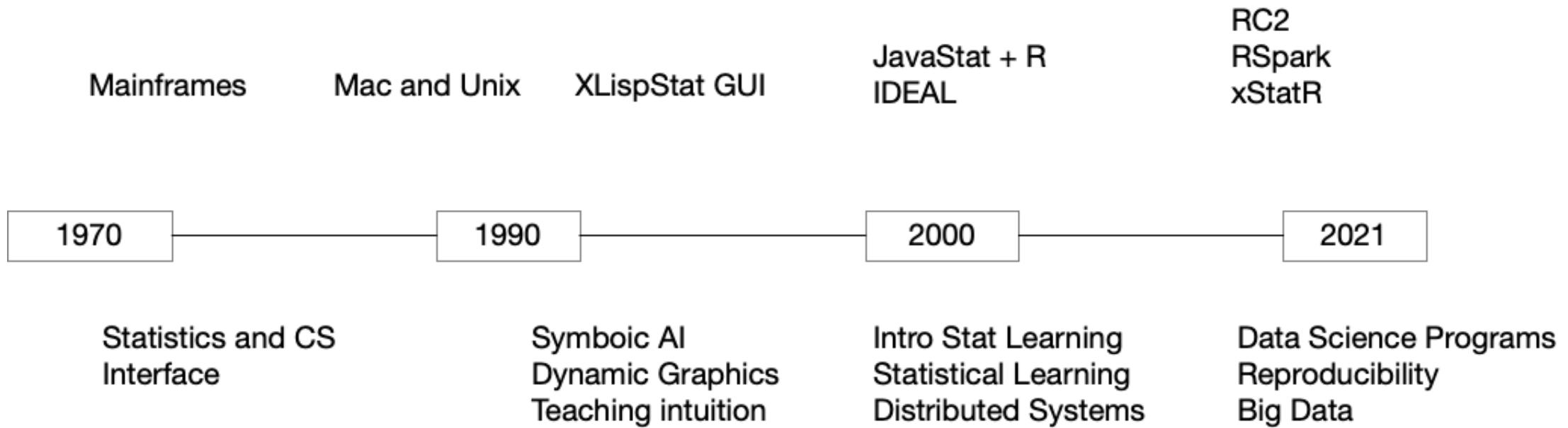


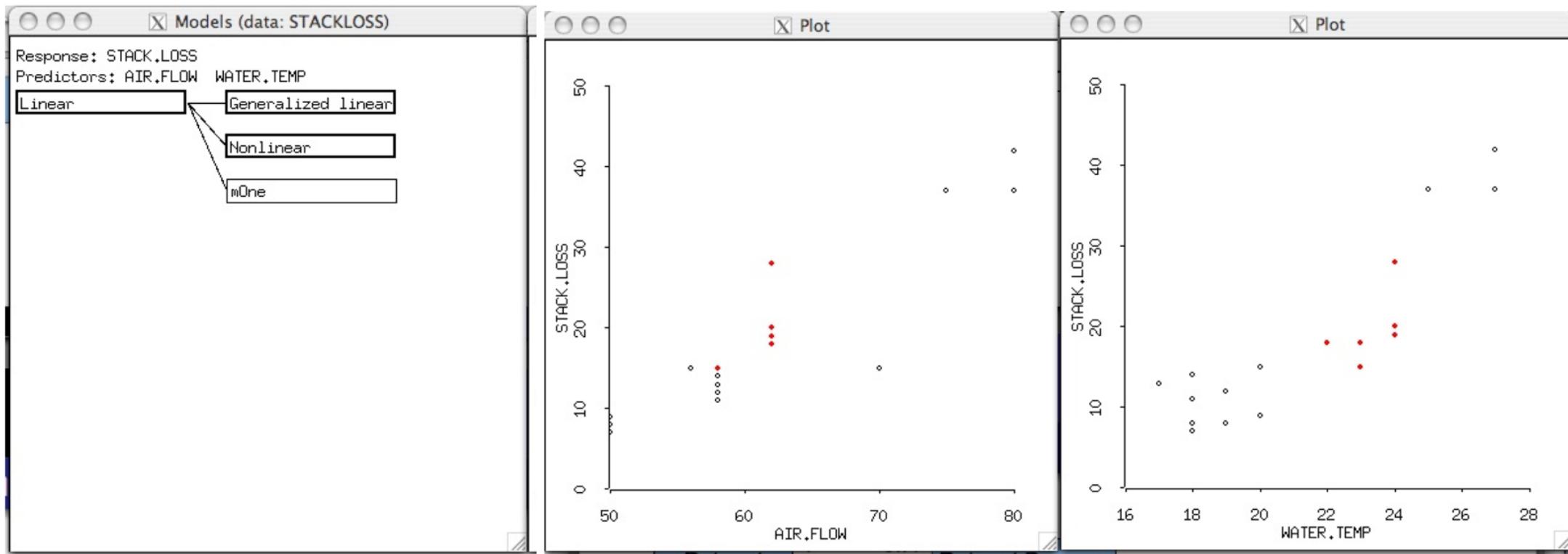
How Usability and Reproducibility in Software Improves Teaching and Research

