Data Version Control (DVC) New User Orientation

Jelai Wang

Aug 19, 2020

UAB CENTER FOR CLINICAL AND TRANSLATIONAL SCIENCE
Knowledge that will change your world

INFORMATICS INSTITUTE

Outline

Slides

- What problem are we trying to solve?
- What are possible solutions?
- Challenges with DVC as a solution.
- Early proof-of-concept work.
- Current implementation.

Hands-on Lab

- See handout.

What problem are we trying to solve?

Context #1

Because we have strong, local HPC (Cheaha), much of our applied data analysis is performed there.

Naturally, we choose to represent our pipelines and data with abstractions that fit HPC.

Context #2

In particular:

- Pipelines are represented as scripts that can be run by SLURM.
- Data are represented and organized as ***flat files***.

Problem statement

We wish to make more specific guarantees about ***data provenance*** and ***reproducible research***, but are limited by not looking beyond the “files and folders” level of abstraction.

Data Provenance

siis.cse.psu.edu/provenance.html

Systems and Internet Infrastructure Security

Information & Resources

- Home
- News
- People
- Grants
- Publications
- Tools
- SIIS wiki
- Certificate Program

Research

- Overview
- Authorization Hook Placement
- Cloud Verifier
- Control Systems and Smart Grid
- Data Provenance
- Interdomain Routing
- Name Resolution Security
- Smartphone
- Application Analysis
- CRA Research

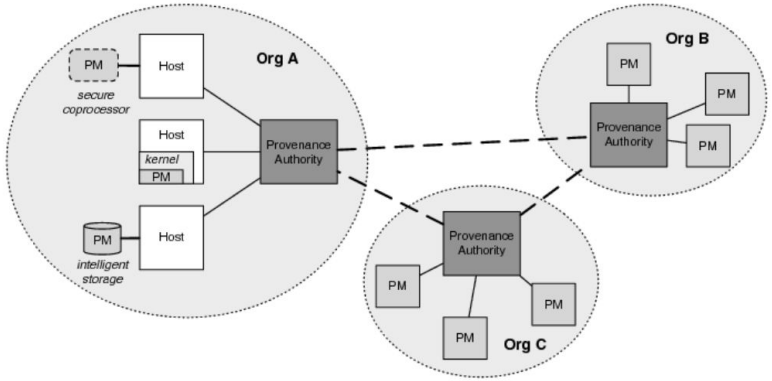
Past Projects

- Access Control
- Attribute Systems
- Hardware Security
- Mobile Phones
- Secure Languages
- Storage Security
- Telecommunications
- Virtual Machines

Data Provenance

Data provenance documents the inputs, entities, systems, and processes that influence data of interest, in effect providing a historical record of the data and its origins. The generated evidence supports essential forensic activities such as data-dependency analysis, error/compromise detection and recovery, and auditing and compliance analysis.

This collaborative project is focused on theory and systems supporting practical end-to-end provenance in high-end computing systems. Here, systems are investigated where provenance authorities accept host-level provenance data from validated provenance monitors, to assemble a trustworthy provenance record. Provenance monitors externally observe systems or applications and securely record the evolution of data they manipulate. The provenance record is shared across the distributed environment.



In support of this vision, tools and systems are explored that identify policy (what provenance data to record), trusted authorities (which entities may assert provenance information), and

"Data provenance documents the inputs, entities, systems, and processes that influence data of interest, in effect providing a historical record of the data and its origins."

Excerpt from http://siis.cse.psu.edu/provenance.html.

What are possible solutions?

Possible Solution #1

Write Custom Software

Pros

- We (might) get exactly what we want.

Cons

- Large up-front investment in development cost (people and time).

Possible Solution #2

Buy Off-the-shelf Software

Pros

- Someone else wrote what we want. Maybe.

Questions

- *How much does it cost?*
- *How much time do we need to invest to customize to meet our needs?*

Example: Univ of Rochester, LabKey, BLISS, team of maintainers

Possible Solution #3

Adapt Existing Software to Address Biggest Needs

Questions

- *Can we adapt specialized software, like **version control** and **wikis** to directly address our biggest needs?*

Case Study

- SVN + Confluence wiki, 2008 - present, UAB SSG
  - Microarray to GWAS to NGS*.
  - Dozens of data analysis projects.

