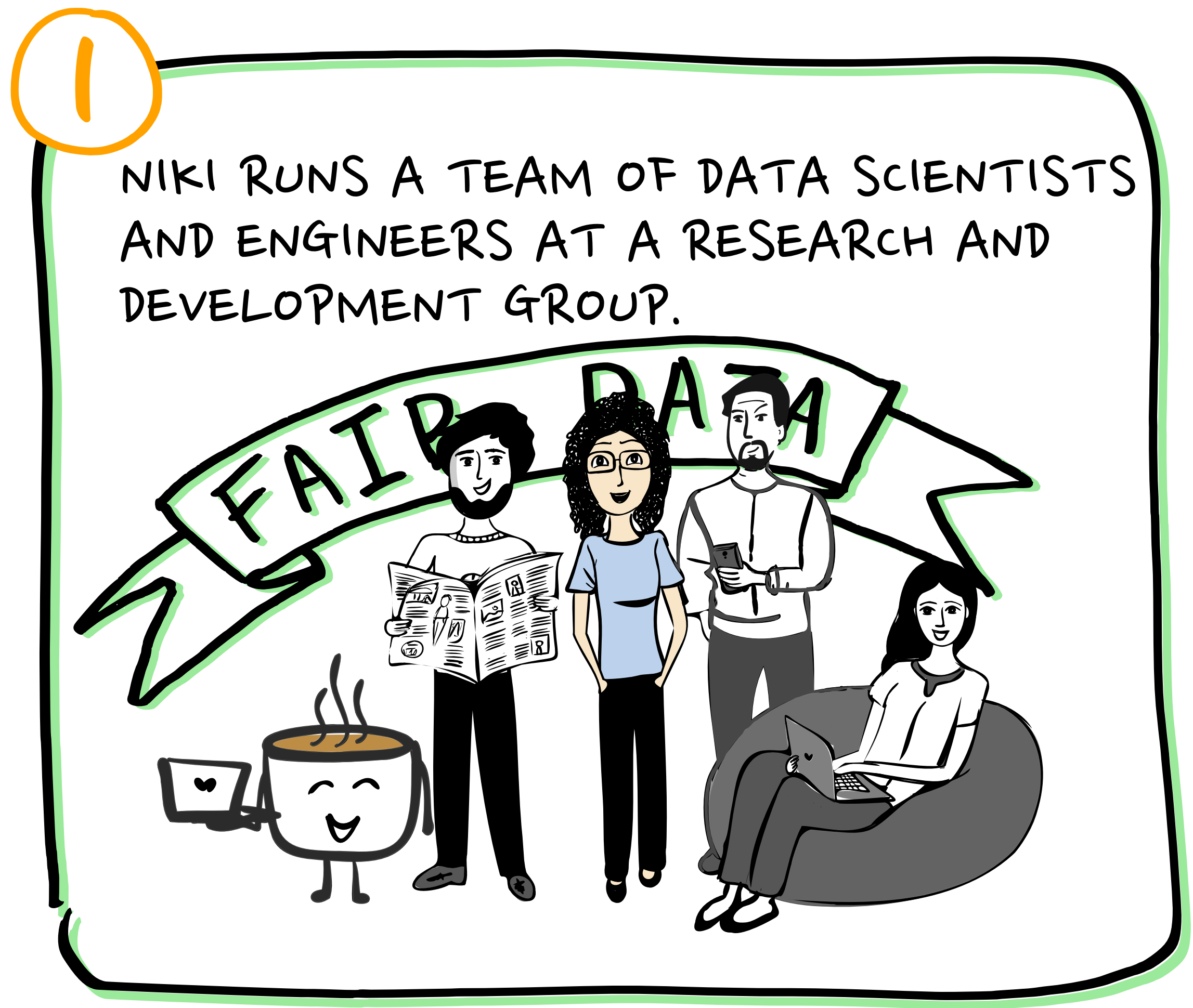
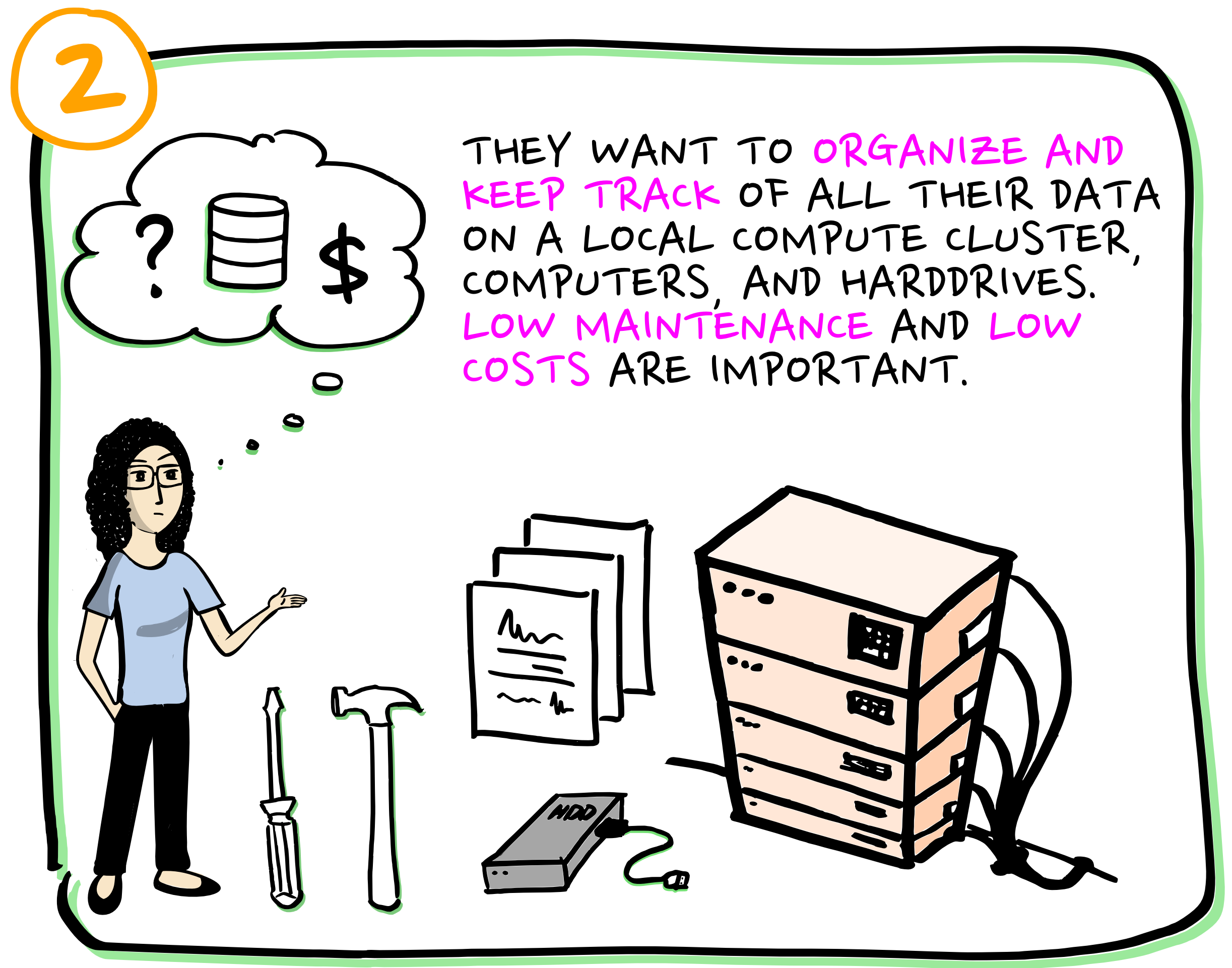
