

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 - Two days before the submission is due, you realize (in horror) that you analyzed an outdated dataset from Canvas.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 - Two days before the submission is due, you realize (in horror) that you analyzed an outdated dataset from Canvas.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.

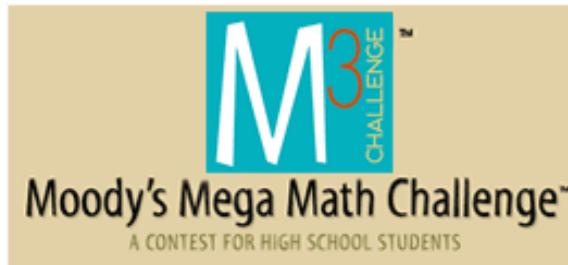
- Clarity:
 - After returning to your old analysis after 6 months, how easily can you understand how your analysis works?
- Replication:
 - If you re-ran your analysis (or received an updated dataset), how easily can you get the same (or updated) figures and results?

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

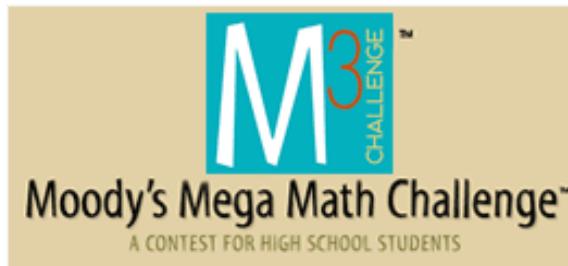
Recall the (possible) data analyses you've done in the past.
Were they reproducible?



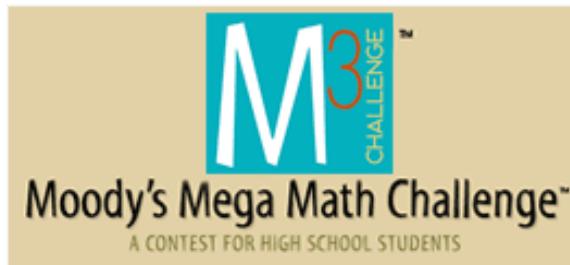
Recall the (possible) data analyses you've done in the past.
Were they reproducible?



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?

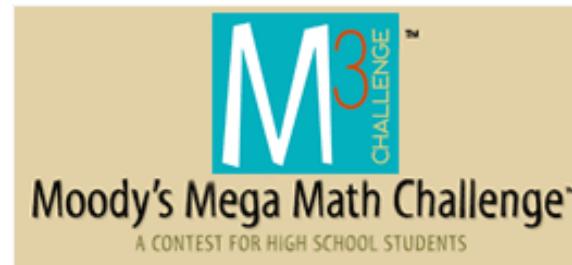


WiDS Datathon



WOMEN IN DATA SCIENCE
@ STANFORD UNIVERSITY

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



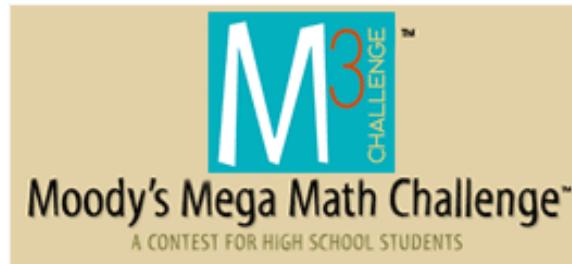
WiDS Datathon



WOMEN IN DATA SCIENCE
@ STANFORD UNIVERSITY



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



WiDS Datathon



Bloomberg

 CITADEL

 twitter

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.

