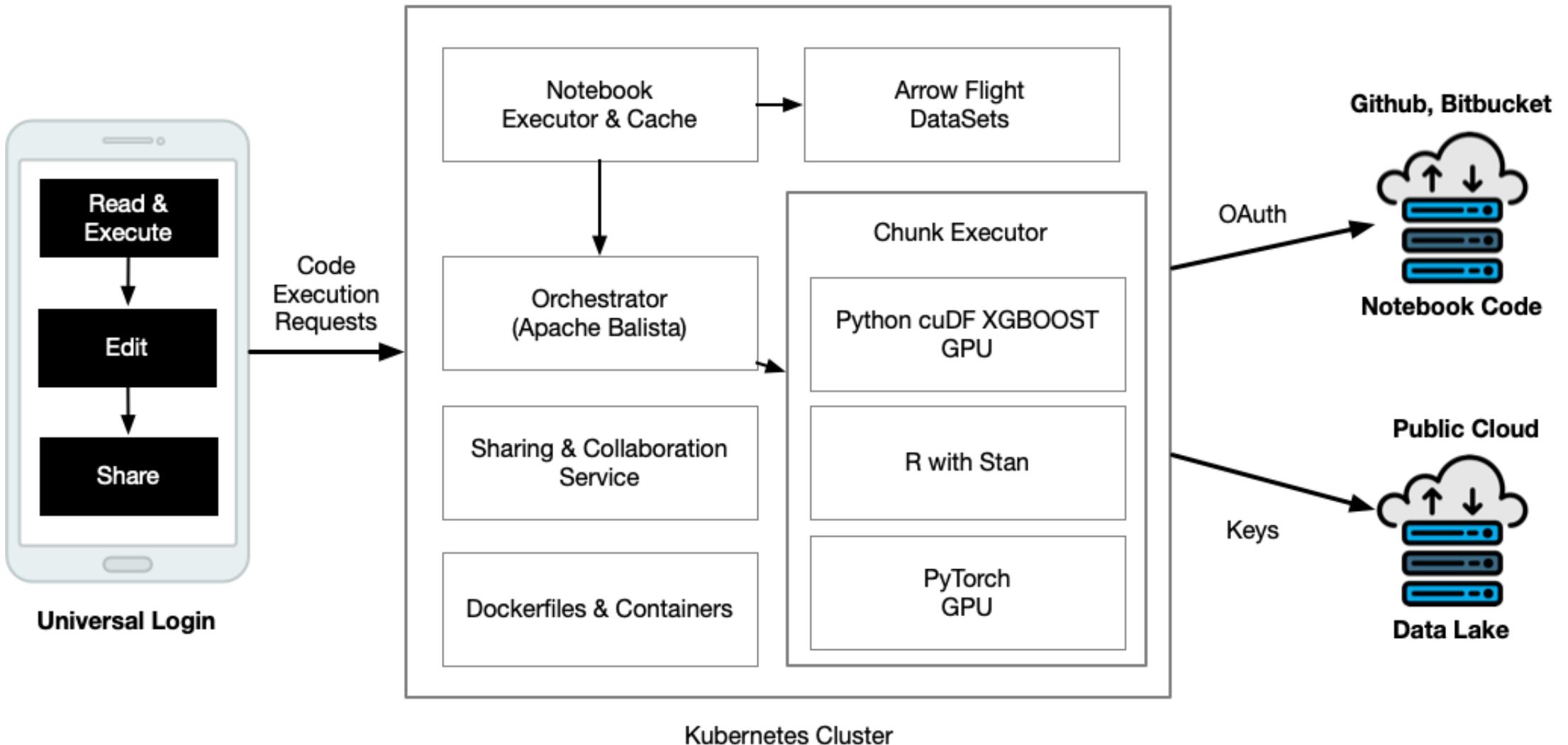
- Read, share, re-use interactive notebooks
- Mobile and Web
- Obviates the calculator
- Better than static books and reports
- Verifiable, tamperproof