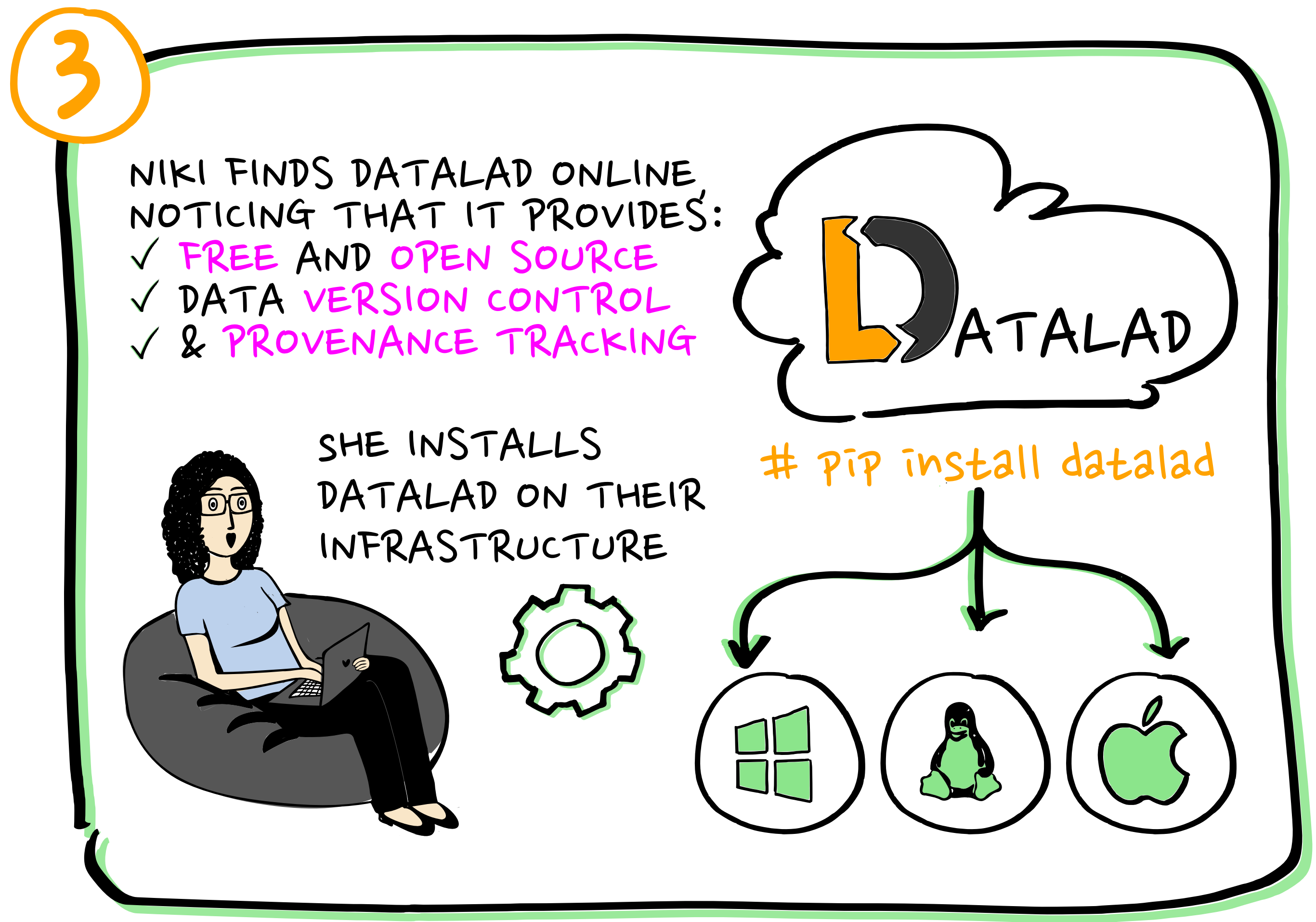
F.A.I.R. DATA MANAGEMENT WITH DATALAD