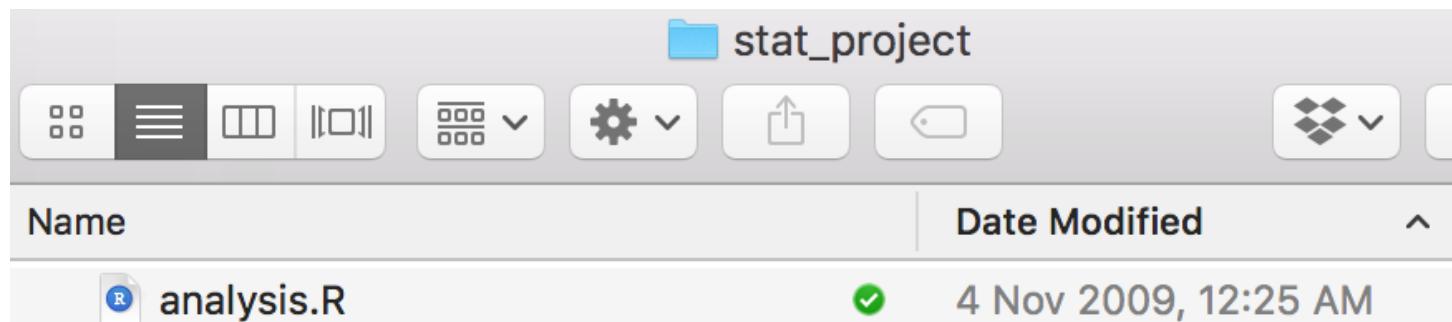
R Markdown offers some solutions, but it's not the end of the story either.

- 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?)

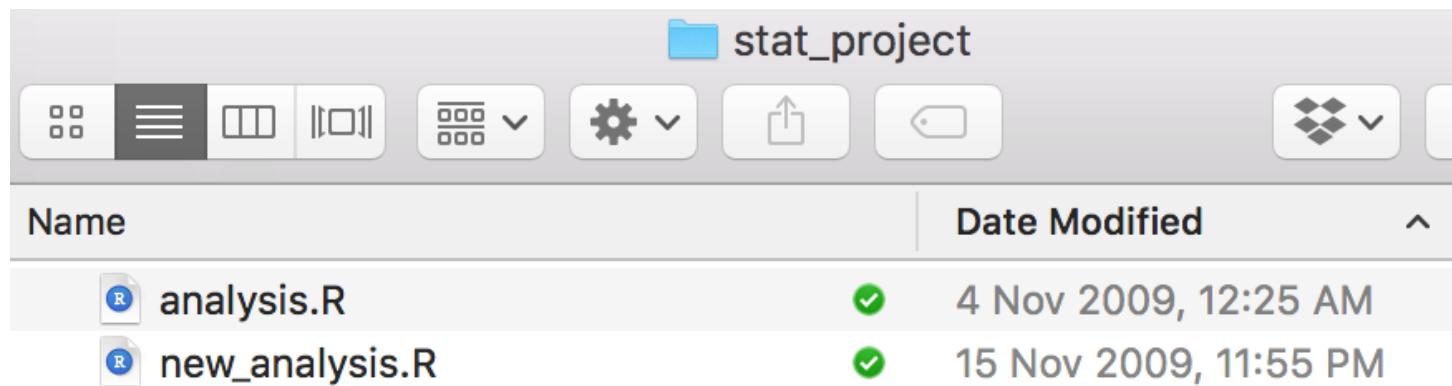
R Markdown offers some solutions, but it's not the end of the story either.

- 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?)