Soren Harner, RC2AI

Jim's 50 years in statistical computing



1990s Interactive Graphics in XLispStat



2000s IDEAL and JavaStat: GUI with R Backend

JavaStat: a Java/R-based statistical computing environment

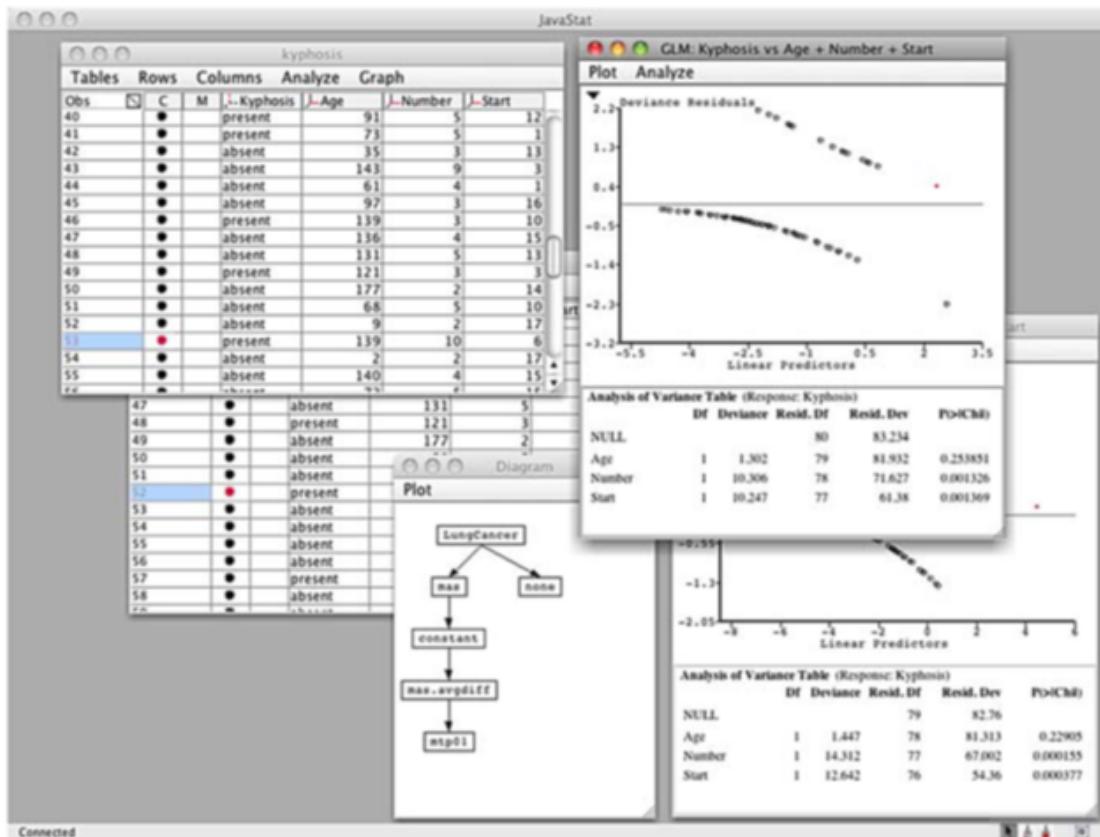


Fig. 3 JavaStat Java application

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 [xStartR](#)

Rc^2 Reproducibility for Everyone

The screenshot shows two side-by-side RStudio environments comparing the output of an R Markdown file (Rethinking.Rmd) and its corresponding HTML output (Rethinking.html).

Left Panel (Rethinking.Rmd):

- Code:**

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

```
- Console Output:** Shows the loading of the rethinking package and the structure of the data frame.
- Data Frame View:** A table showing summary statistics for height, weight, age, and male variables.
- Code:**

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

```
- Plot View:** A normal distribution plot centered at 178 with a standard deviation of 20.

Right Panel (Rethinking.html):

- Title:** Building the Regression Model
- Code:** The same R code as the Rmd file.
- Data Frame View:** A summary table and histograms for height, weight, age, and male.
- Plot View:** The same normal distribution plot as the Rmd file.