NIKI RUNS A TEAM OF DATA SCIENTISTS AND ENGINEERS AT A RESEARCH AND DEVELOPMENT GROUP.



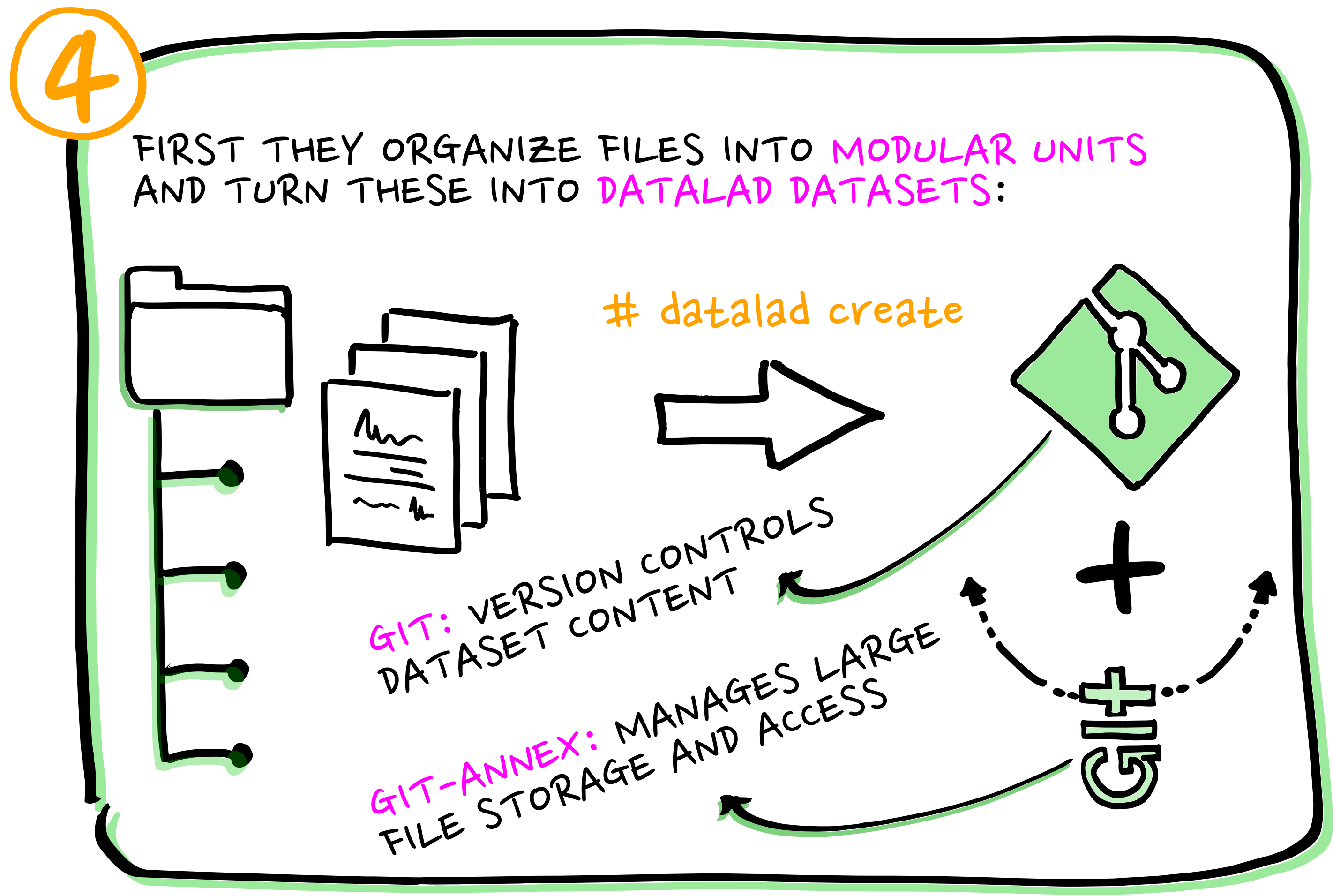
THEY WANT TO ORGANIZE AND KEEP TRACK OF ALL THEIR DATA ON A LOCAL COMPUTE CLUSTER, COMPUTERS, AND HARDDRIVES. LOW MAINTENANCE AND LOW COSTS ARE IMPORTANT.



