Coding Reproducibility: Clarity and replication

Kevin Lin
@linnylin92

Coding allows us to perform and automate data analysis.

- Consider the following hypothetical:
  - You analyze a dataset for a class's final.

Coding allows us to perform and automate data analysis.

- Consider the following hypothetical:
  - You analyze a dataset for a class's final.
  - You spend a week performing and writing up your report.

Coding allows us to perform and automate data analysis.

- Consider the following hypothetical:
  - You analyze a dataset for a class's final.
  - You spend a week performing and writing up your report.
  - Two days before the submission is due, you realize (in horror) that you analyzed an outdated dataset from Canvas. You now download the correct dataset.

Coding allows us to perform and automate data analysis.

- Consider the following hypothetical:
  - You analyze a dataset for a class's final.
  - You spend a week performing and writing up your report.
  - Two days before the submission is due, you realize (in horror) that you analyzed an outdated dataset from Canvas. You now download the correct dataset.
  - How easily can you redo your entire analysis?

Coding reproducibility is about clarity and replication.

- Clarity:
  - After returning to your old analysis after 6 months, how easily can you understand how your analysis works?

Coding reproducibility is about clarity and replication.

- Clarity:
  - After returning to your old analysis after 6 months, how easily can you understand how your analysis works?

- Replication:
  - If you re-ran your analysis (or received an updated dataset), how easily can you get the same (or updated) figures and results?

Recall the (possible) data analyses you've done in the past.
Were they reproducible?

Recall the (possible) data analyses you've done in the past.
Were they reproducible?

Recall the (possible) data analyses you've done in the past. Were they reproducible?

Recall the (possible) data analyses you've done in the past. Were they reproducible?

Recall the (possible) data analyses you've done in the past. Were they reproducible?

AP STATISTICS

M3 CHALLENGE™
Moody's Mega Math Challenge™
A CONTEST FOR HIGH SCHOOL STUDENTS

kaggle

WiDS Datathon

WOMEN IN DATA SCIENCE
@ STANFORD UNIVERSITY

Recall the (possible) data analyses you've done in the past. Were they reproducible?

AP STATISTICS

M3 CHALLENGE™
Moody's Mega Math Challenge™
A CONTEST FOR HIGH SCHOOL STUDENTS

kaggle

WiDS Datathon

WOMEN IN DATA SCIENCE
@ STANFORD UNIVERSITY

CITADEL

Recall the (possible) data analyses you've done in the past.
Were they reproducible?


Excel
R Markdown
GitHub
Domino Data Lab

Excel (and other click/formula-based softwares) are a good start, but have flaws that are not easy to overcome.

- Benefits:
  - Data, analysis, plots, tables, and results all in one file.

Excel (and other click/formula-based softwares) are a good start, but have flaws that are not easy to overcome.

- Benefits:
  - Data, analysis, plots, tables, and results all in one file.

- Potential problems:
  - New dataset? (New formatting might break your existing formulas)

Excel (and other click/formula-based softwares) are a good start, but have flaws that are not easy to overcome.

- Benefits:
  - Data, analysis, plots, tables, and results all in one file.

- Potential problems:
  - New dataset? (New formatting might break your existing formulas)
  - Sequential logic of analysis? (No clear ordering of which steps were first, second, etc. in your analysis.)

Excel (and other click/formula-based softwares) are a good start, but have flaws that are not easy to overcome.

- Benefits:
  - Data, analysis, plots, tables, and results all in one file.

- Potential problems:
  - New dataset? (New formatting might break your existing formulas)
  - Sequential logic of analysis? (No clear ordering of which steps were first, second, etc. in your analysis.)
  - Complex, custom functions? (What if you wanted to do something beyond the basic formulas?)

R Markdown offers some solutions, but it's not the end of the story either.

- Solutions:
  - R separates data from analysis, easily set up code to adapt to data.

R Markdown offers some solutions, but it's not the end of the story either.

- Solutions:
  - R separates data from analysis, easily set up code to adapt to data.
  - R Markdown weaves in code with plots, can be read from top to bottom.

R Markdown offers some solutions, but it's not the end of the story either.

- Solutions:
  - R separates data from analysis, easily set up code to adapt to data.
  - R Markdown weaves in code with plots, can be read from top to bottom.
  - R is a programming language.

R Markdown offers some solutions, but it's not the end of the story either.

- Solutions:
  - R separates data from analysis, easily set up code to adapt to data.
  - R Markdown weaves in code with plots, can be read from top to bottom.
  - R is a programming language.