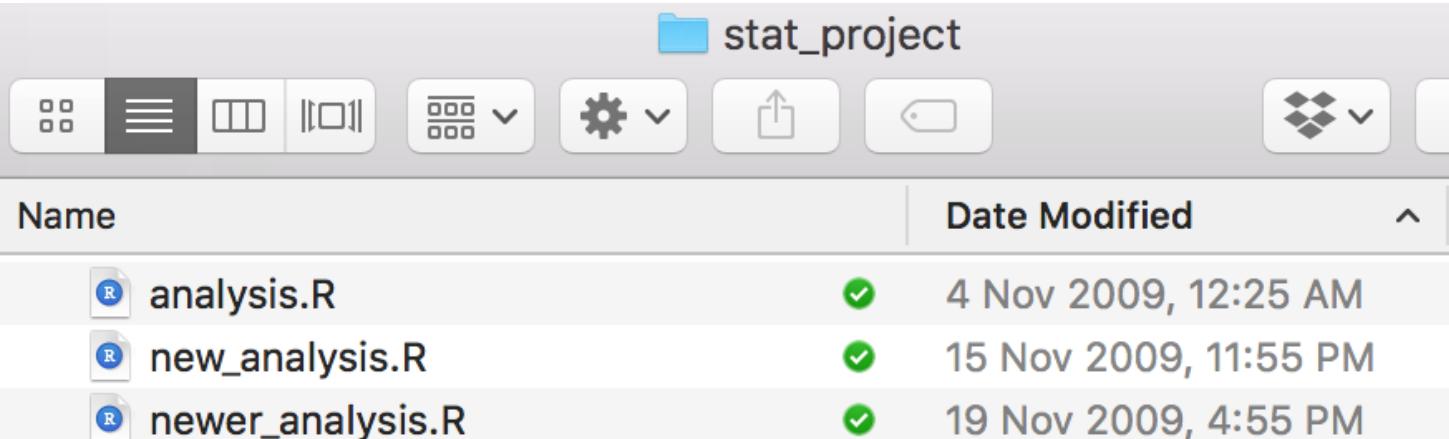
A repository is system that allows you to (painlessly) keep an annotated history of your files.



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

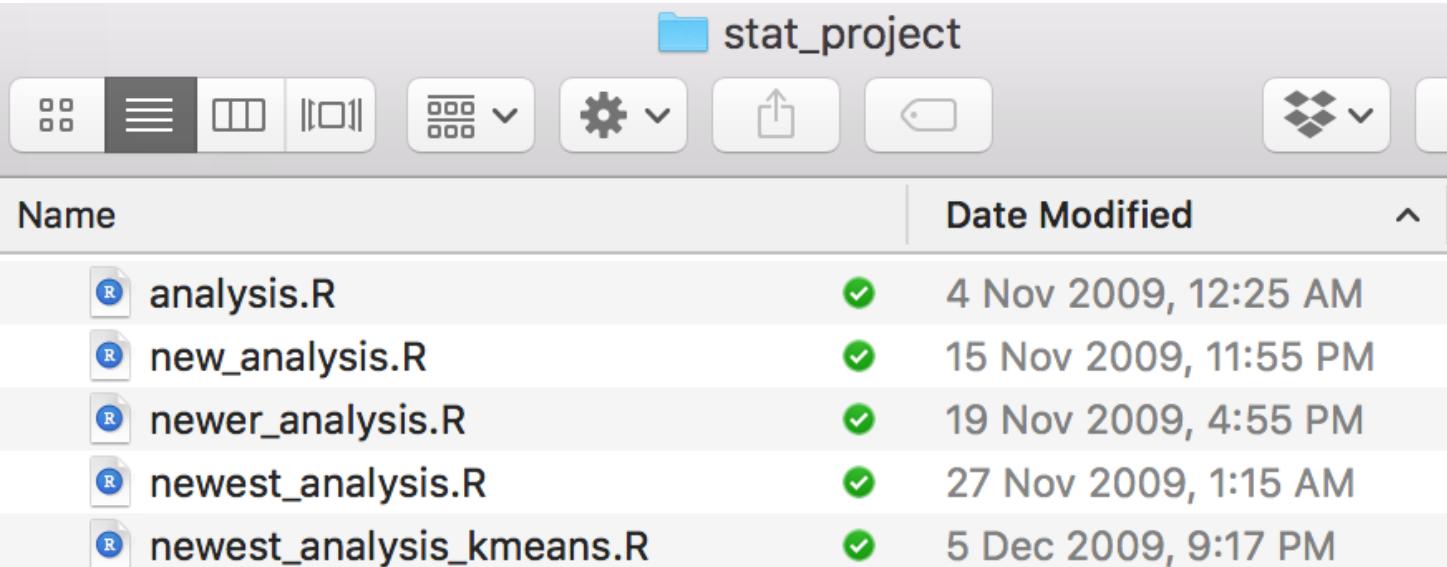


A repository is system that allows you to (p painlessly) keep an annotated history of your files.