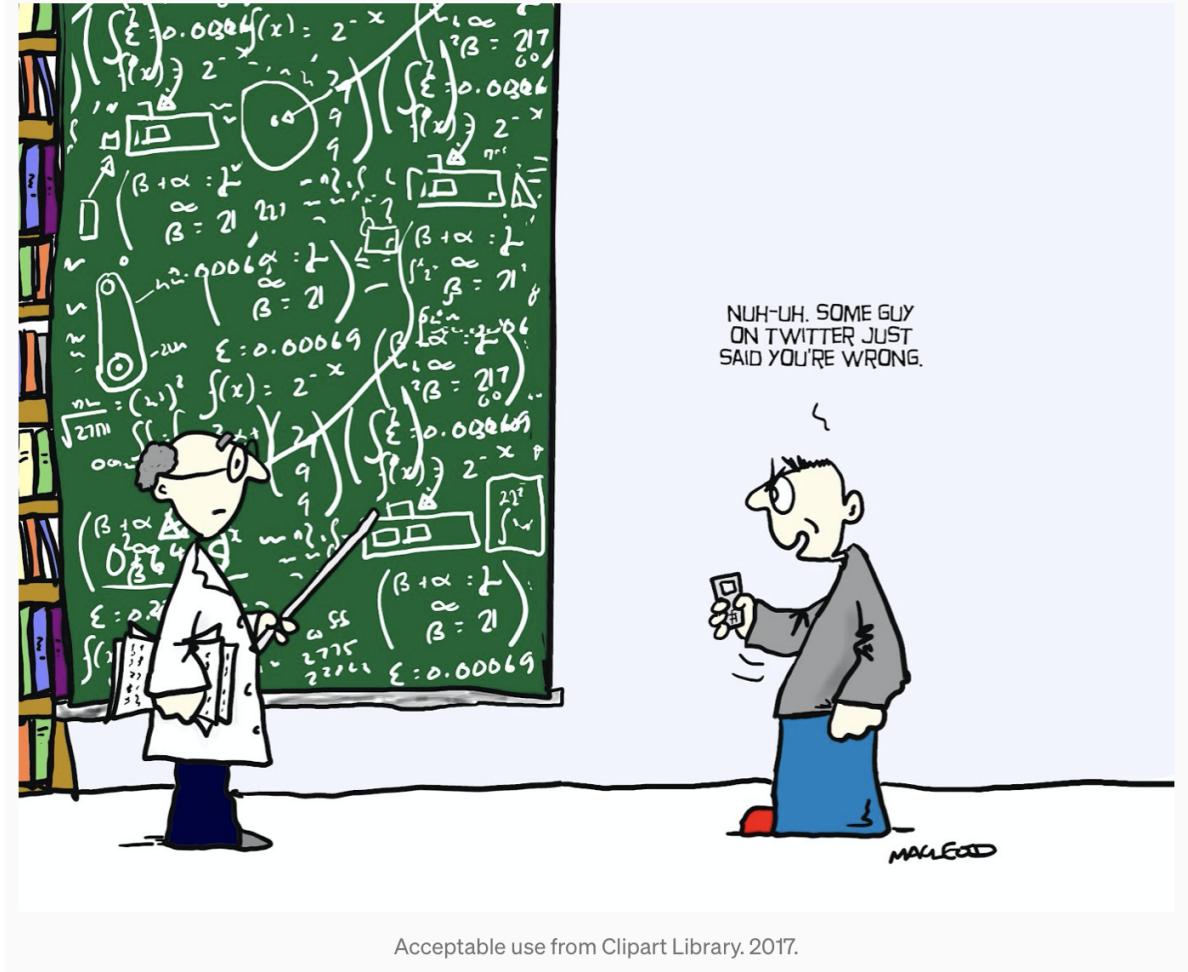


Reproducibility

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?