Challenges with DVC

Data Version Control (DVC) Challenges

Challenge 1

Analyst Workflow

Can we *instrument* existing analyst workflow with DVC such that it is at least as efficient?

How can DVC improve analyst efficiency and effectiveness?

Challenge 2

Data are Big

Can we even store NGS data analyses, which can be TB in size, in VC?

Will it be too cumbersome to interact with large files in VC?

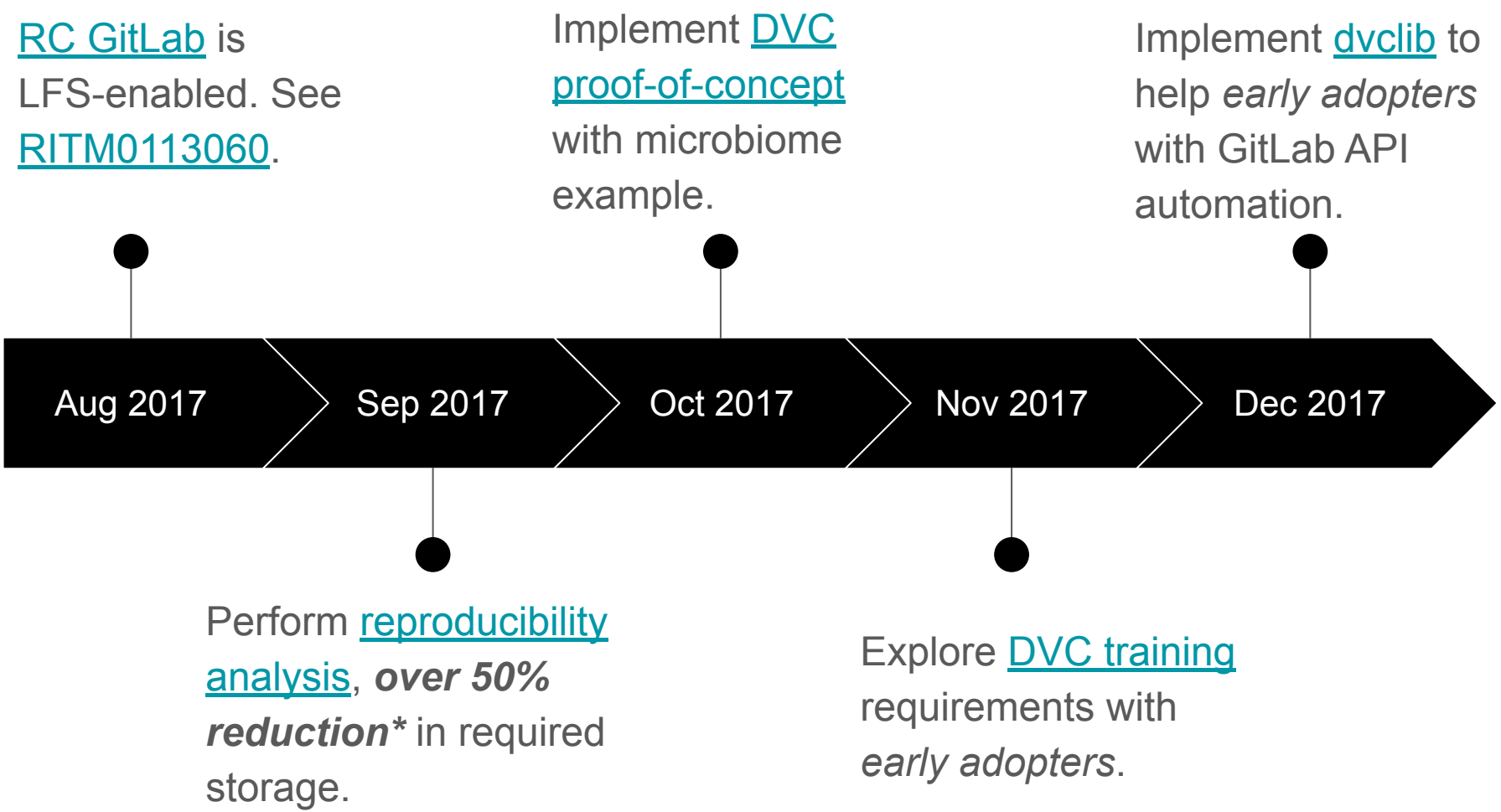
Challenge 3

DVC Training

Using DVC requires knowledge of and practical skills in modern VC tools and practices.