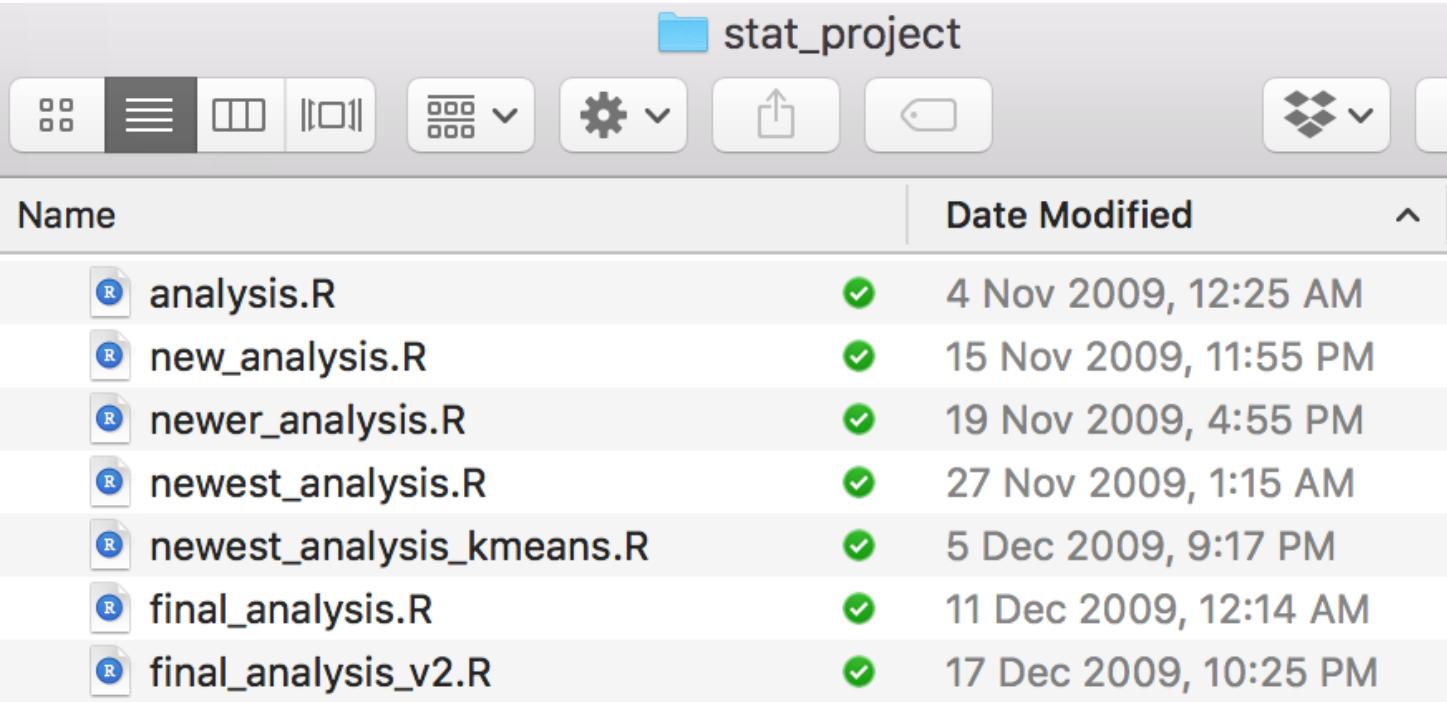
Name	Date Modified
analysis.R	4 Nov 2009, 12:25 AM
new_analysis.R	15 Nov 2009, 11:55 PM
newer_analysis.R	19 Nov 2009, 4:55 PM

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



Name	Date Modified
analysis.R	4 Nov 2009, 12:25 AM
new_analysis.R	15 Nov 2009, 11:55 PM
newer_analysis.R	19 Nov 2009, 4:55 PM
newest_analysis.R	27 Nov 2009, 1:15 AM
newest_analysis_kmeans.R	5 Dec 2009, 9:17 PM