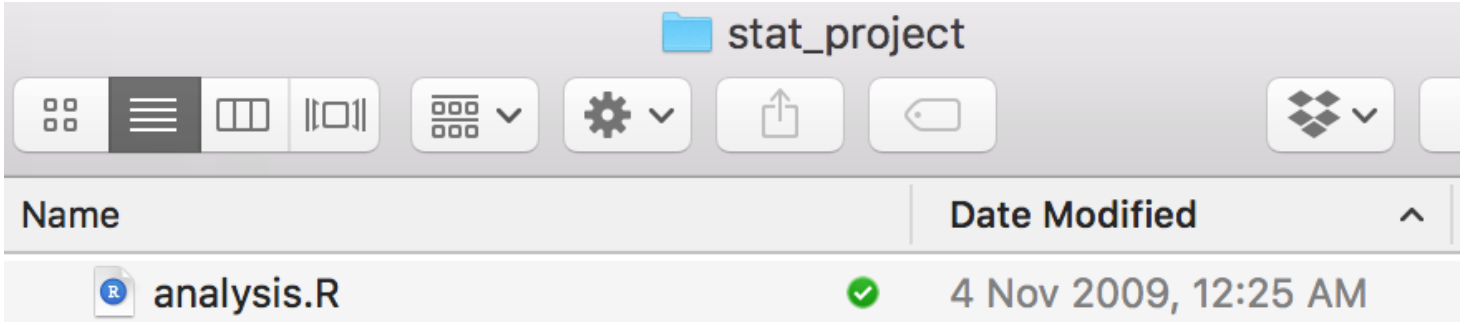
- Potential problems:
  - Clarity? (What if you start defining custom functions?)

R Markdown offers some solutions, but it's not the end of the story either.

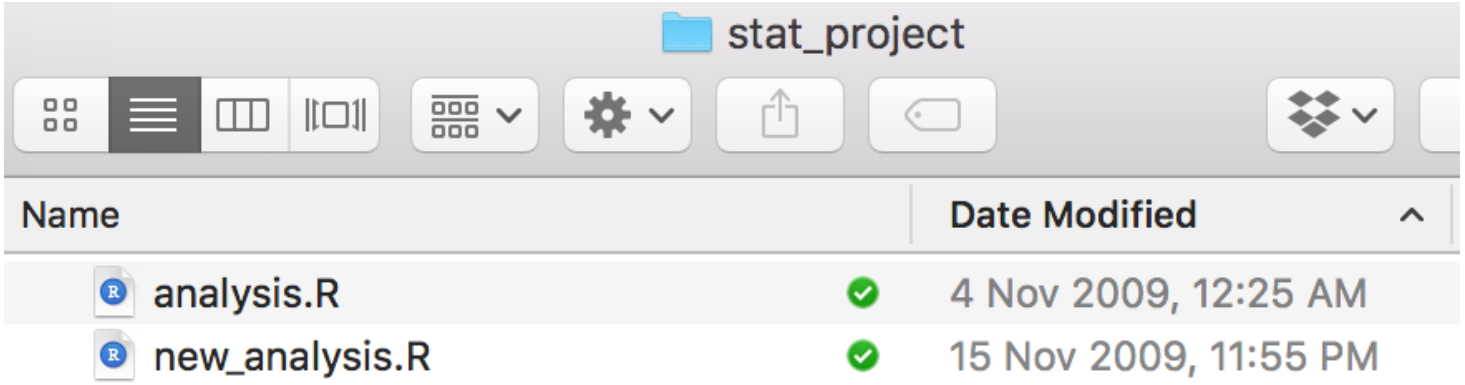
- Solutions:
  - R separates data from analysis, easily set up code to adapt to data.
  - R Markdown weaves in code with plots, can be read from top to bottom.
  - R is a programming language.

- Potential problems:
  - Clarity? (What if you start defining custom functions?)
  - Replication? (What if you have been trying many different ways to analyze data?)

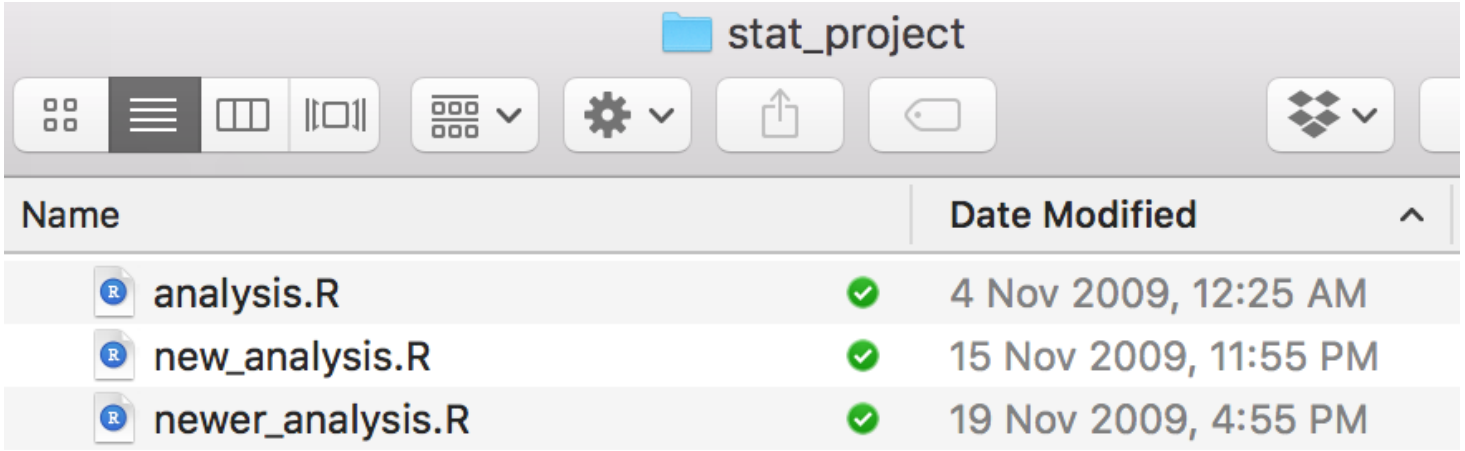
A repository is system that allows you to (painlessly) keep an annotated history of your files.