What are the associated training resources and costs?

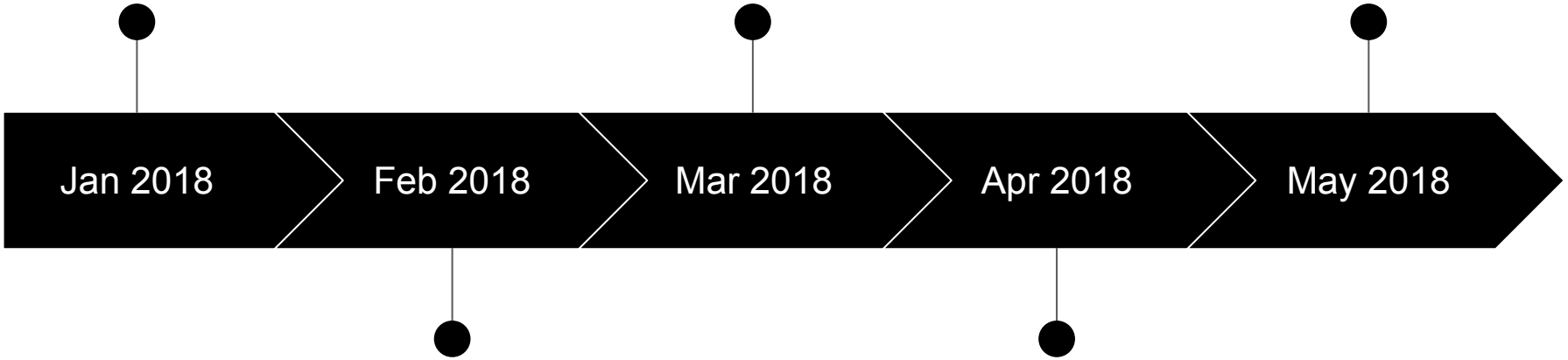
Early Proof-of-Concept Work


RC GitLab is LFS-enabled. See RITM0113060.
Implement DVC proof-of-concept with microbiome example.
Implement dvclib to help early adopters with GitLab API automation.
Aug 2017
Sep 2017
Oct 2017
Nov 2017
Dec 2017
Perform reproducibility analysis, over 50% reduction* in required storage.
Explore DVC training requirements with early adopters.

Research SOP and convention, as exemplar for *early adopters*.

Establish Confluence for DVC.

Release dvctools 0.2 as "*guaranteed DVC stack*".



Produce YouTube training videos for *early adopters*. See DVC demo and dvclib demo.

Write Quick Start documentation for helping DVC newcomers get started.

62

investigators

119

data analysis projects

CCTS Microbiome Case Study

May 2018 - May 2019

Project Summary Statistics

- Max Size = **102 GB**
- Avg Size = **6.2 GB**
- Max Num Files = **17,574**
- Avg Num Files = **1,847**

Liam

Dongquan

Current Implementation

A distributed version control system (DVCS).

Features

- Huge, existing ecosystem
- Widely-adopted, well-documented
- Well-understood data model and concepts
- Mature client tooling and processes

Benefits

- Increases network effect
- Reduces startup cost
- Reduces training cost

GitLab

An open-source git repo hosting solution.