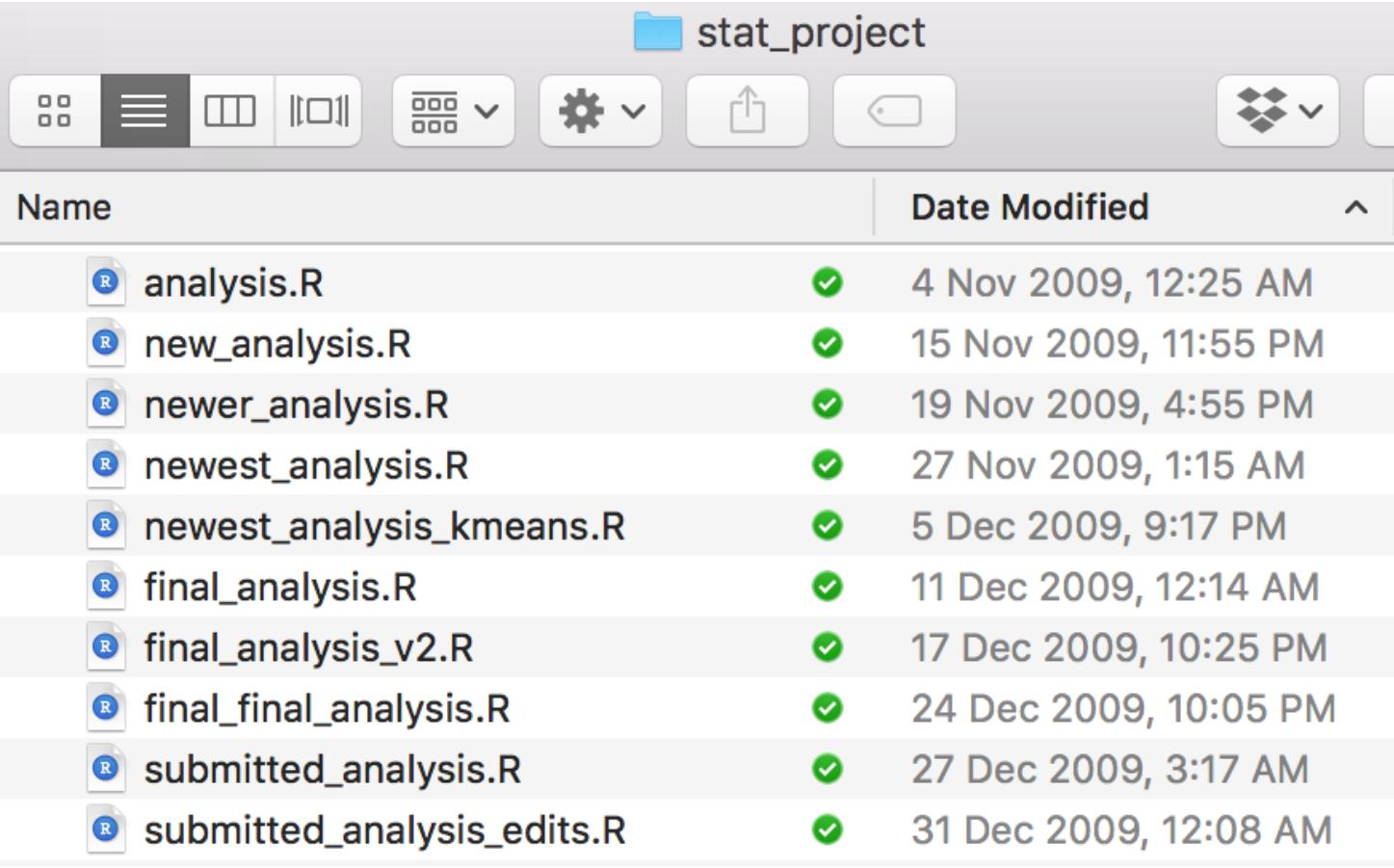
A repository is system that allows you to (painlessly) keep an annotated history of your files.



