Using Containers to Valid
Research on Confidential D
Scale

Lars Vilhuber

May 2024

Introduction

Concerns

Concerns about confidentiality in statistical products ha
in the past several years:

- New disclosure avoidance techniques in the Decennia
garnered much attention (an understatement...),
- also concerns about formal disclosure avoidance tech
public-use microdata files (PUMFs) (see Census Burea
implementation of such methods for the American Co
Survey)

Creating synthetic data

- Much effort put into creating privacy-protected or syn
(this conference!)
- Goal of each of these: **release and forget**

But what if users don´t trust the data?

Direct access

- Many different RDC-style systems have been stood u
past 35+ years in multiple countries
- Provide direct access to confidential (pseudonymous)
- Still need output disclosure avoidance measures (mo
- Expensive for stat agencies to maintain, expensive fo

Alternative: validate analyses run on synthetic a
against the confidential data

Validation and Verification

- long-running pilot projects with (non-formal) syntheti
products (SynLBD, SIPP Synthetic Beta) came to an er
(End of life for the Cornell Synthetic Data Server Septe
2022)
- not sure how active OPM Verification server was (Bar
2018)
- scale an issue
- not done or planned for most synthetic data product

Scaling up

These pilot projects were not set up to scale, and
demonstrated that there is a need for such a pr

Background

Anecdotal evidence with SD

From conversations/informal surveys:

- researchers were happy with the ability to access dat
  having to request a full-blown project in an FSRDC
- somewhat frustrated by the process (slowness)

Reproducibility and SDS

SDS validation required typically substantial human

- Reason: problems with the reproducibility of code in t
sciences despite similarity of environment.

Intermediate causes

- no strong pre-testing of reproducibility, often intense interactive programming practices
- divergence in environments over time
- divergence of data schemas over time

Failure to maintain strong links

Some broader evidence

In a sample of over **8,000 replication packages** asso
high-profile economics articles, **only 30% had some sc
script**.

Other systems

Statistical agencies and research institutes have explo
ways to scale up access to confidential data, without f
access to confidential data.

- Statistics Canada: Real Time Remote Access (RTRA
- Norway: Microdata.no system,
- Germany/IAB: JoSuA system

Access restrictions

Most such processes have limitations, including in the general purpose analysis

Most still have some **strong access limitatio**

- RTRA: organizational application process
- microdata.no: Institutional MOU (and only Norwegia
- IAB: proposal process

Analysis restrictions

Many systems strongly **limit the type of analysis** that

- RTRA: restricting the software keywords that can be u
  of SAS allowing to “calculate frequencies, means, perc
  percent distribution, proportions, ratios and shares.
- microdata.no: by creating a structured new statistical
  (albeit with increasingly sophisticated capabilities)

Comparison

The comparison researchers and analysts make is (f
wrong) to the **unfettered use of public-use data** that

The quest

Direct access is expensive

Remote-access or local secure access in the form of virtual secure data enclaves is still the dominant - **but** way to access confidential data.

*The dominant method of access thus forces researchers to choose between **lower quality data** in an environment that corresponds to th preferred computing method (public-use data and **higher quality confidential data** in environments that are expensive for researche data providers, or both.*

Possible solution

Containers

Containers are lightweight, standalone, executable pa
contain everything needed to run an application, includ
a runtime, libraries, environment variables, and cor

Containerized validation

- **Containers**,
  - hosted on public cloud platform or run on researc
  - provide access to synthetic or “plausible” data, anc resources
  - mechanism to ensure authors can validate reprod analysis
- Then submitted to the **confidential computing envi**
  - analysis modified to use confidential data
  - enables a wide spectrum of plug-in disclosure avoi measures as well
    - similar in spirit: IAB JoSuA system, but without hos

Containers in the wild

One of the first mentions of containers for scientific r
Boettiger (2015).

- CodeOcean is a commercial service facilitating that pr
making the resources available through a web brows
- Wholetale and MyBinder are other (academically orie
services that provide similar functionality[2]
- Many universities HPC clusters provide some support
more popular than `Docker`)

Containers in Social Sciences are
challenge!

In a sample of over **8,000 replication packages** assoc
high-profile economics articles, **only 11 had a `Docker`**
build script for containers).

(That's n=11, not 11% - in fact, it's **0.13%** of replicatior

What's new

The use of containers in this way is novel as a system
provide **scalable, potentially high-throughput vali**
differs in usage from previous methods, such as th
Synthetic Data Server.