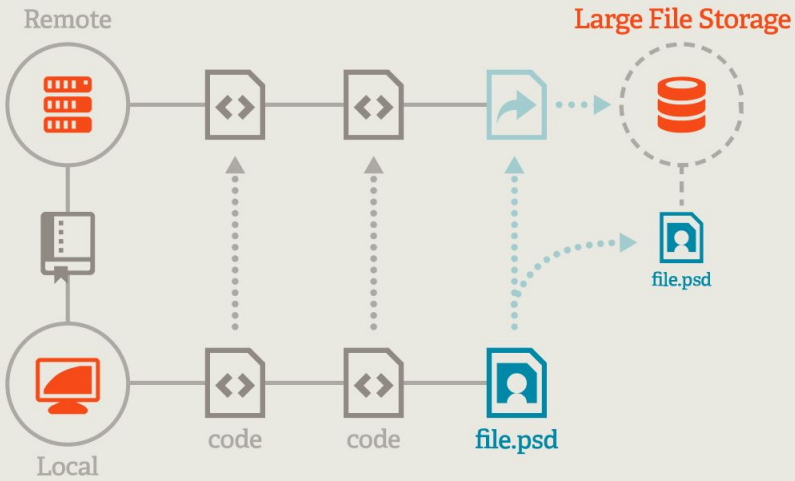
Features

- Integrated issue tracking, wiki, CI/CD, and more.
- Modern authn/authz options.
- Nice web-based user interface.
- Nice REST API, dev SDK.

DVC early adopter prototyping on GitLab instance at https://gitlab.rc.uab.edu hosted by UAB IT's Research Computing.

Git Large File Storage

An open source Git extension for versioning large files.



DVC prototype uses 50 TB **Git LFS** backing store on high-performance, parallel GPFS storage.

Confluence

A modern WYSIWYG wiki.

Notes

- Pros: It's a very good wiki.
- Cons: It's not free.

DVC prototype for CCTS Informatics use cases pairs git with a dedicated **Confluence** VM for *data analysis documentation* hosted by UAB IT at https://wiki.genome.uab.edu.