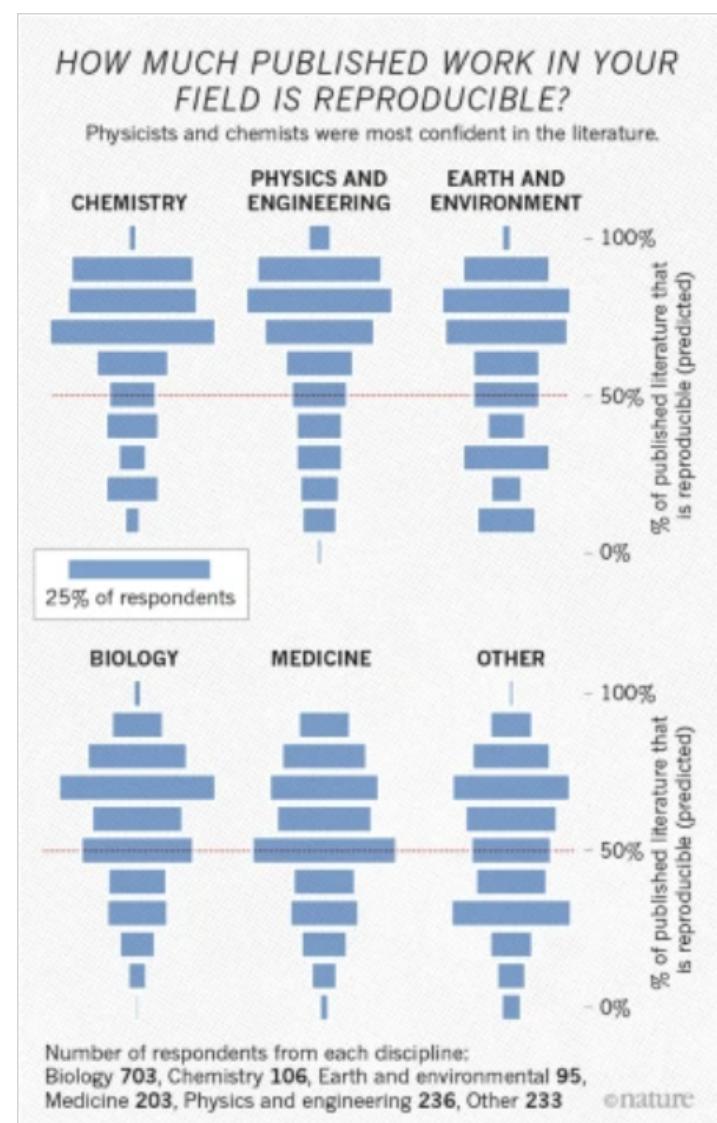
- Building on others' work
- Transparency over authority



Acceptable use from Clipart Library. 2017.

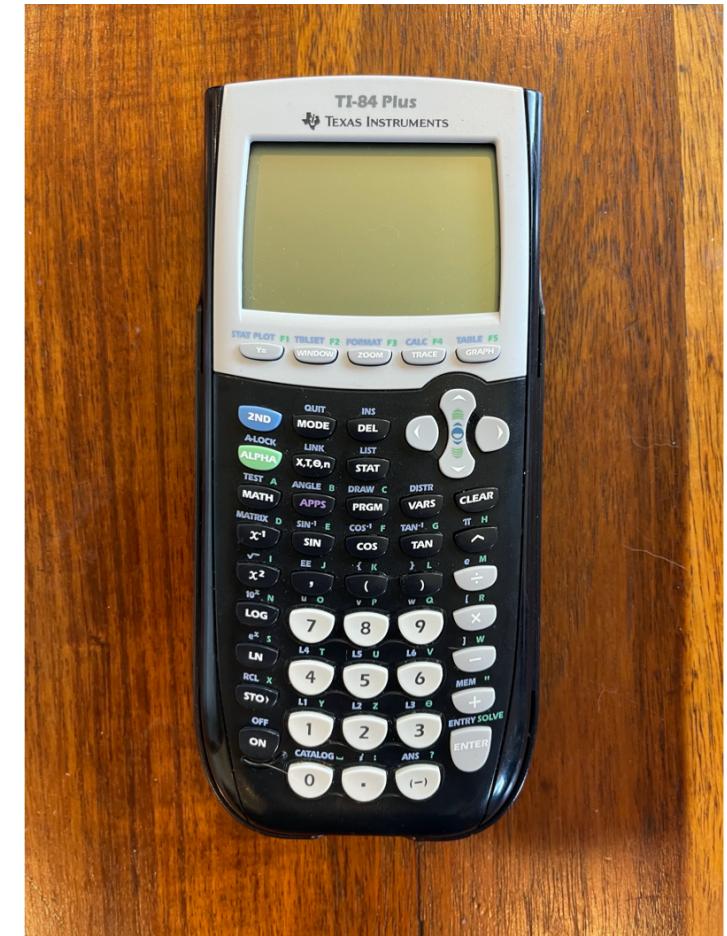
Reproducibility in research

- [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
- Accessible tools and data
- Everything is versioned
- Sharing and verifying 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) | | | | | | | | | |
| $R, N = \text{integer}$ | | | | | | | | | |
| $R, N \geq 0$ | | | | | | | | | |
| | | Razor | Navajo | | | | | | |
| Profit Contribution (\$/unit) | | 150 | 160 | Total Profit (\$) | | | | | |
| Units to Make | | 840 | 450 | 198000 [1] | | | | | |
| Resource requirements | | | | | | | | | |
| Frame Manufacturing | | 4 | 5 | E14 [2] | 5610 | | | | |
| Wheels and Deck Assembly | | 1.5 | 2 | | 2160 | | | | |
| QA and Packaging | | 1 | 0.8 | | 1200 | | | | |
| Available (hours) | | | | | | | | | |
| | | | | | | | | | |