*I believe that it is promising as a modern way
implementing validation when data are
confidential.*

User perspective

Use provided container with pre-pro
data

Possibilities:

- use directly (*safer*)
- use as input to build own container (addition of co

Critically

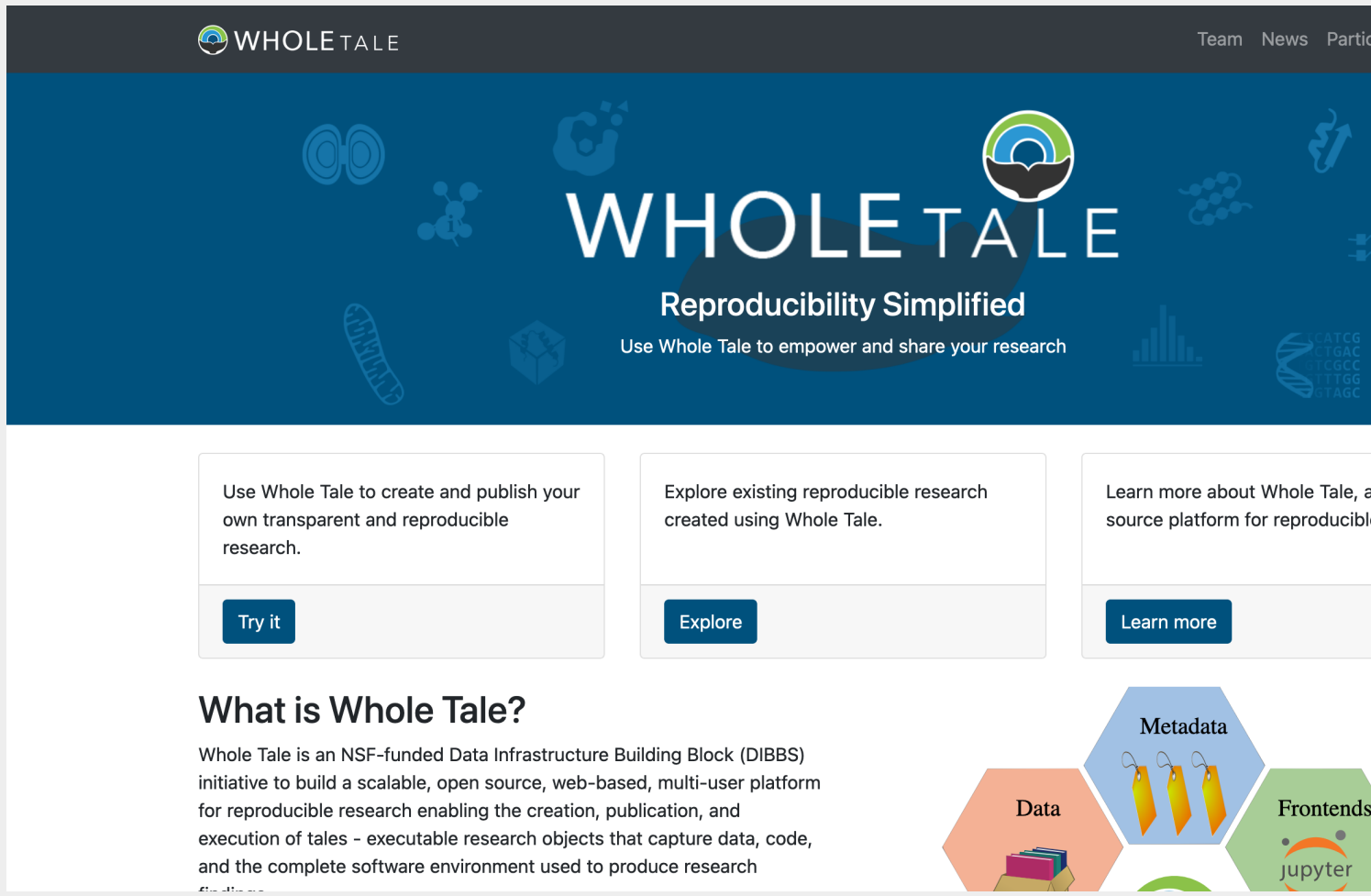
Pre-provisioned data does not need to be “**analytically**
only be “**plausible**”!