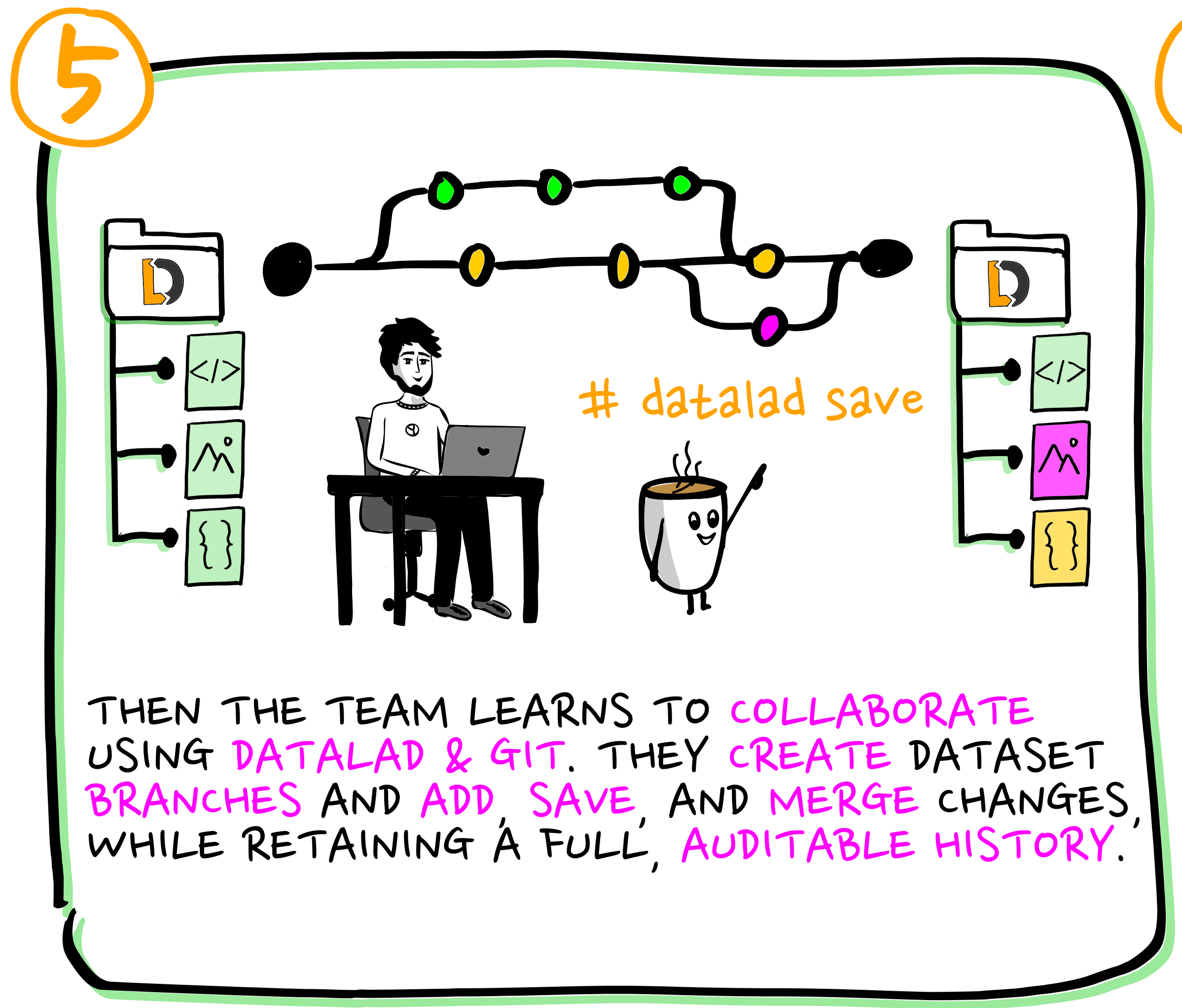
NIKI FINDS DATALAD ONLINE NOTICING THAT IT PROVIDES:

✓ FREE AND OPEN SOURCE
✓ DATA VERSION CONTROL
✓ & PROVENANCE TRACKING

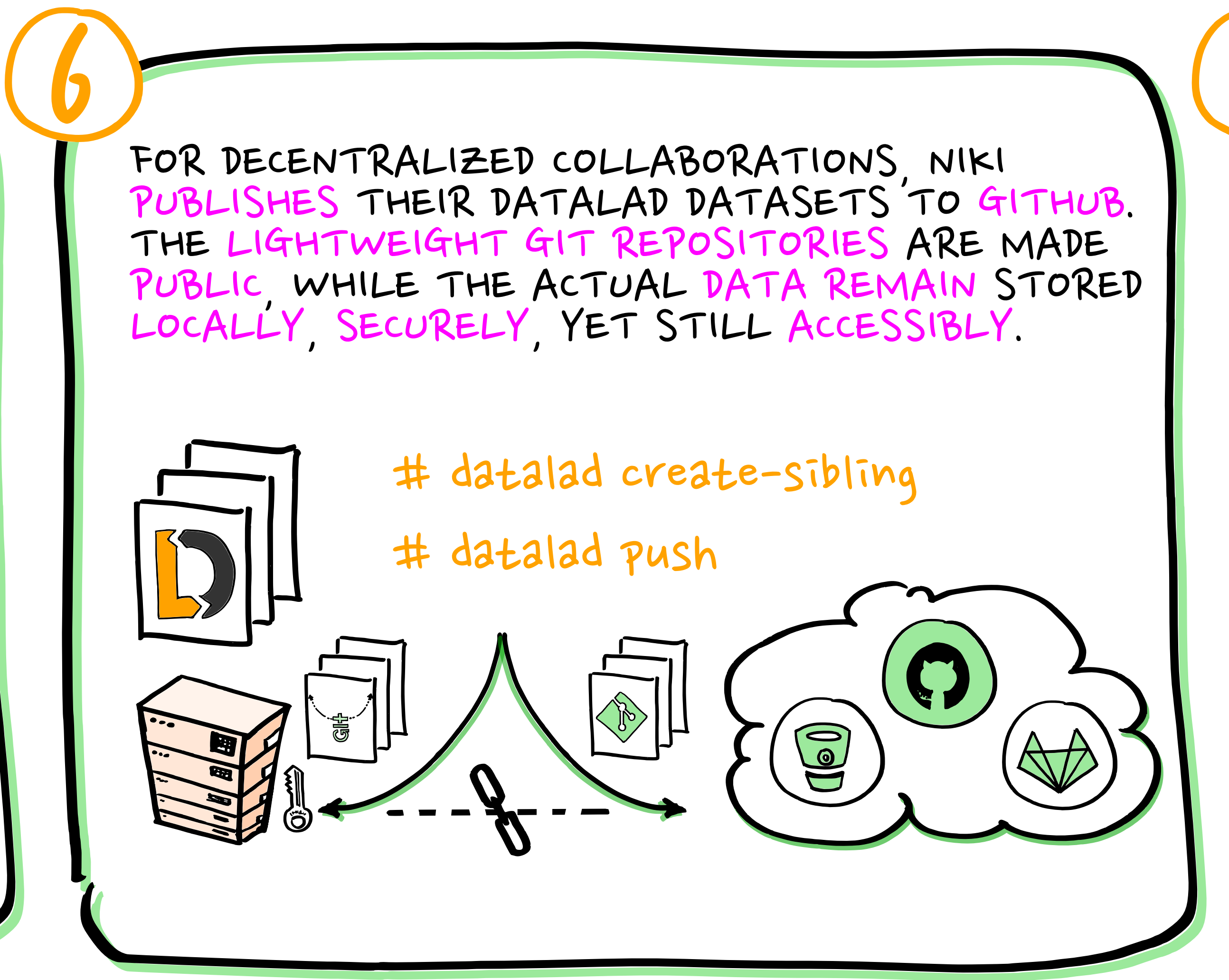
SHE INSTALLS DATALAD ON THEIR INFRASTRUCTURE


4
# datalad create
GIT: VERSION CONTROLS DATASET CONTENT
GIT-ANNEX: MANAGES LARGE FILE STORAGE AND ACCESS
+

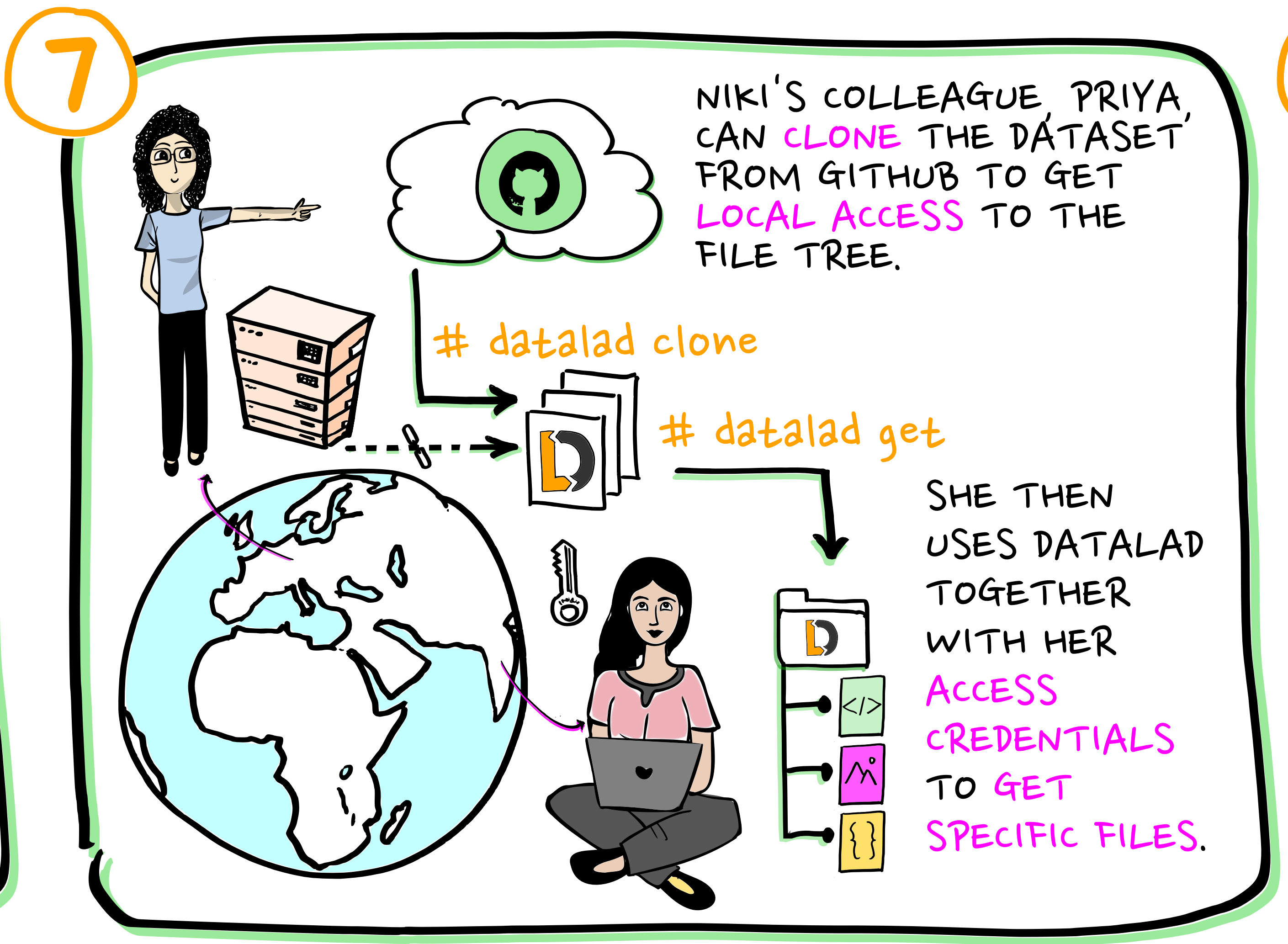
FIRST THEY ORGANIZE FILES INTO MODULAR UNITS AND TURN THESE INTO DATALAD DATASETS:



THEN THE TEAM LEARNS TO COLLABORATE USING DATALAD & GIT. THEY CREATE DATASET BRANCHES AND ADD, SAVE, AND MERGE CHANGES, WHILE RETAINING A FULL, AUDITABLE HISTORY.