[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**



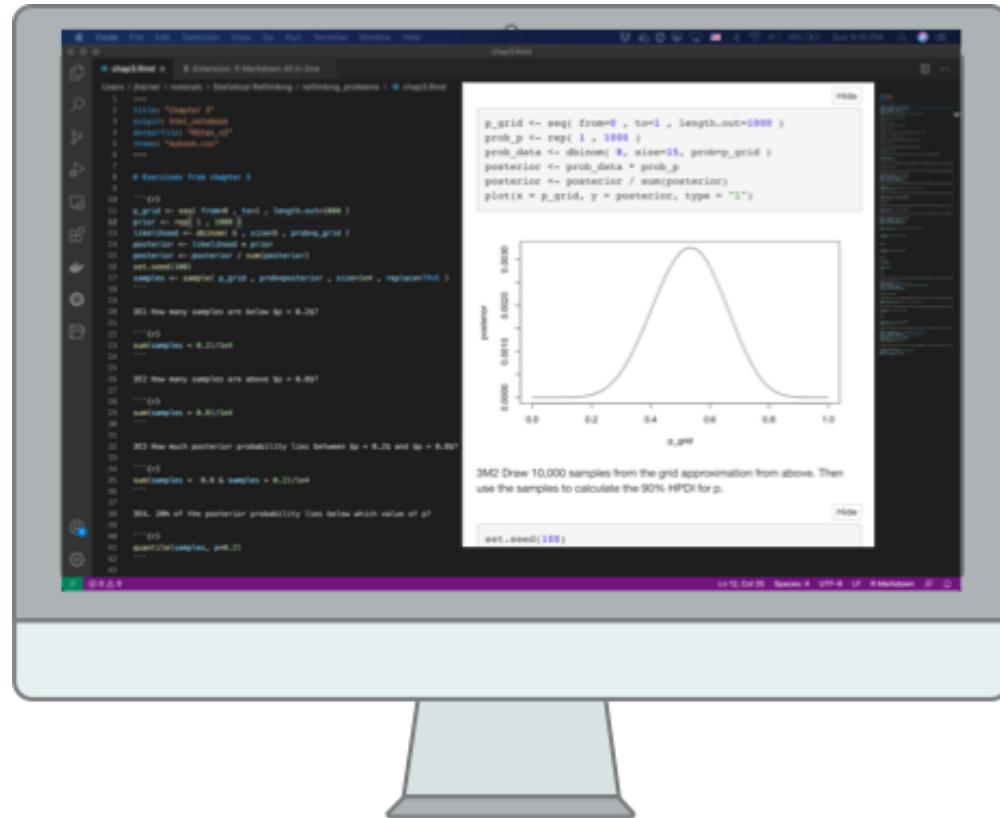
JIM HARNER



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

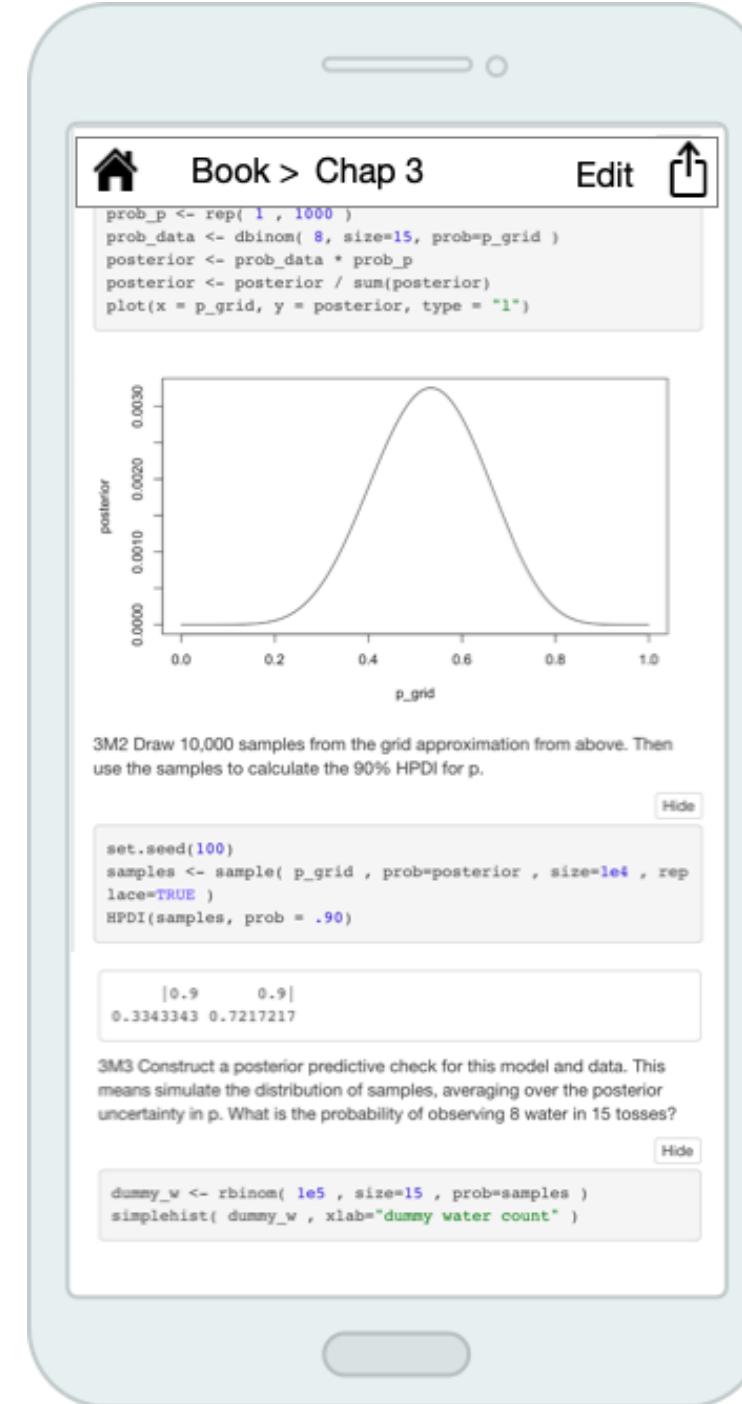


The screenshot shows a Jupyter Notebook interface with a single code cell containing R code. The code generates a probability distribution plot for a parameter π_1 . The plot shows a bell-shaped curve centered around 0.5, with the