Develop where feasible

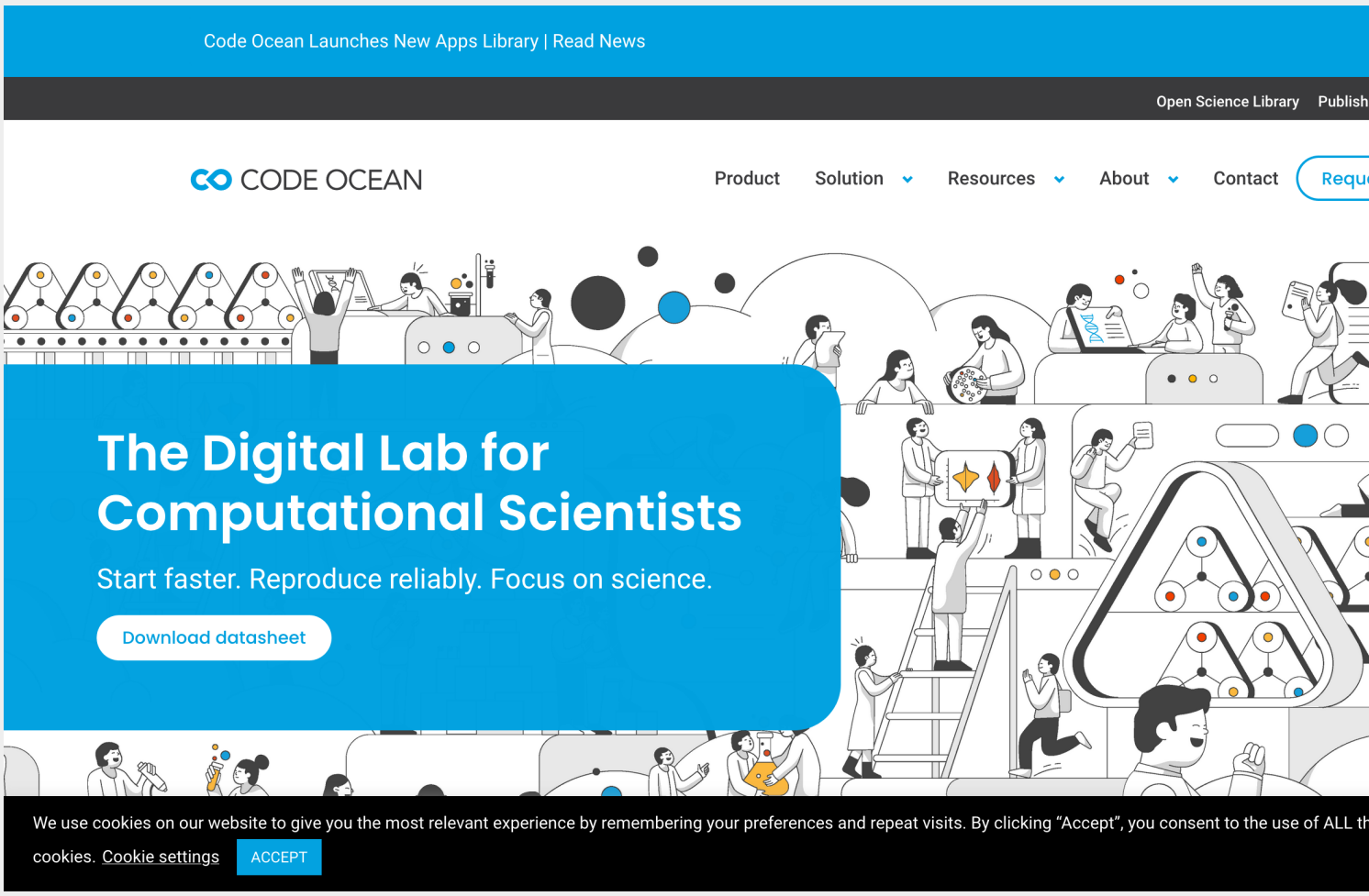
Containers are generalized technology

- can be run on provisioned university computing infra
(most HPC systems can run containers)
- can run on desktops as needed (free container softwa
for all major operating systems for non-commercial u
- can run on generic cloud infrastructure (AWS, Google
Azure)
- can run on custom cloud infrastructure specialized in
containers (Nuvolos, Codeocean, Onyxia, etc.)
- can be prepared by research institutions for use on th
infrastructure (e.g., NSF-funded Whole Tale project, se
Onyxia)

Whole Tale


WHOLE TALE
Team
News
WHOLE TALE
Reproducibility Simplified
Use Whole Tale to empower and share your research
Use Whole Tale to create and publish your own transparent and reproducible research.
Try it
Explore existing reproducible research created using Whole Tale.
Explore
Learn more about Whole Tale, a
source platform for reproducibl
Learn more
What is Whole Tale?
Whole Tale is an NSF-funded Data Infrastructure Building Block (DIBBS) initiative to build a scalable, open source, web-based, multi-user platform for reproducible research enabling the creation, publication, and execution of tales - executable research objects that capture data, code, and the complete software environment used to produce research
Metadata
Data
Frontends
jupyter