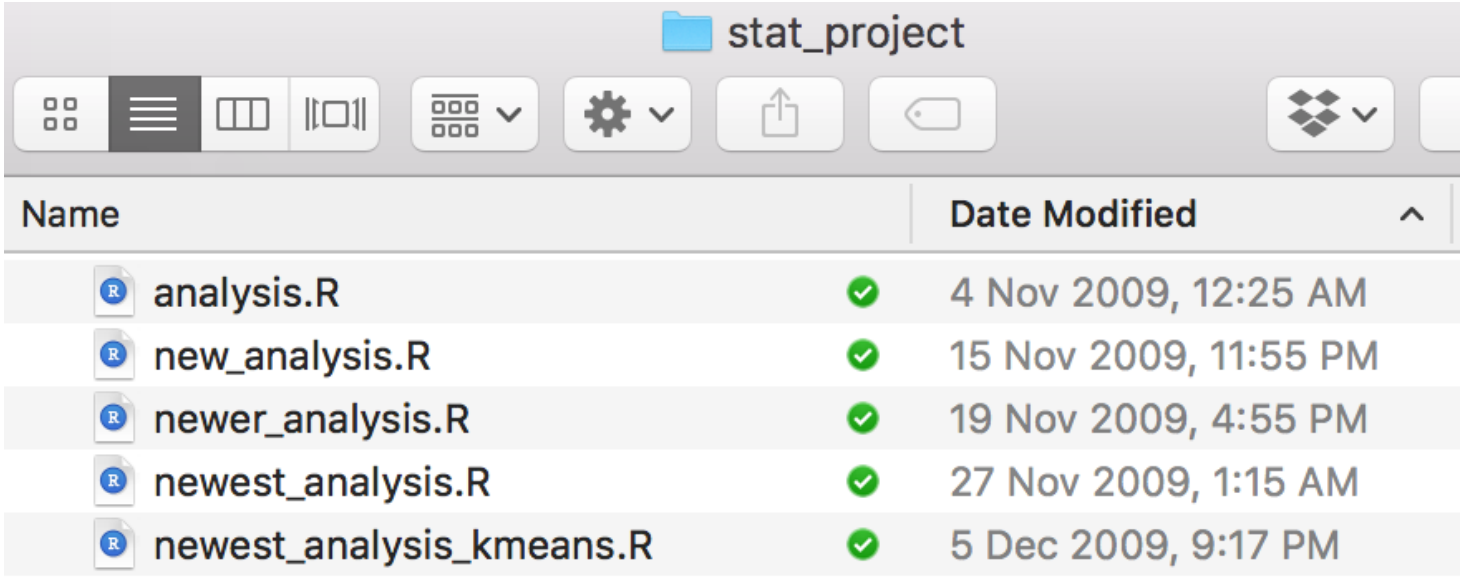
A repository is system that allows you to (painlessly) keep an annotated history of your files.



