REDIVIS

Data analysis and machine learning workflows on Redivis

Ian Mathews · ian@redivis.com

Stanford

Center for Population Health Sciences

Today’s agenda

- Brief introduction of Redivis at Stanford
- Redivis core concepts and analysis walkthrough
- Let’s do some Machine Learning!
- Q&A

Our mission

Redivis enables research centers to distribute rich datasets, and provides scientists with the means to understand them.

We strive to reduce barriers in working with data, and to create intuitive tools that make data science accessible and reproducible.

Redivis, a.k.a “The Stanford Data Farm”

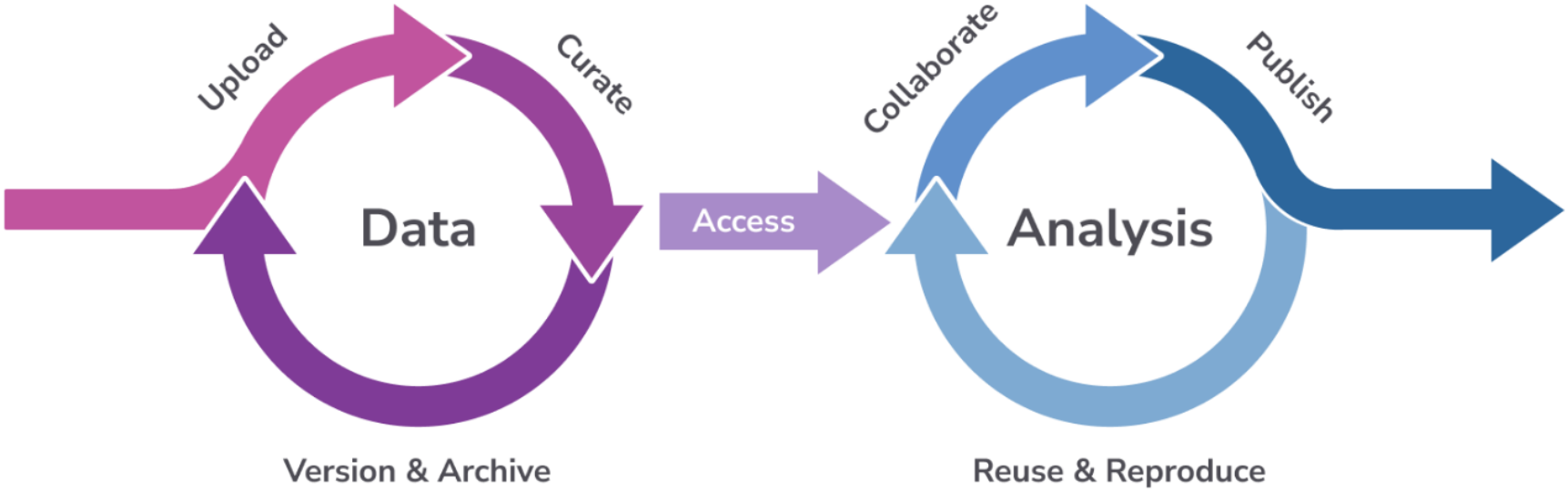
- The deployment of Redivis at Stanford: https://redivis.com/Stanford
- Home to 12 organizations, 363 datasets, 194 TB of data.
- Groups from medicine, sustainability, business, education, and more.

Stanford | Center for Population Health Sciences
Community Engagement | Data | Education | Research

STANFORD BUSINESS GRADUATE SCHOOL OF

Stanford | LIBRARIES

Redivis is a comprehensive data platform



Analyzing data on Redivis

- **High-performance compute:** Use SQL, R, Python, Stata, SAS, or a no-code interface.
- **Interdisciplinary:** Combine any dataset you have access to – even from other institutions.
- **Collaborative:** Work with your peers in real time, and build off of others' work, all within a secure environment.
- **Reproducible:** Full code history and ability to revert to previous state. Aligned with data + code sharing funder guidelines.

Data analysis walkthrough

Let’s talk machine learning

- What do I actually do?
  - Inference with a pre-trained model -> “classify these patient notes”
  - Fine-tuning a pre-trained model with labeled data
  - Creating a novel model from scratch (hardest / most $$)
- Where do I start?
  - Hugging Face
    - Tutorials
    - Pre-trained models

Machine learning on Redivis

- Computational notebooks in **Python**, R, Stata, and SAS
- Easily port between systems (e.g., import from Colab)
- The default, free notebook has 2 CPUs, 32GB RAM
- Can provision any other VM on Google Cloud, at-cost
  - Up to 16xA100 GPUs, 96 vCPUs, 1360GB RAM
- Do I need a GPU?
  - Just testing things out? Not necessarily! Try the free notebook.
  - Doing real work, at scale? Probably, it's just so much faster.

Machine learning walkthrough

Analysis on Redivis: summary

- On large datasets, always do as much preprocessing as possible via transforms.
- Make sure your code takes advantage of your hardware! Always check that CPU, memory, GPU are being full utilized.
- Customizable VMs allow for you to load, apply, and fine-tune machine-learning models.
- Leverage common languages and libraries to allow for code portability across systems.

Where does Redivis excel?

- Large datasets
- High-risk datasets
- Collaborative workflows
- Reproducibility
- Low barrier to entry, can do a lot for free

05

Q&A

06

Additional resources

Additional resources

Learn more

Redivis for researchers

Redivis for organizations

Additional resources

Support materials

Documentation

API documentation

Security overview