dvctools

A guaranteed DVC stack via container technology.

Uses **GitLab API** to extend functionality and automate tedious, multi-step processes to improve reproducibility.

Version 1.4 simg for Cheaha

- Git 1.8.3.1
- Git LFS 2.10.0
- Python 3
- python-gitlab 1.6.0
- dvclib

DVC Training and Documentation

DVC Orientation (8 hours)

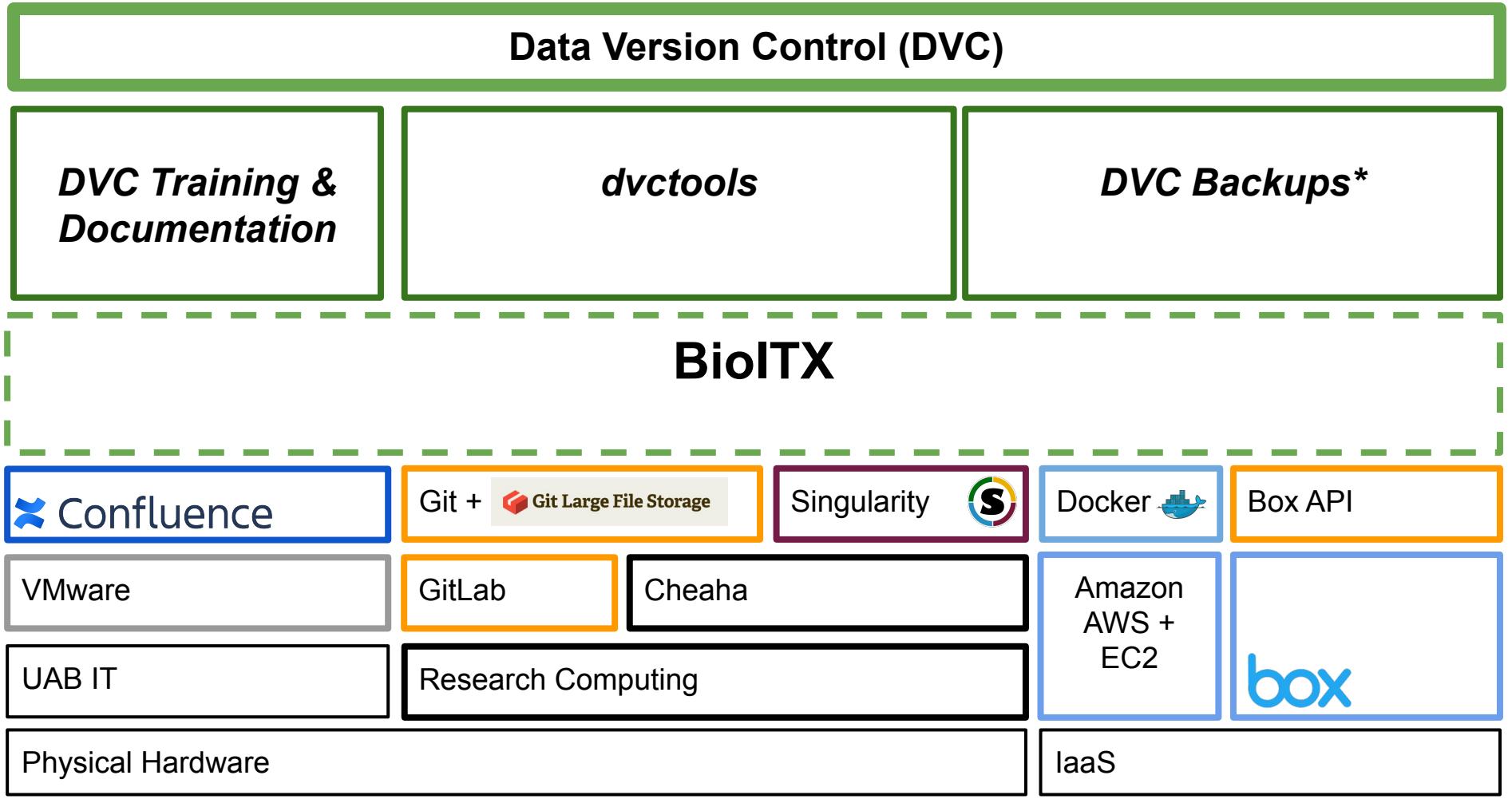
- Version Control Concepts
- `git` fundamentals
  - `git add/commit/status/log/diff`
  - `git branch/checkout/merge`
  - `git push/pull/remote`
- Common `git` workflows
  - Centralized, Feature Branch, GitFlow, Forking
- Git LFS specifics
  - `git lfs track/ls-files/status/clone`
- Examples of Group/Subgroup/Project conventions and SOP.