x-axis labeled π_1 and the y-axis labeled "prob". The distribution is symmetric and centered at 0.5, with most of the probability density between 0.2 and 0.8.

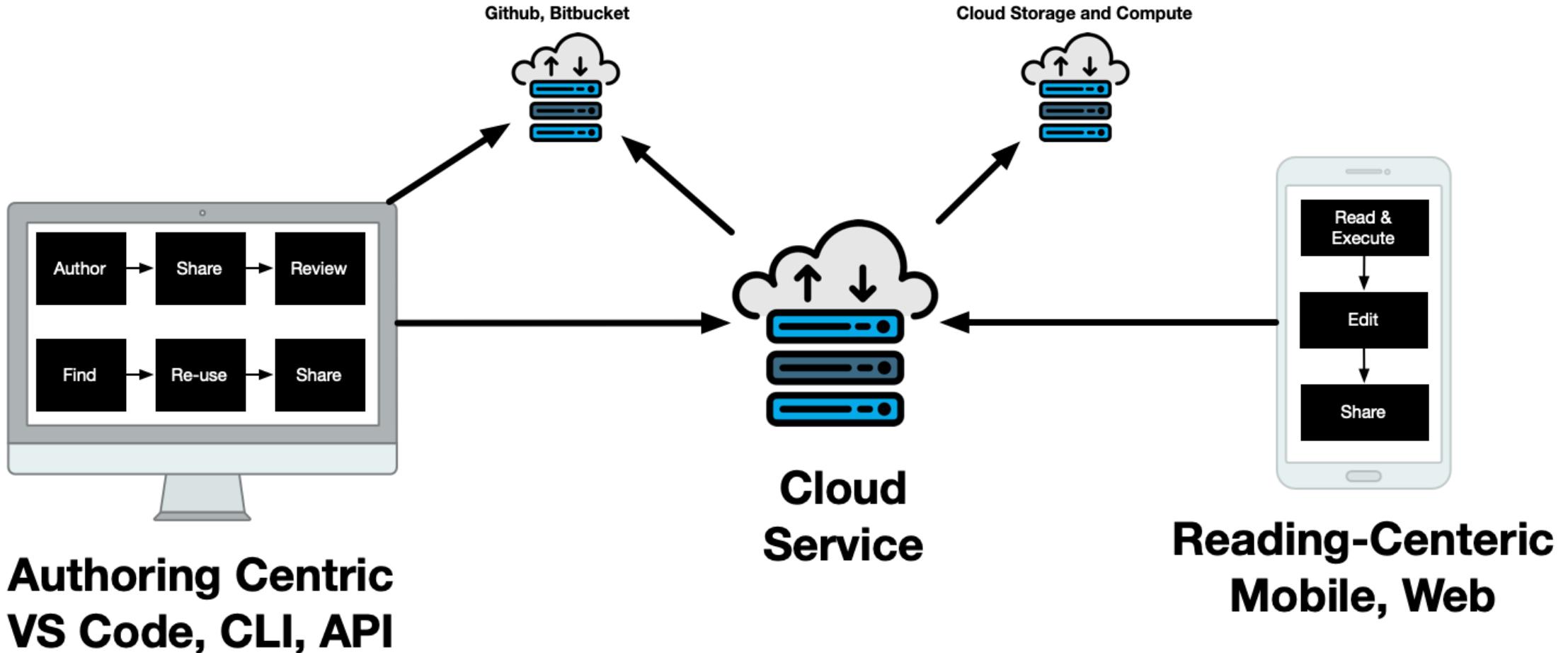
```
p_grid <- seq(from=0, to=1, length.out=1000)
prob_p <- rep(1, 1000)
prior <- dnorm(0.5, 0.04)
posterior <- dnorm(0.5, prior)
set.seed(100)
samples <- rnorm(p_grid, posterior, sd=sqrt(posterior))
plot(x = p_grid, y = posterior, type = "l")
```