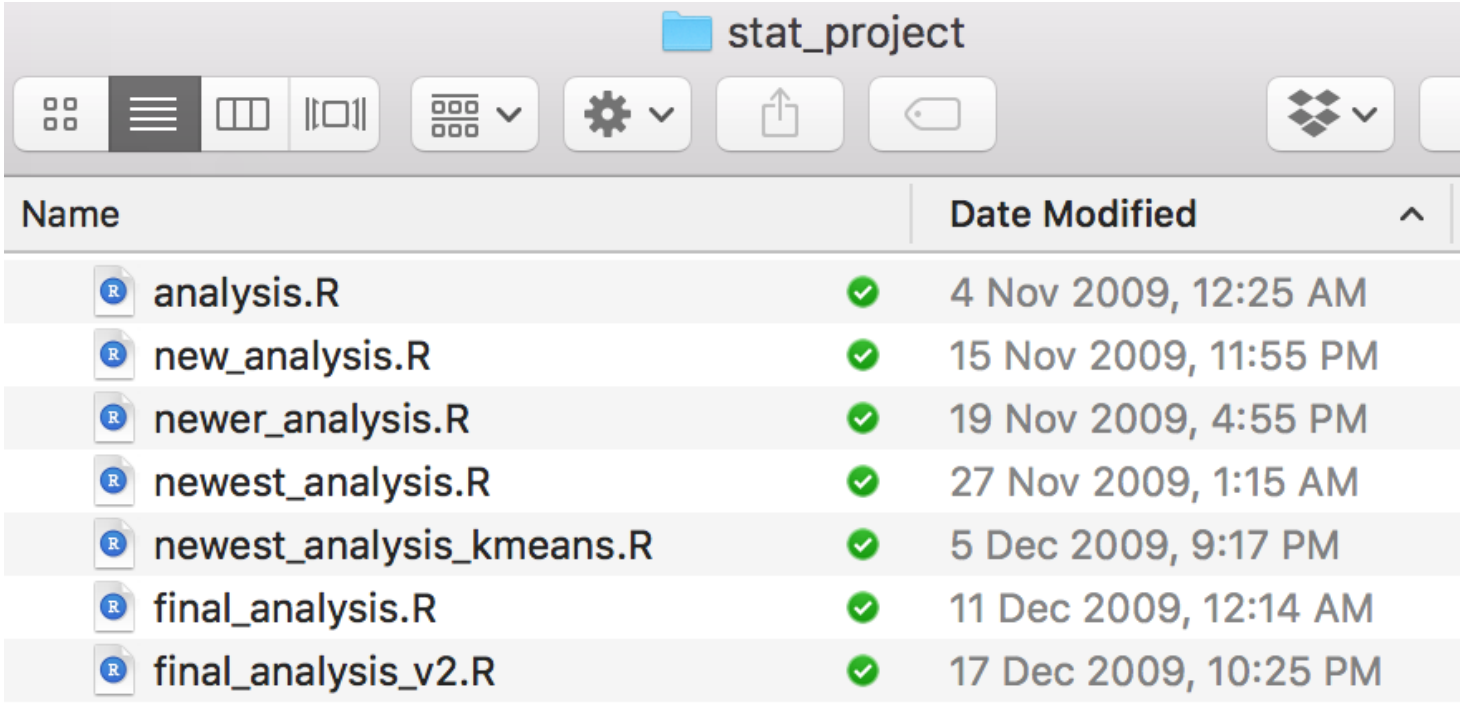
A repository is system that allows you to (painlessly) keep an annotated history of your files.



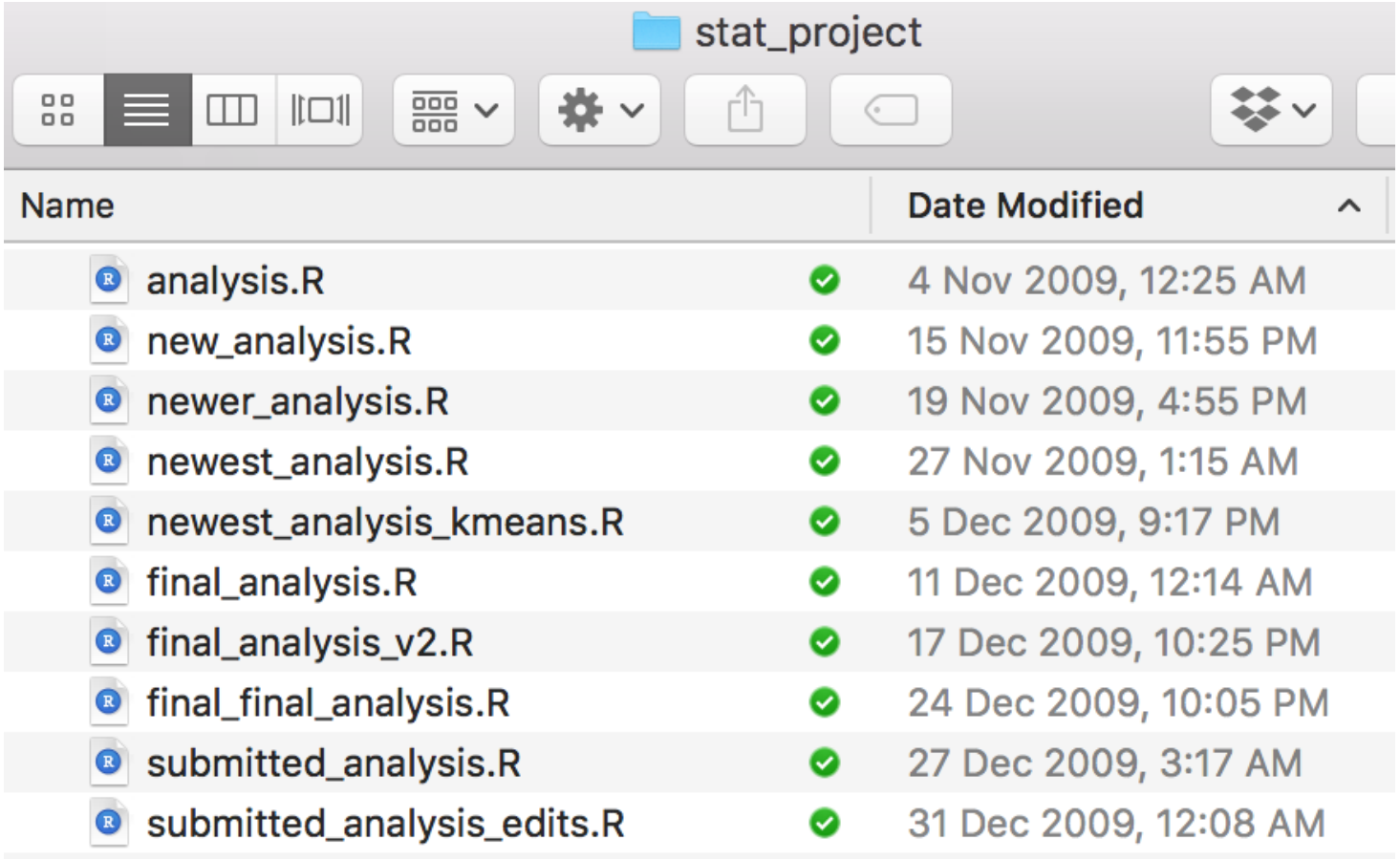
A repository is system that allows you to (painlessly) keep an annotated history of your files.