DVC Documentation

- Quick Start wiki page in Data Version Control space on Confluence instance at https://wiki.genome.uab.edu/x/L4Ae.
- DVC SOP Demo for standard CCTS Informatics microbiome pipeline on Cheaha at https://wiki.genome.uab.edu/x/PZ0e.

Links

- Early adopter DVC training GitLab issue

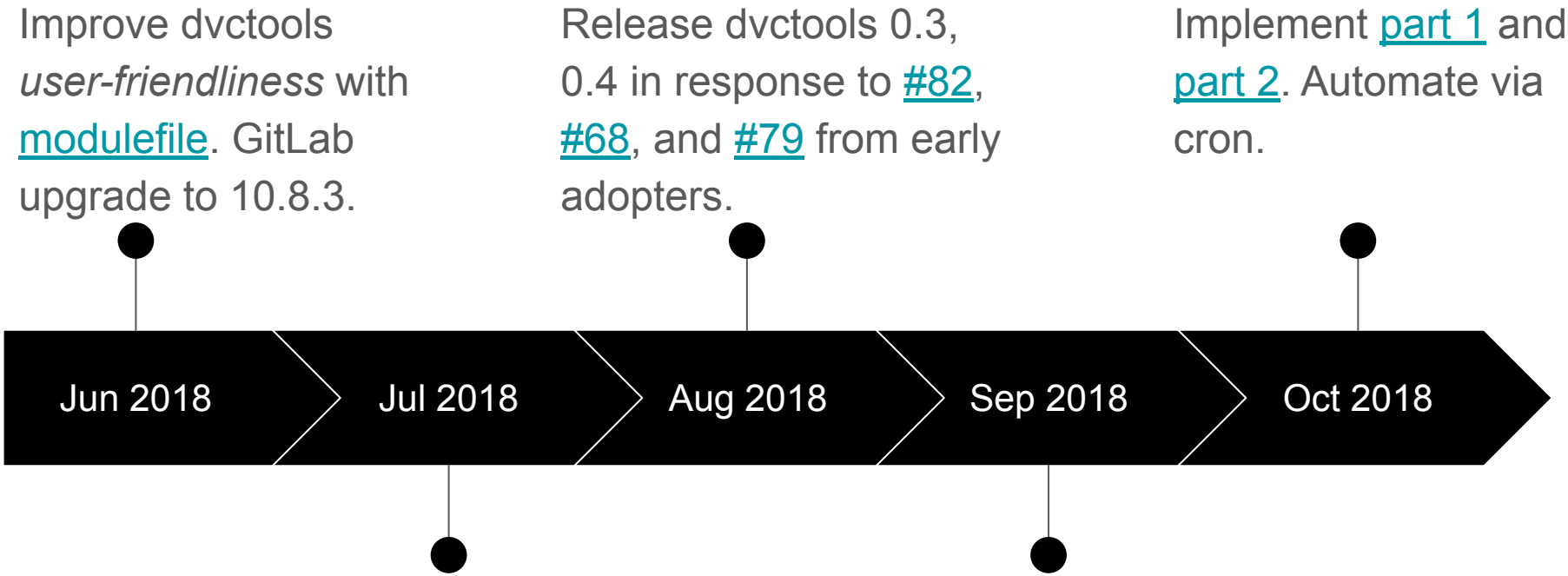

Data Version Control (DVC)
DVC Training & Documentation
dvctools
DVC Backups*
BioITX
Confluence
Git +
Git Large File Storage
Singularity
Docker
Box API
VMware
GitLab
Cheaha
Amazon AWS + EC2
box
UAB IT
Research Computing
Physical Hardware
IaaS

Early Adopter Case Study
CCTS Informatics

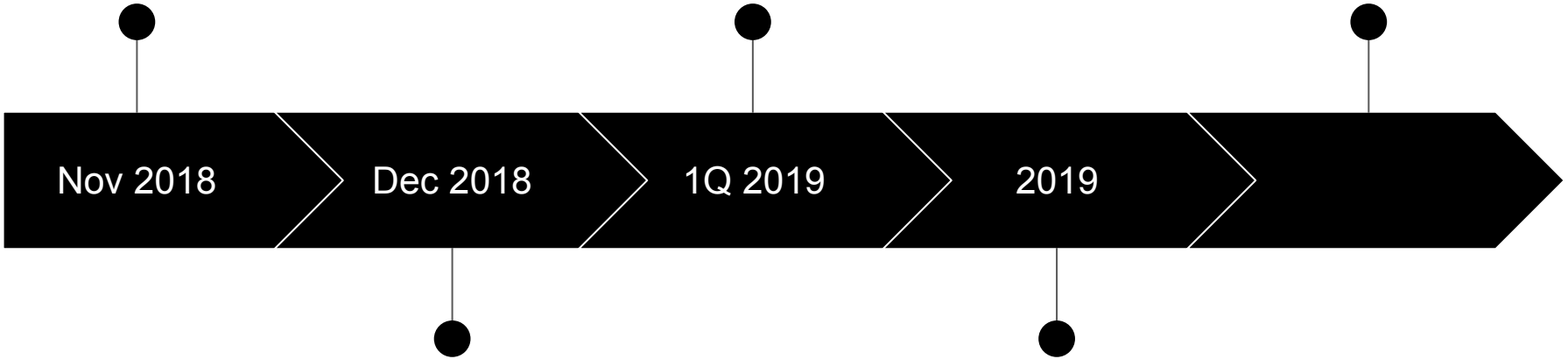
*An automated backup contingency.**

Desired Features

- Research Computing handles high availability of GitLab and Git LFS backing store.
- Research Computing handles primary backup.
- Per-git repo *automated backup contingency* to UAB **Box** cloud storage via **Globus** Box connector.