Share, read, tinker

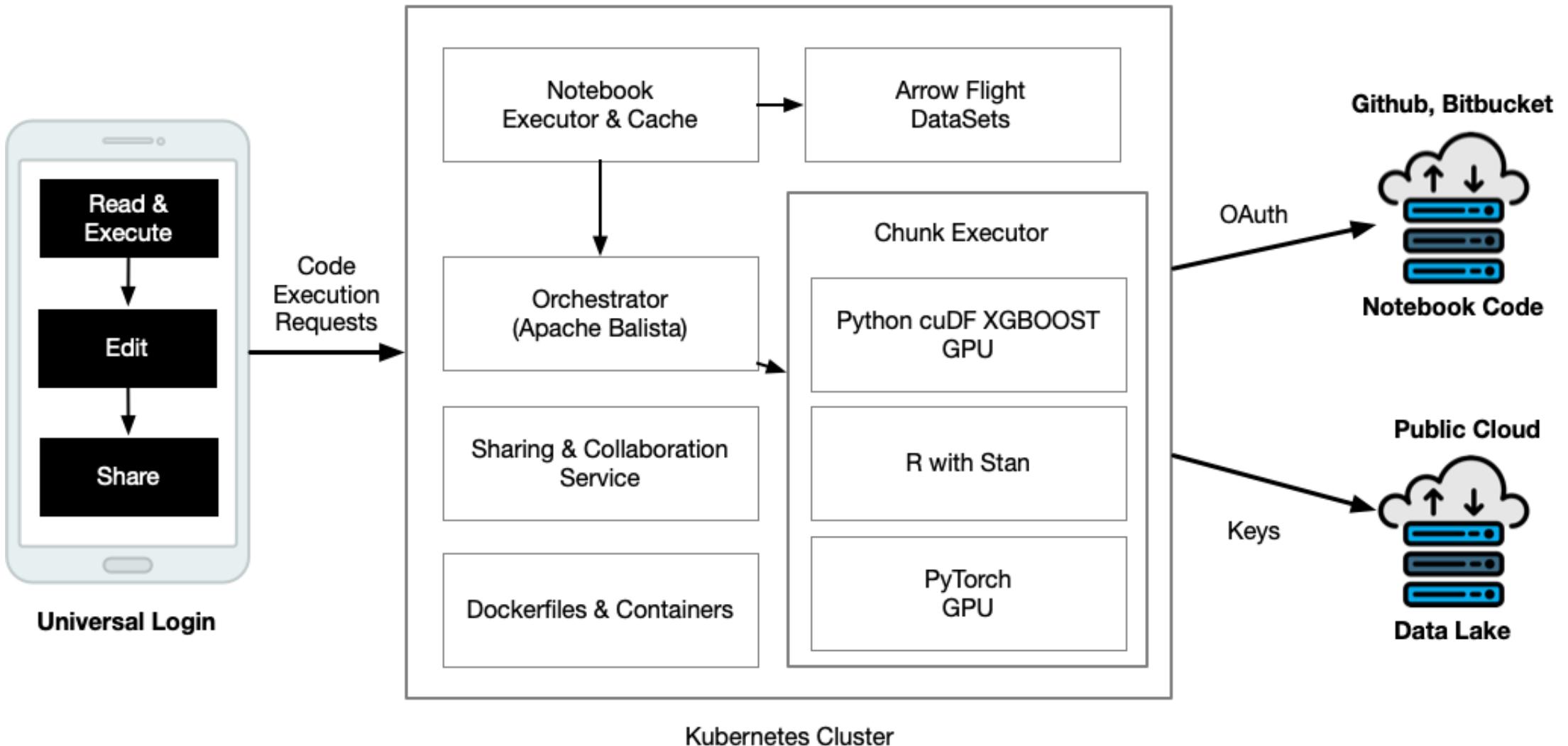
- Read, share, re-use interactive notebooks
- Mobile and Web
- Obviate the calculator
- Better than static books and reports
- Verifiable, tamperproof