A repository is system that allows you to (painlessly) keep an annotated history of your files.



A repository is system that allows you to (painlessly) keep an annotated history of your files.



GitHub allows you to save an “annotated history” (called a repository) of your code.

- Solutions:
  - Tracks changes (behind the scenes) so you don’t need to duplicate files.
  - Allows you to annotate what you changed.

GitHub allows you to save an “annotated history” (called a repository) of your code.

- Solutions:
  - Tracks changes (behind the scenes) so you don’t need to duplicate files.
  - Allows you to annotate what you changed.

- Potential problems:
  - Different computers running same code? (Think of our HW2 with different text encodings)
  - Collaboration and benchmarking? (How do you share results and communicate ideas during design?)

GitHub allows you to save an “annotated history” (called a repository) of your code.

- Solutions:
  - Tracks changes (behind the scenes) so you don’t need to duplicate files.
  - Allows you to annotate what you changed.

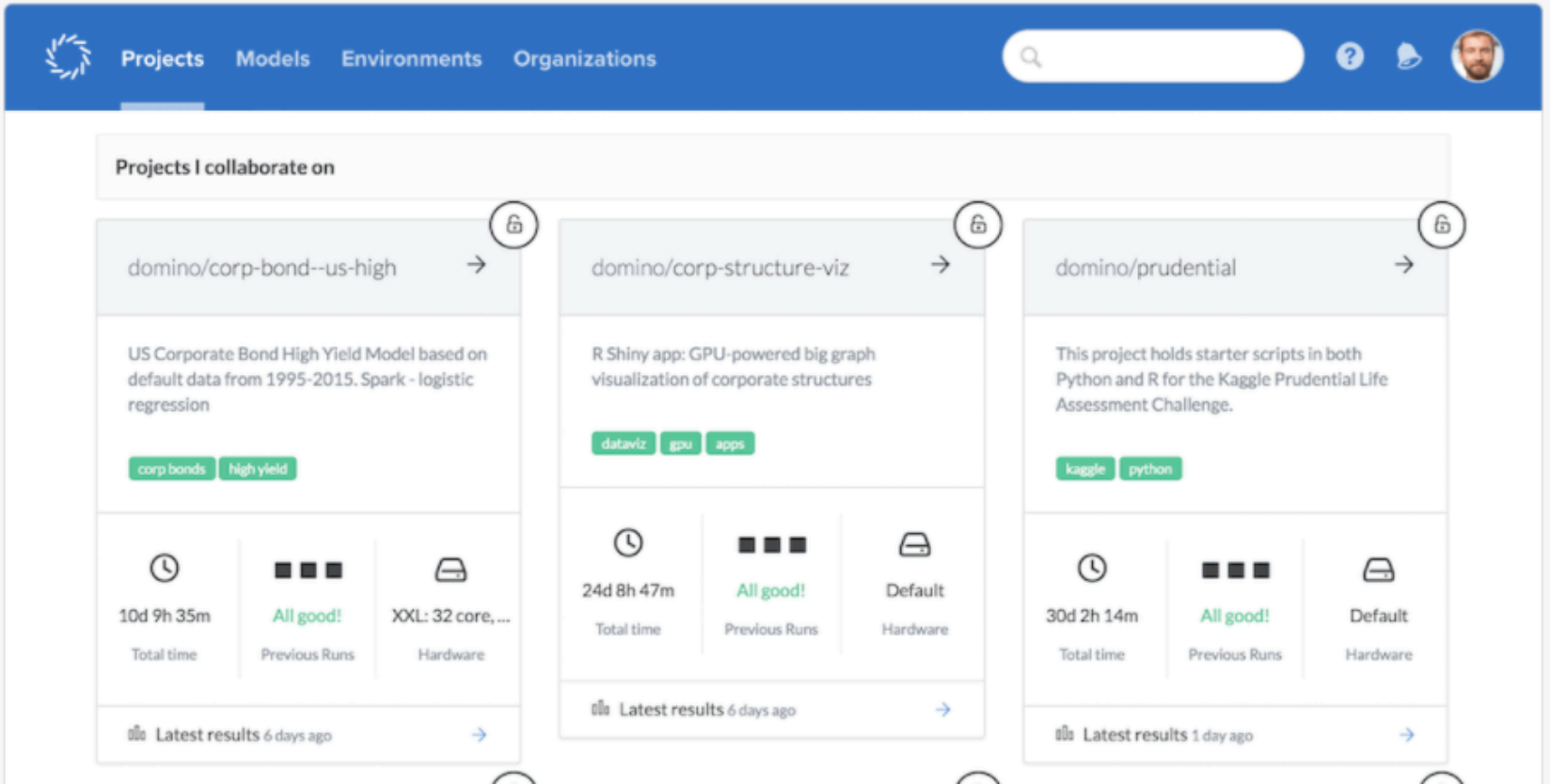
- Potential problems:
  - Different computers running same code? (Think of our HW2 with different text encodings)
  - Collaboration and benchmarking? (How do you share results and communicate ideas during design?)
  - Scalability? (What if you needed a lot of computational power?)
  - Publishing products? (How do you distribute your software to clients?)