The screenshot shows a file browser window titled "stat_project". The interface includes a toolbar with various icons for file operations like copy, paste, and search. Below the toolbar is a table listing files. The columns are "Name" and "Date Modified". The "Date Modified" column is sorted in descending order, indicated by an upward arrow icon. The table lists seven R script files:

Name	Date Modified
analysis.R	4 Nov 2009, 12:25 AM
new_analysis.R	15 Nov 2009, 11:55 PM
newer_analysis.R	19 Nov 2009, 4:55 PM
newest_analysis.R	27 Nov 2009, 1:15 AM
newest_analysis_kmeans.R	5 Dec 2009, 9:17 PM
final_analysis.R	11 Dec 2009, 12:14 AM
final_analysis_v2.R	17 Dec 2009, 10:25 PM

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



The screenshot shows a file browser window titled "stat_project". The toolbar includes icons for file operations like New, Open, Save, and Delete, along with a gear icon for settings and a sync icon. The main area is a table listing files:

Name	Date Modified
analysis.R	4 Nov 2009, 12:25 AM
new_analysis.R	15 Nov 2009, 11:55 PM
newer_analysis.R	19 Nov 2009, 4:55 PM
newest_analysis.R	27 Nov 2009, 1:15 AM
newest_analysis_kmeans.R	5 Dec 2009, 9:17 PM
final_analysis.R	11 Dec 2009, 12:14 AM
final_analysis_v2.R	17 Dec 2009, 10:25 PM
final_final_analysis.R	24 Dec 2009, 10:05 PM
submitted_analysis.R	27 Dec 2009, 3:17 AM
submitted_analysis_edits.R	31 Dec 2009, 12:08 AM

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 knowledge management with all projects stored, searchable, and forkable.
- Avoid a cold start by using native integrations with popular source control systems like GitHub.

The screenshot shows the interface of the Collaboration Hub & Reproducibility Engine. At the top, there is a navigation bar with icons for Projects, Models, Environments, and Organizations, a search bar, and user profile icons. Below the navigation bar, the main content area displays three project cards under the heading "Projects I collaborate on".

- domino/corp-bond--us-high**: US Corporate Bond High Yield Model based on default data from 1995-2015. Spark - logistic regression. Tags: corp bonds, high yield. Status: All good!. Total time: 10d 9h 35m. Previous Runs: 3. Hardware: XXL: 32 core, ... Latest results: 6 days ago.
- domino/corp-structure-viz**: R Shiny app: GPU-powered big graph visualization of corporate structures. Tags: dataviz, gpu, apps. Status: All good!. Total time: 24d 8h 47m. Previous Runs: 3. Hardware: Default. Latest results: 6 days ago.
- domino/prudential**: This project holds starter scripts in both Python and R for the Kaggle Prudential Life Assessment Challenge. Tags: kaggle, python. Status: All good!. Total time: 30d 2h 14m. Previous Runs: 3. Hardware: Default. Latest results: 1 day ago.

Data Science Workbench

Faster iteration and experimentation

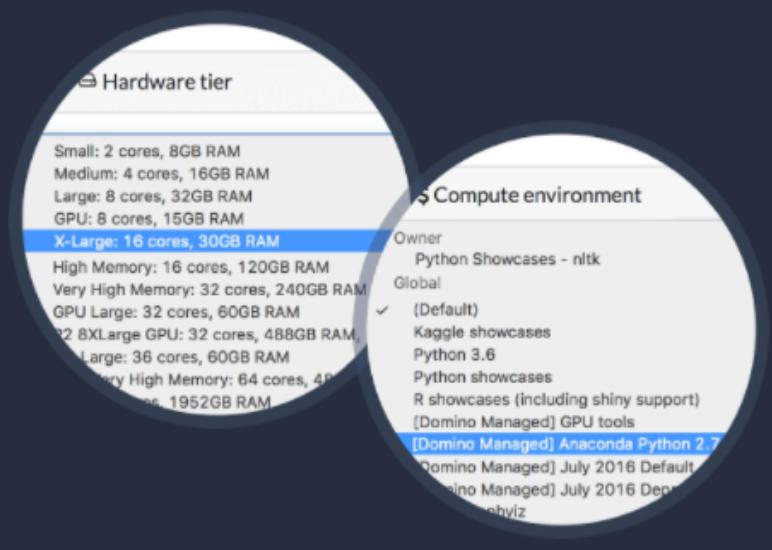
- Work instantly by spinning up interactive workspaces with one click – using the tools you already know and love, e.g., Jupyter, RStudio, SAS, and Zeppelin.
- Tackle complex problems by running, tracking, and comparing batch experiments in parallel with any language, even commercial languages like SAS 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.