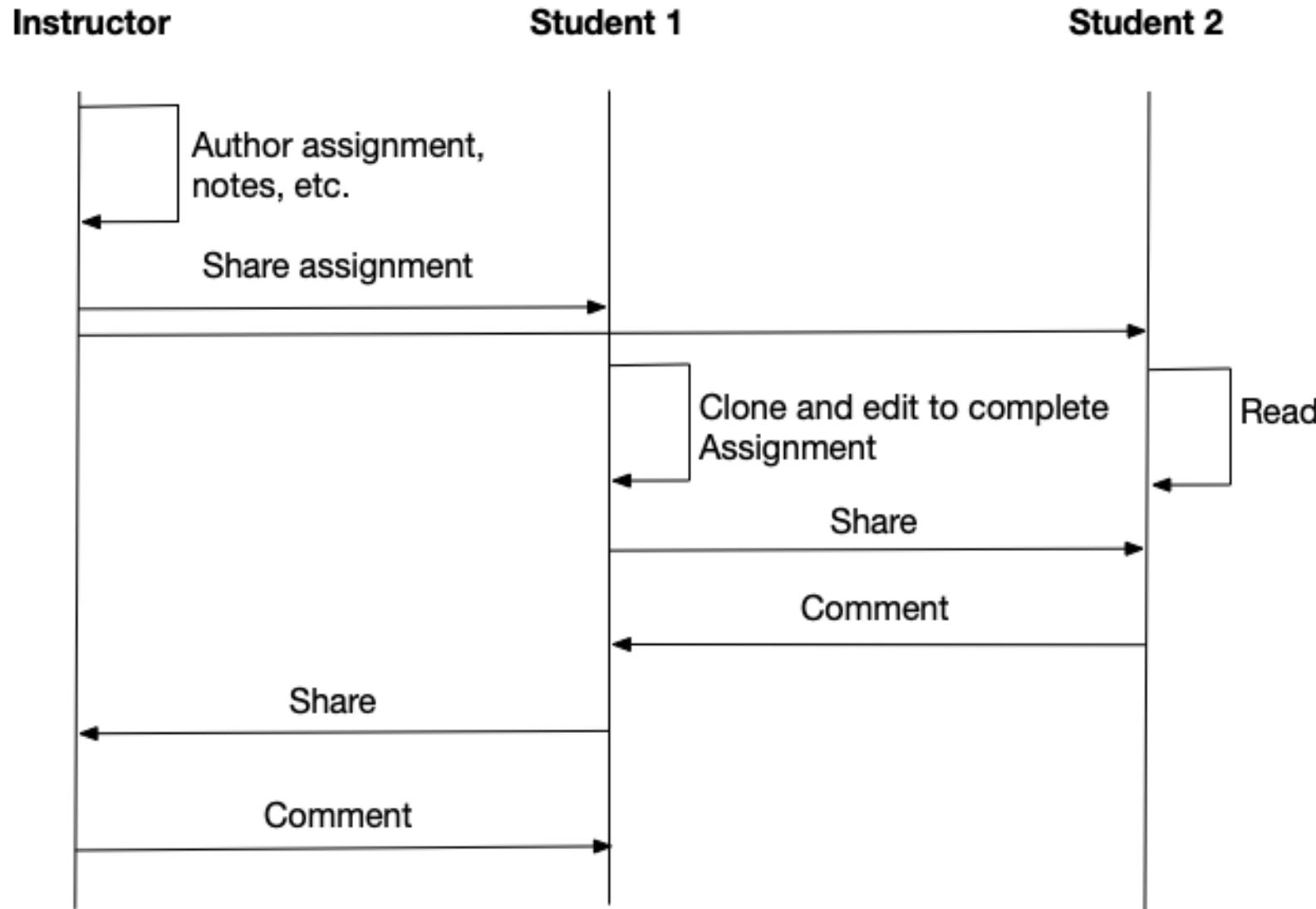
Hosted as a cloud service



Accessible, scalable cloud containers



Use-case: instructor student sharing



Next steps

- Now: Collect feedback from the community
- Sept 2021: early user testing
- Early 2022: conduct pilots in representative courses
- Develop as Open Source
- Finalize commercial component

Please get involved

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

Thank you

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

[Feedback and sign-up](#)



Feedback Form



Presentation