Domino Data Labs offers business solutions to keep track of all changes across the company in a central platform.

Collaboration Hub & Reproducibility Engine

Increase productivity and reduce risk, together

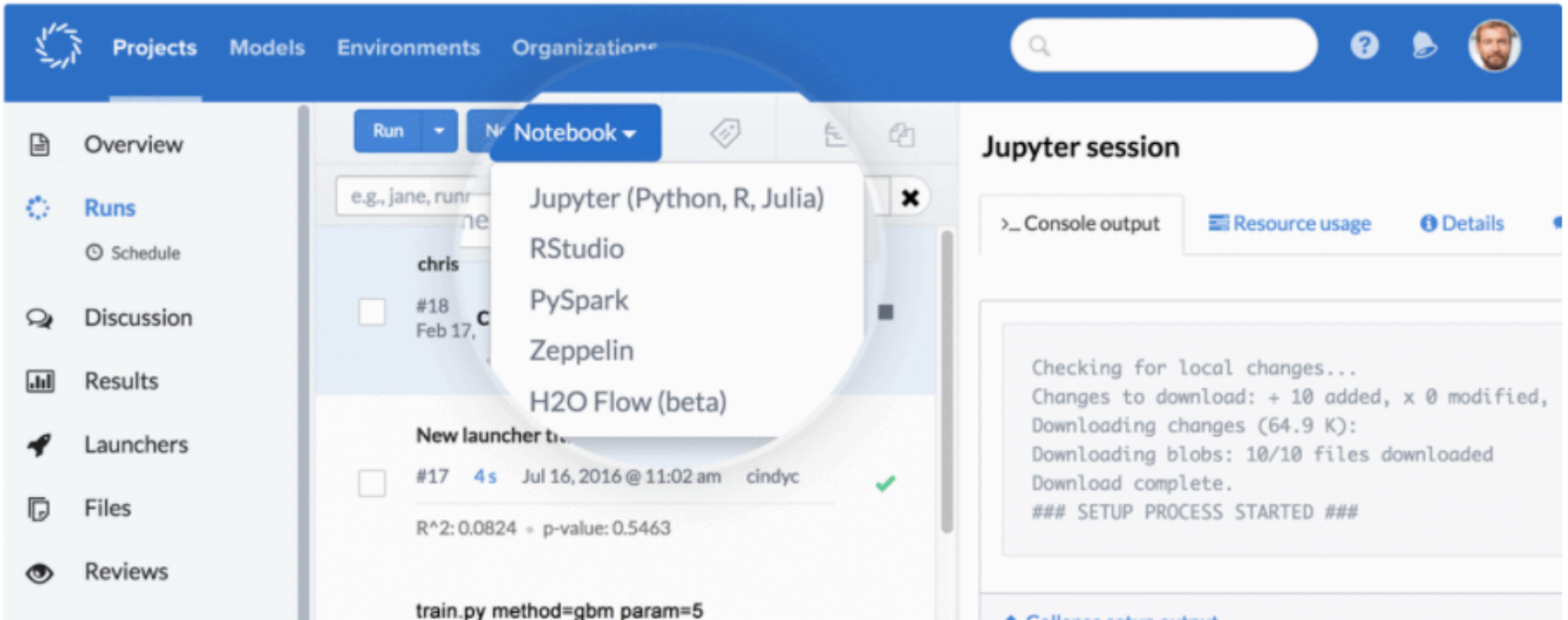
- **Reduce key-man and operational risk** by automatically preserving all key project information with Domino's containerization and dependency map– save data, software configurations, code, parameters, results, discussion, and delivered artifacts as they happen.
- **Streamline knowledgement management** with all projects stored, searchable, and forkable.
- **Avoid a cold start** by using native integrations with popular source control systems like GitHub.