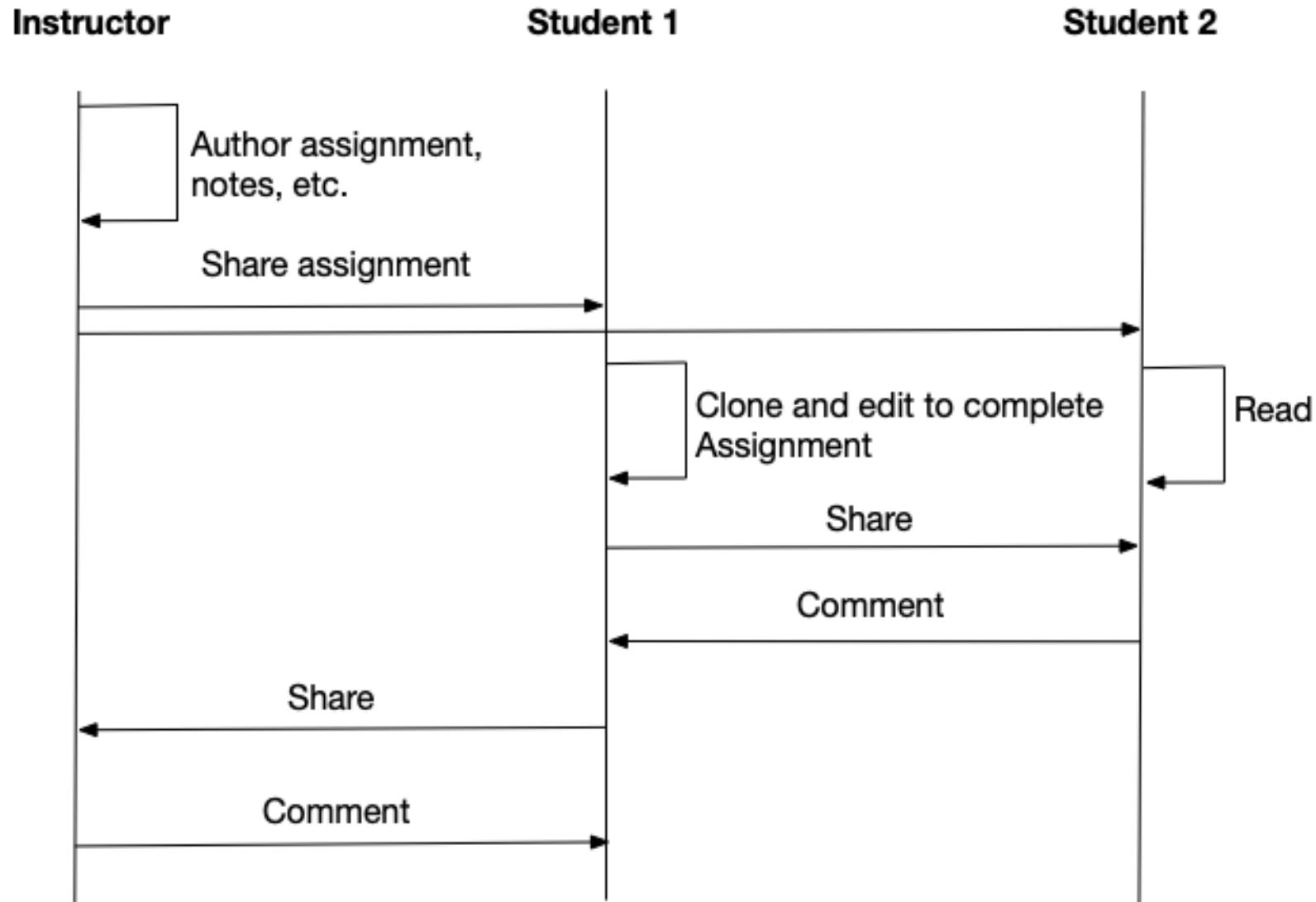
Hosted as a cloud service



Accessible, scalable cloud containers



Use-case: instructor, student sharing



Next steps

- Now: Collect feedback from the community
- Sept 2021: early user testing of Rc^2
- Early 2022: conduct pilots in representative courses
- Develop as Open Source
- Finalize free tier and commercial component

Please get involved with Rc^2

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

Thank you for remembering Jim

[Link to links on Jim's work, tributes, presentation, references, etc.](#)

[Feedback and sign-up](#)



**Links to Presentation,
Sign Up, Survey**