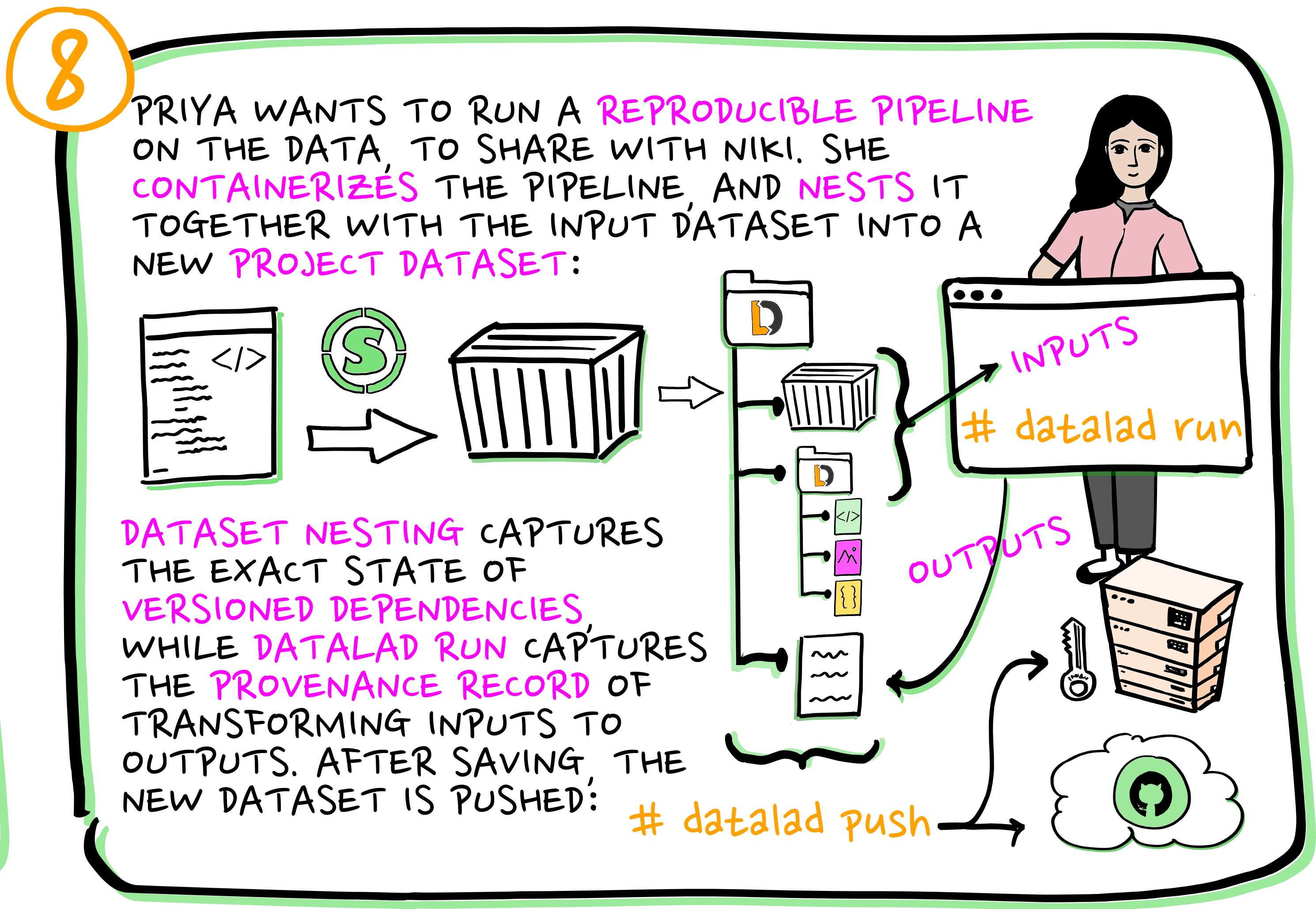


FOR DECENTRALIZED COLLABORATIONS, NIKI PUBLISHES THEIR DATALAD DATASETS TO GITHUB. THE LIGHTWEIGHT GIT REPOSITORIES ARE MADE PUBLIC, WHILE THE ACTUAL DATA REMAIN STORED LOCALLY, SECURELY, YET STILL ACCESSIBLY.



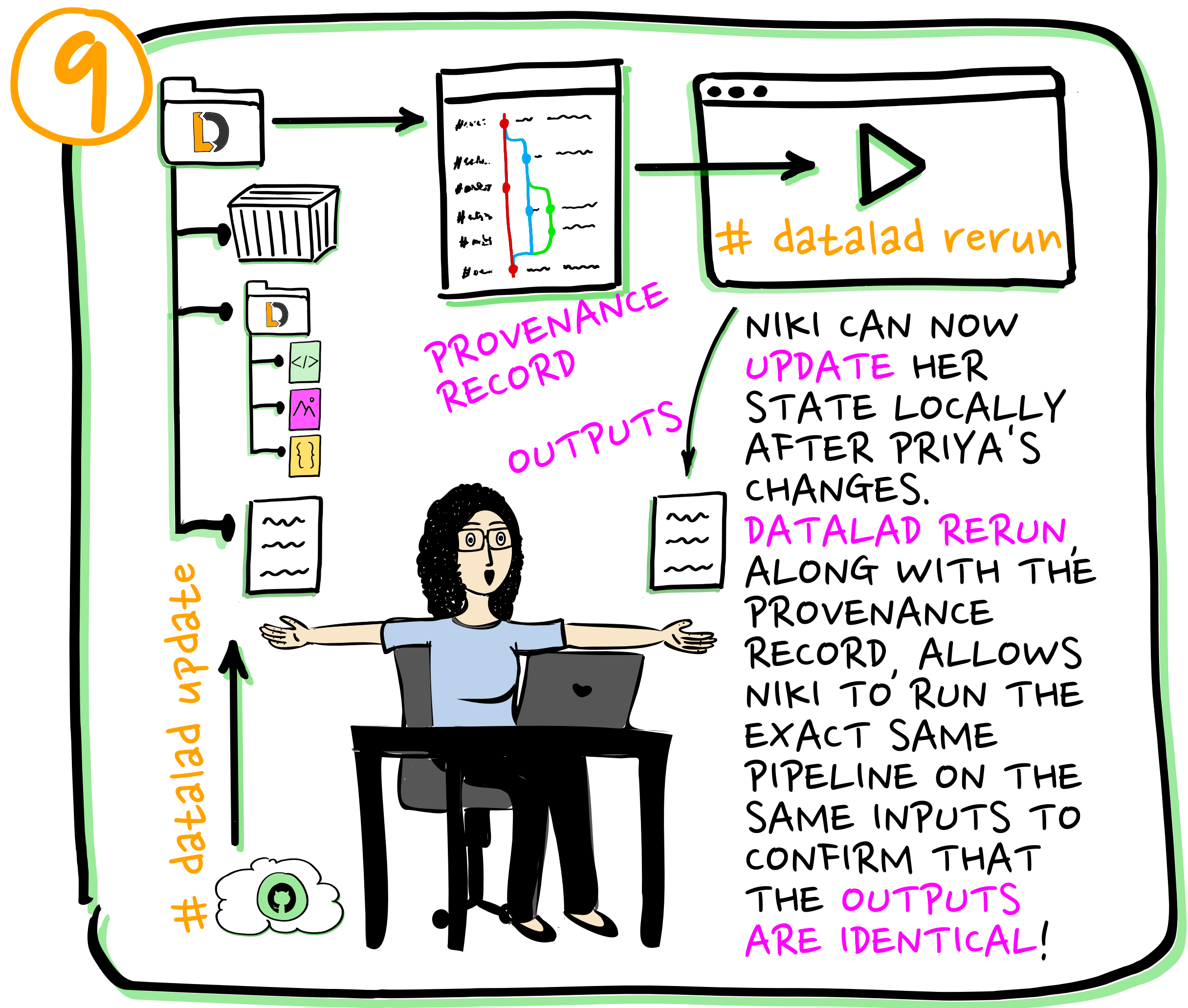
NIKI'S COLLEAGUE, PRIYA, CAN CLONE THE DATASET FROM GITHUB TO GET LOCAL ACCESS TO THE FILE TREE.