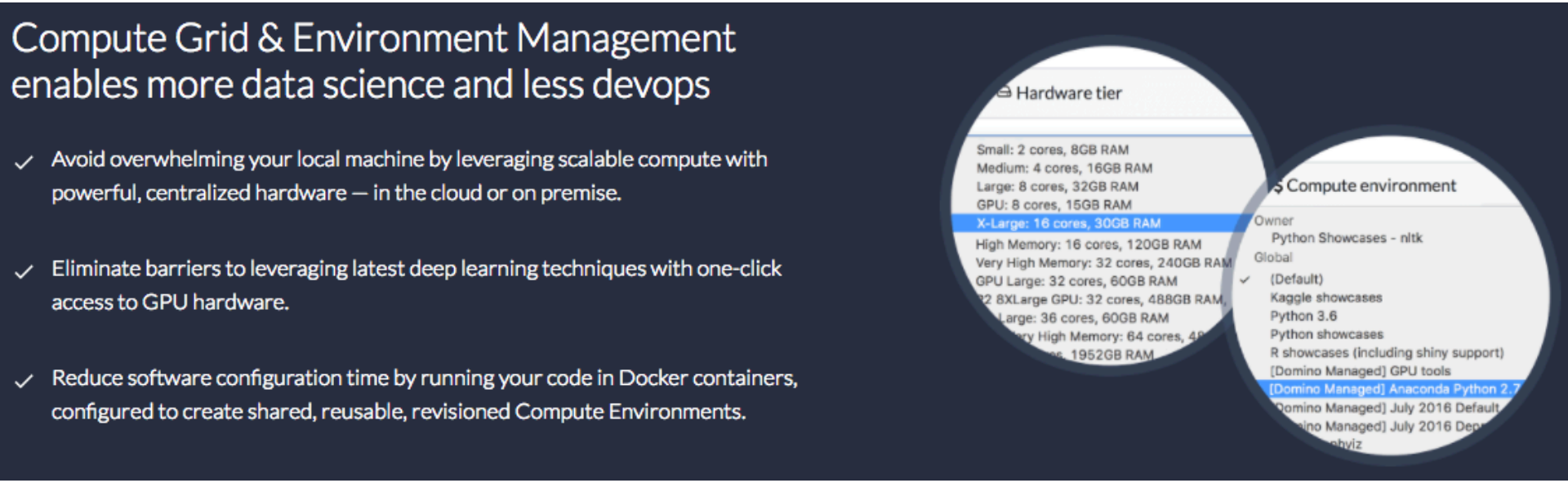
Data Science Workbench

Faster iteration and experimentation

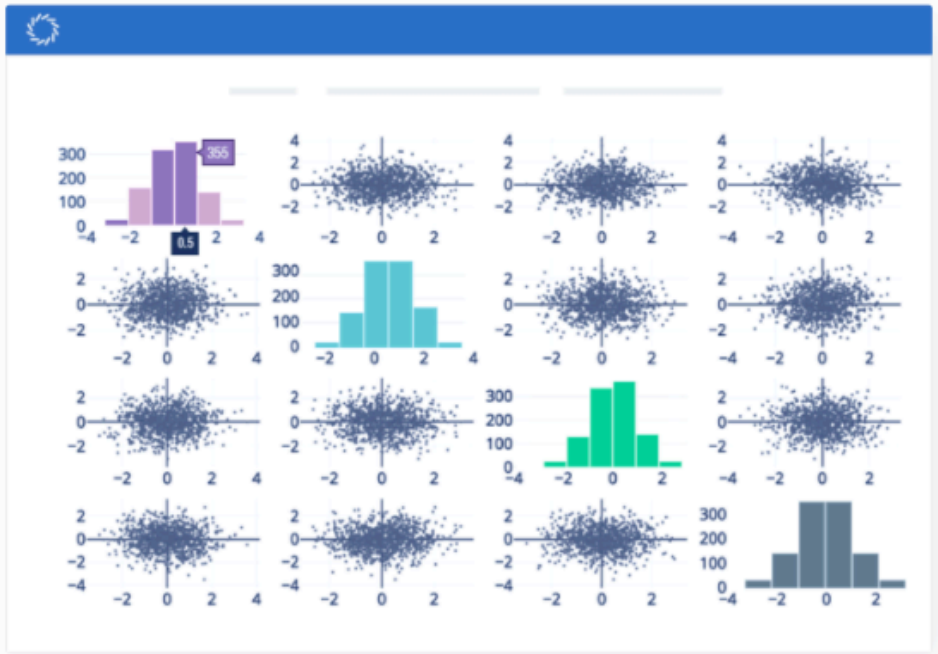
- **Work instantly by spinning up interactive workspaces with one click** — using the tools you already know and love, e.g., Jupyter, RStudio, SAS, and Zeppelin.
- **Tackle complex problems** by running, tracking, and comparing batch experiments in parallel with any language, even commercial languages like SAS or Matlab.
- **Minimize changes to your existing workflow** by connecting to any data including cloud databases and distributed systems like Hadoop and Spark
- **Instantly leverage all popular tools** with the our pre-packaged Domino Analytics Distribution (includes database drivers, Anaconda Python, popular deep learning packages, visualization packages, etc.) Or customize your own environment without risk of affecting other users.



Domino Data Labs offers business solutions to keep track of all changes across the company in a central platform.