The screenshot shows the Data Science Workbench interface. At the top, there's a navigation bar with icons for Projects, Models, Environments, and Organizations, along with a search bar and user profile icons. The main area is titled "Jupyter session". On the left, there's a sidebar with links for Overview, Runs, Discussion, Results, Launchers, Files, and Reviews. The central part of the screen displays a "Notebook" dropdown menu with options like "Run", "New Notebook", "e.g., jane, runn...", "chris", "#18 Feb 17, 2016", "New launcher tr...", "H2O Flow (beta)", and "R^2: 0.0824 • p-value: 0.5463". To the right of the notebook dropdown is a "Jupyter session" panel with tabs for "Console output", "Resource usage", and "Details". The "Console output" tab shows logs: "Checking for local changes... Changes to download: + 10 added, x 0 modified, Downloading changes (64.9 K): Downloading blobs: 10/10 files downloaded Download complete. ### SETUP PROCESS STARTED ###".

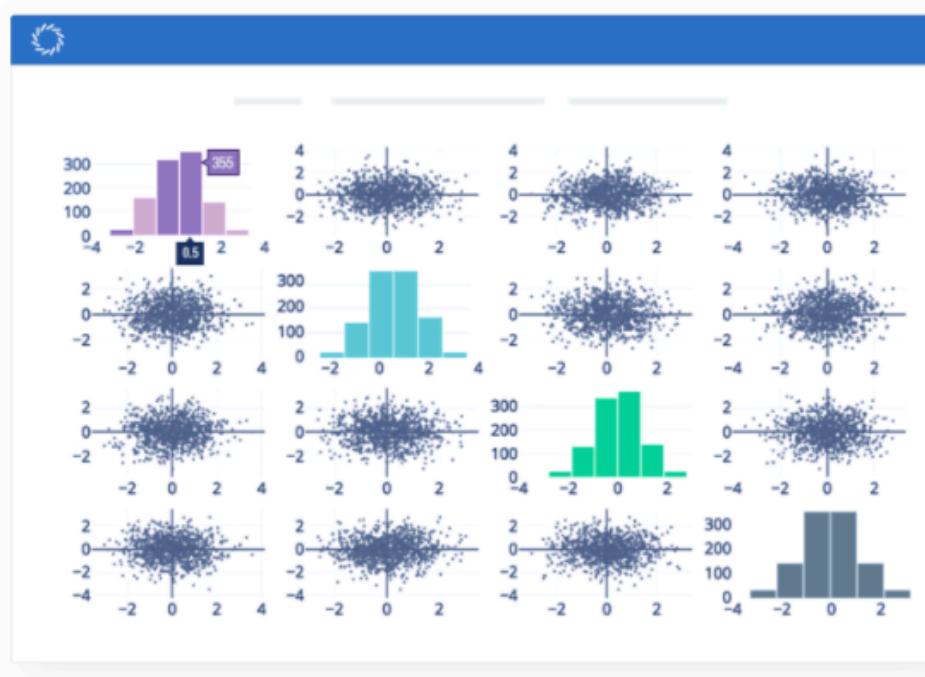
Domino Data Labs offers business solutions to keep track of all changes across the company in a central platform.

Compute Grid & Environment Management enables more data science and less devops

- ✓ Avoid overwhelming your local machine by leveraging scalable compute with powerful, centralized hardware – in the cloud or on premise.
- ✓ Eliminate barriers to leveraging latest deep learning techniques with one-click access to GPU hardware.
- ✓ Reduce software configuration time by running your code in Docker containers, configured to create shared, reusable, revisioned Compute Environments.