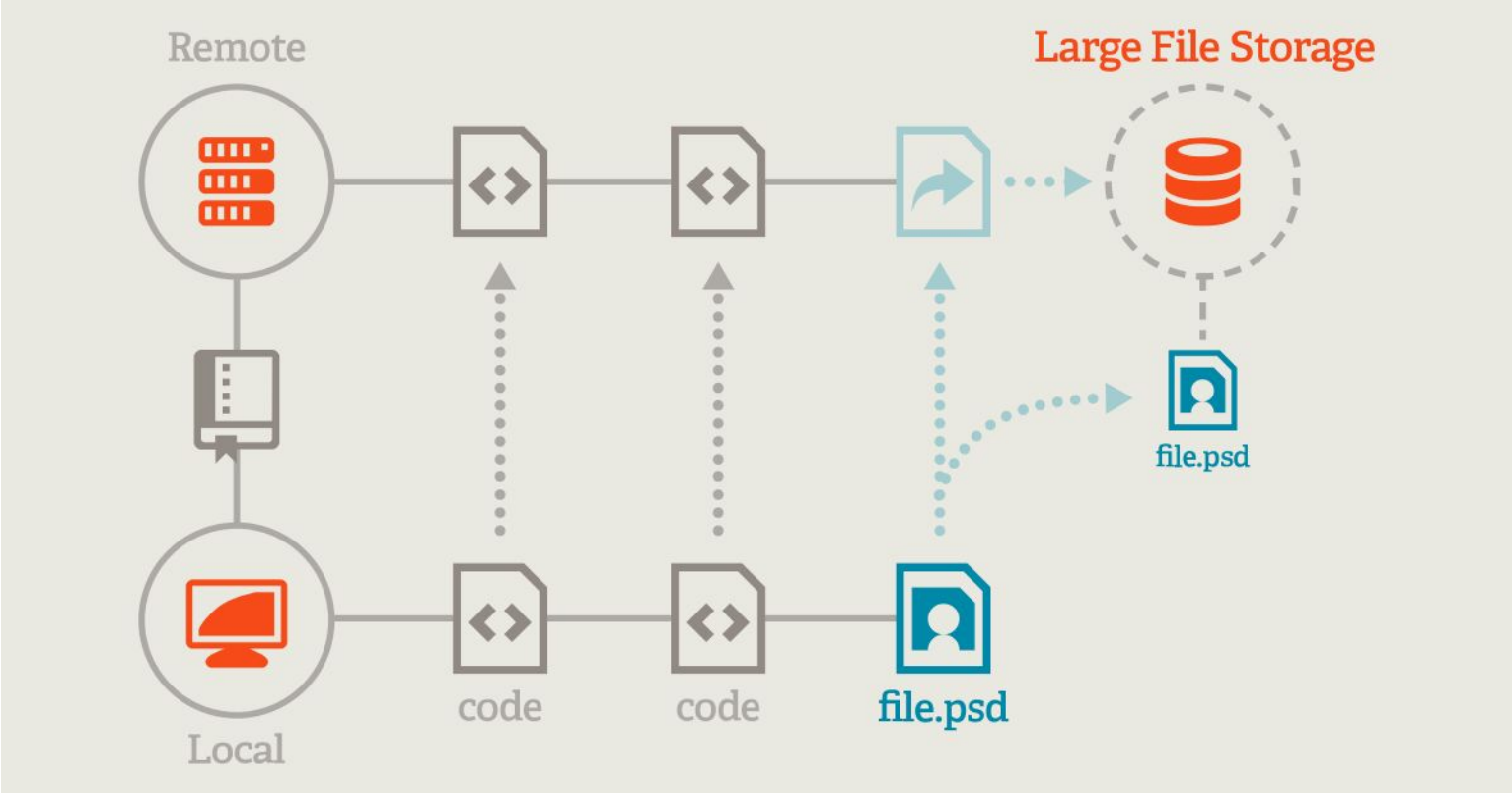

Improve dvctools *user-friendliness* with modulefile. GitLab upgrade to 10.8.3.
Release dvctools 0.3, 0.4 in response to #82, #68, and #79 from early adopters.
Implement part 1 and part 2. Automate via cron.
Jun 2018
Jul 2018
Aug 2018
Sep 2018
Oct 2018
Research **Box API**, especially SHA1 checksum, OAuth2, and OAuth2 + JWT.
Draft DVC backups contingency plan. GitLab upgrade to 11.2.3.

Release dvctools version 0.7 to address giant git lfs push error.



CCTS-Microbiome, CCTS Informatics Pipelines GitLab Group and DVC wiki documentation

Tour


Remote
Large File Storage
file.psd
code
code
file.psd
Local