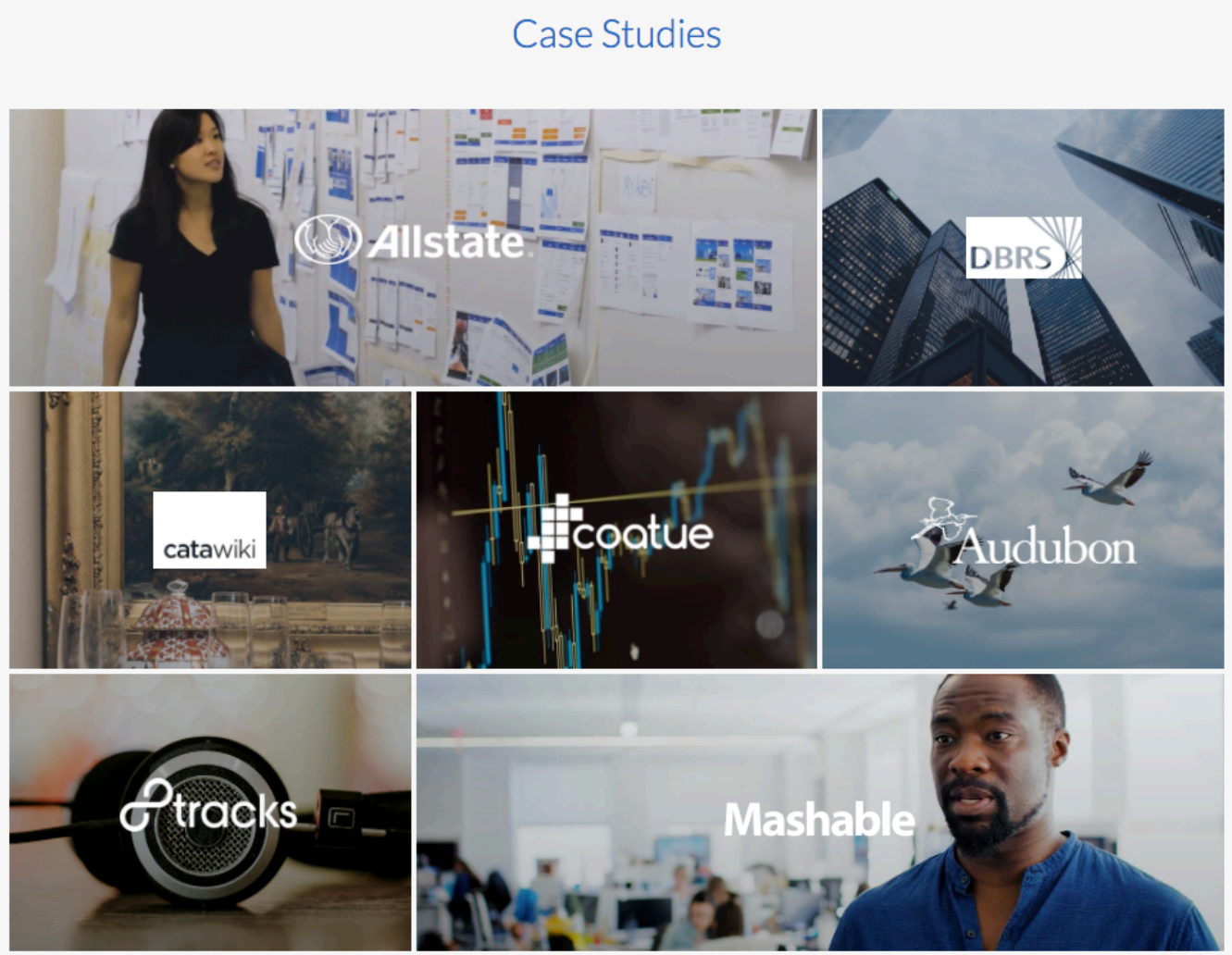
Domino Data Labs offers business solutions to keep track of all changes across the company in a central platform.



Deliver powerful insights to stakeholders

- **Communicate business benefits.** Publish visualizations using open source data science tools, including knitr, Plotly, D3, etc. or for commercial tools like Tableau.
- **Expose complex results in business-friendly manner.** Publish interactive dashboards and web apps using Shiny and Flask.
- **Remove low-value admin work.** Schedule recurring tasks to update reports — serve results through the web or send to stakeholders via email.

Domino Data Labs offers business solutions to keep track of all changes across the company in a central platform.



So what?

So what?

So what?

- People analyze data differently, and while there is not a universal “best” way to do this, some ways are certainly more reproducible than others.

So what?

- People analyze data differently, and while there is not a universal “best” way to do this, some ways are certainly more reproducible than others.

- While this course focuses on teaching you how to code in R, it’s important to be aware there are many ways to use R.

So what?

- People analyze data differently, and while there is not a universal “best” way to do this, some ways are certainly more reproducible than others.
- While this course focuses on teaching you how to code in R, it’s important to be aware there are many ways to use R.
- Be proactive in learning new tools. You never know what will be useful.

So what?

Imagine you had to do all your data analysis in only Excel. Forever.

Thank you!

Email me if you have further questions.

Shameless plug: Follow me on twitter (**@linnylin92**).