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

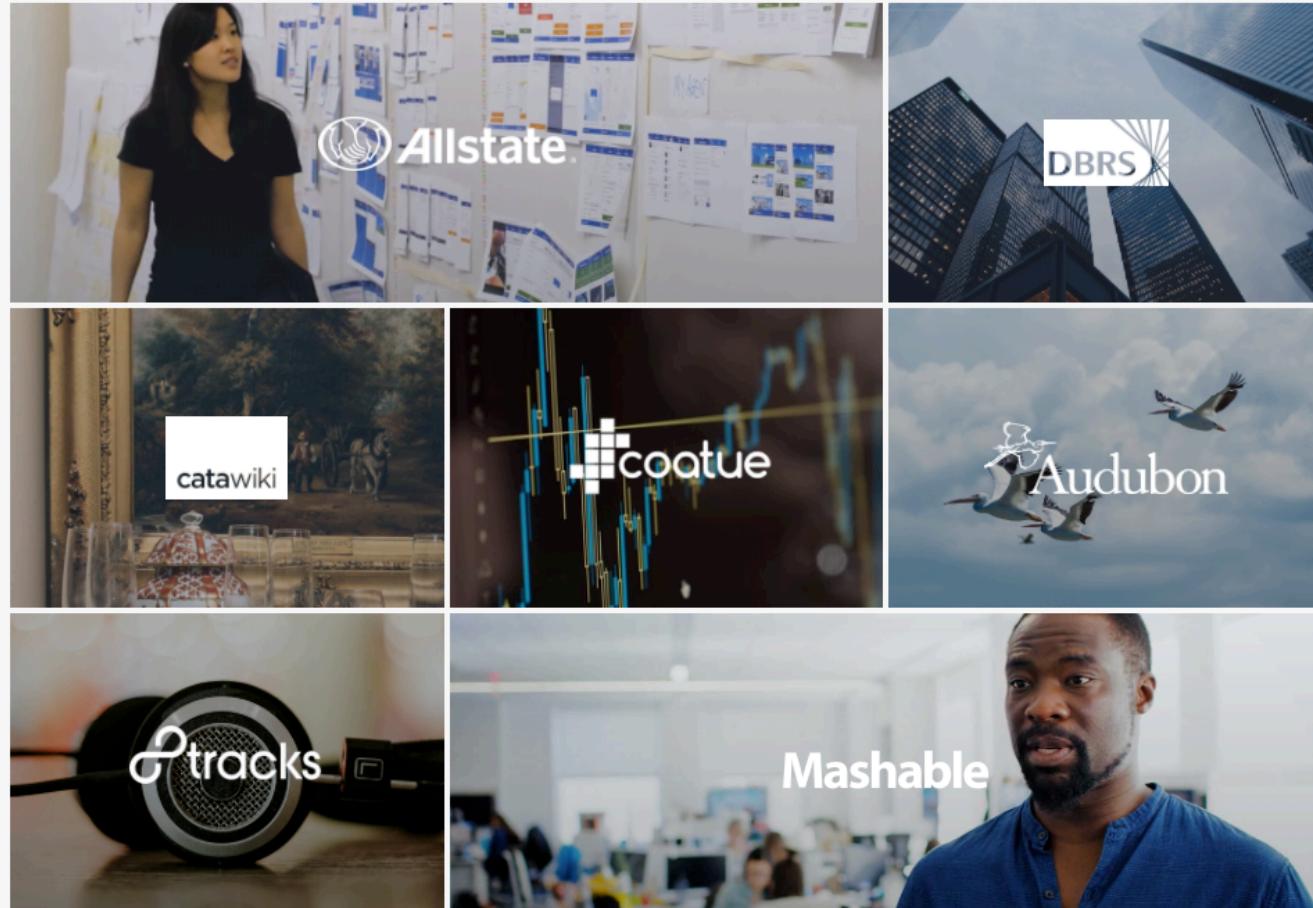


Deliver powerful insights to stakeholders

- **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.

Case Studies



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](https://twitter.com/linnylin92)).