SHE THEN USES DATALAD TOGETHER WITH HER ACCESS CREDENTIALS TO GET SPECIFIC FILES.

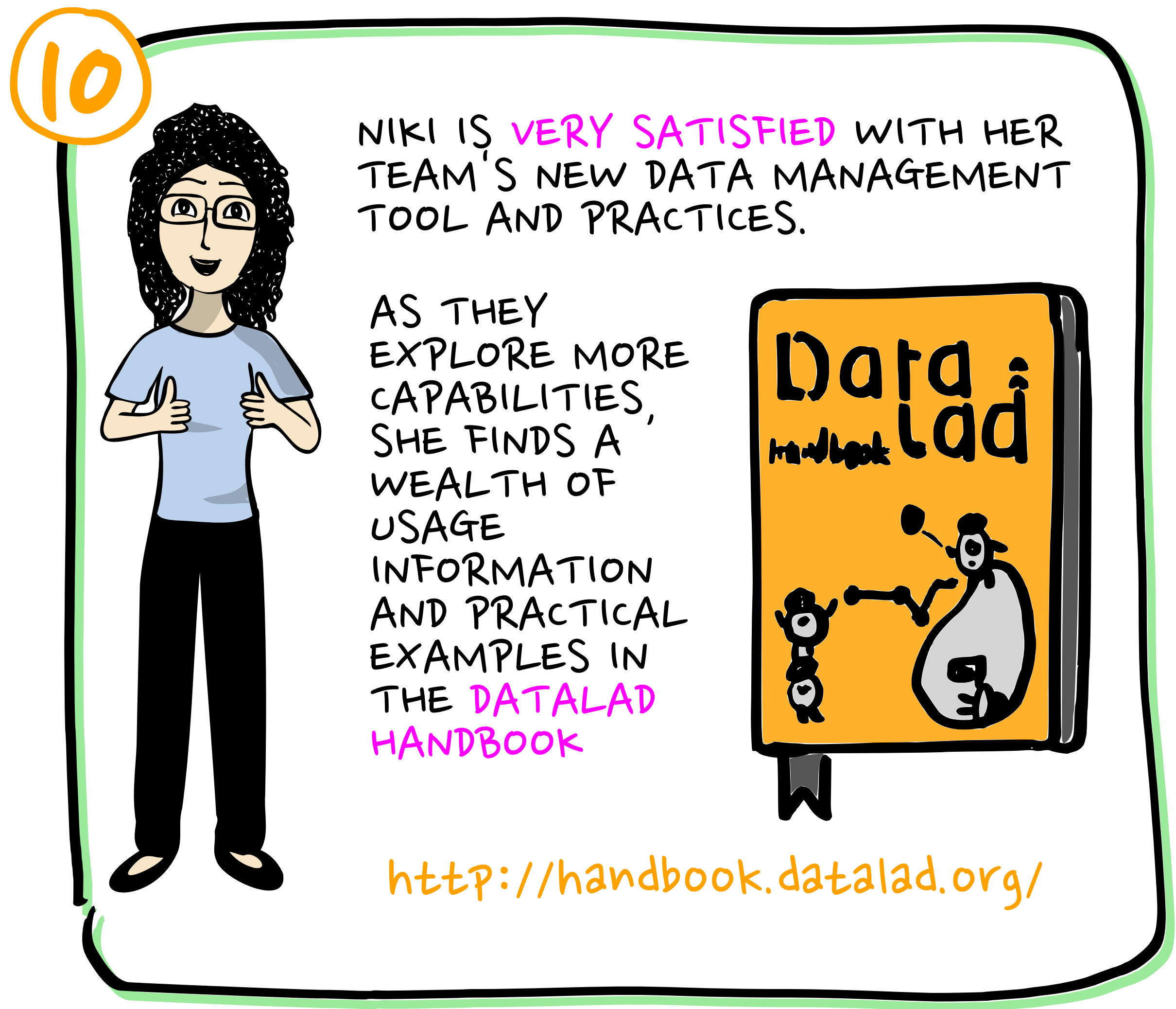


PRIYA WANTS TO RUN A REPRODUCIBLE PIPELINE ON THE DATA, TO SHARE WITH NIKI. SHE CONTAINERIZES THE PIPELINE, AND NESTS IT TOGETHER WITH THE INPUT DATASET INTO A NEW PROJECT DATASET:

DATASET NESTING CAPTURES THE EXACT STATE OF VERSIONED DEPENDENCIES WHILE DATALAD RUN CAPTURES THE PROVENANCE RECORD OF TRANSFORMING INPUTS TO OUTPUTS. AFTER SAVING, THE NEW DATASET IS PUSHED:



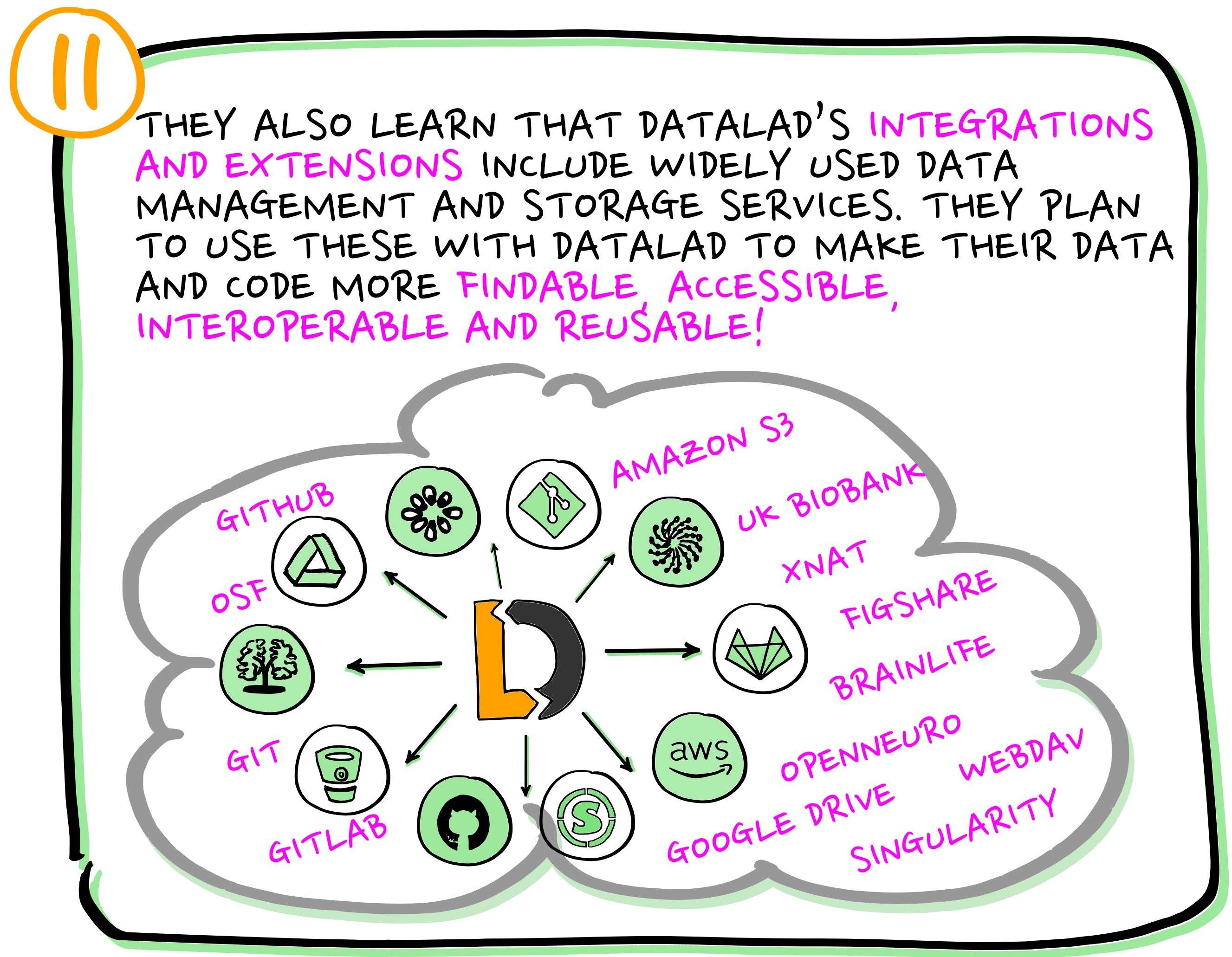
NIKI CAN NOW UPDATE HER STATE LOCALLY AFTER PRIYA'S CHANGES. DATALAD RERUN, ALONG WITH THE PROVENANCE RECORD, ALLOWS NIKI TO RUN THE EXACT SAME PIPELINE ON THE SAME INPUTS TO CONFIRM THAT THE OUTPUTS ARE IDENTICAL!



NIKI IS VERY SATISFIED WITH HER TEAM'S NEW DATA MANAGEMENT TOOL AND PRACTICES.

AS THEY EXPLORE MORE CAPABILITIES, SHE FINDS A WEALTH OF USAGE INFORMATION AND PRACTICAL EXAMPLES IN THE DATALAD HANDBOOK

http://handbook.datalad.org/



THEY ALSO LEARN THAT DATALAD'S INTEGRATIONS AND EXTENSIONS INCLUDE WIDELY USED DATA MANAGEMENT AND STORAGE SERVICES. THEY PLAN TO USE THESE WITH DATALAD TO MAKE THEIR DATA AND CODE MORE FINDABLE, ACCESSIBLE, INTEROPERABLE AND REUSABLE!

12

DATALAD RESOURCES

datalad.org

github.com/datalad

info@datalad.org

youtube.com/datalad

twitter.com/datalad

10.5281/zenodo.6400523