Codeocean


Code Ocean Launches New Apps Library | Read News
Open Science Library
Publish
CODE OCEAN
Product
Solution
Resources
About
Contact
Requ
The Digital Lab for Computational Scientists
Start faster. Reproduce reliably. Focus on science.
Download datasheet
We use cookies on our website to give you the most relevant experience by remembering your preferences and repeat visits. By clicking "Accept", you consent to the use of ALL th
cookies. Cookie settings
ACCEPT

Onyxia

Onyxia

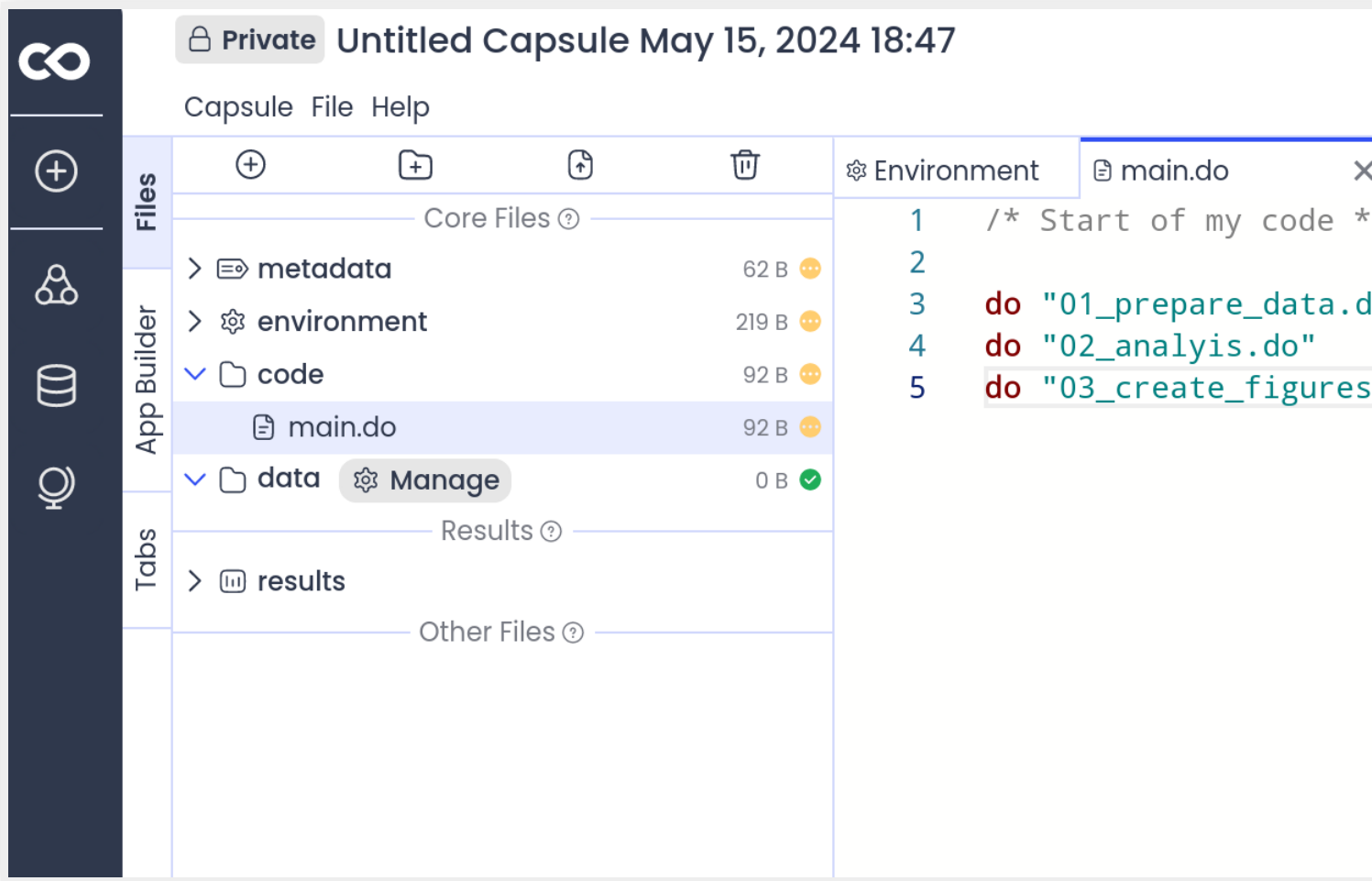
Cost to user

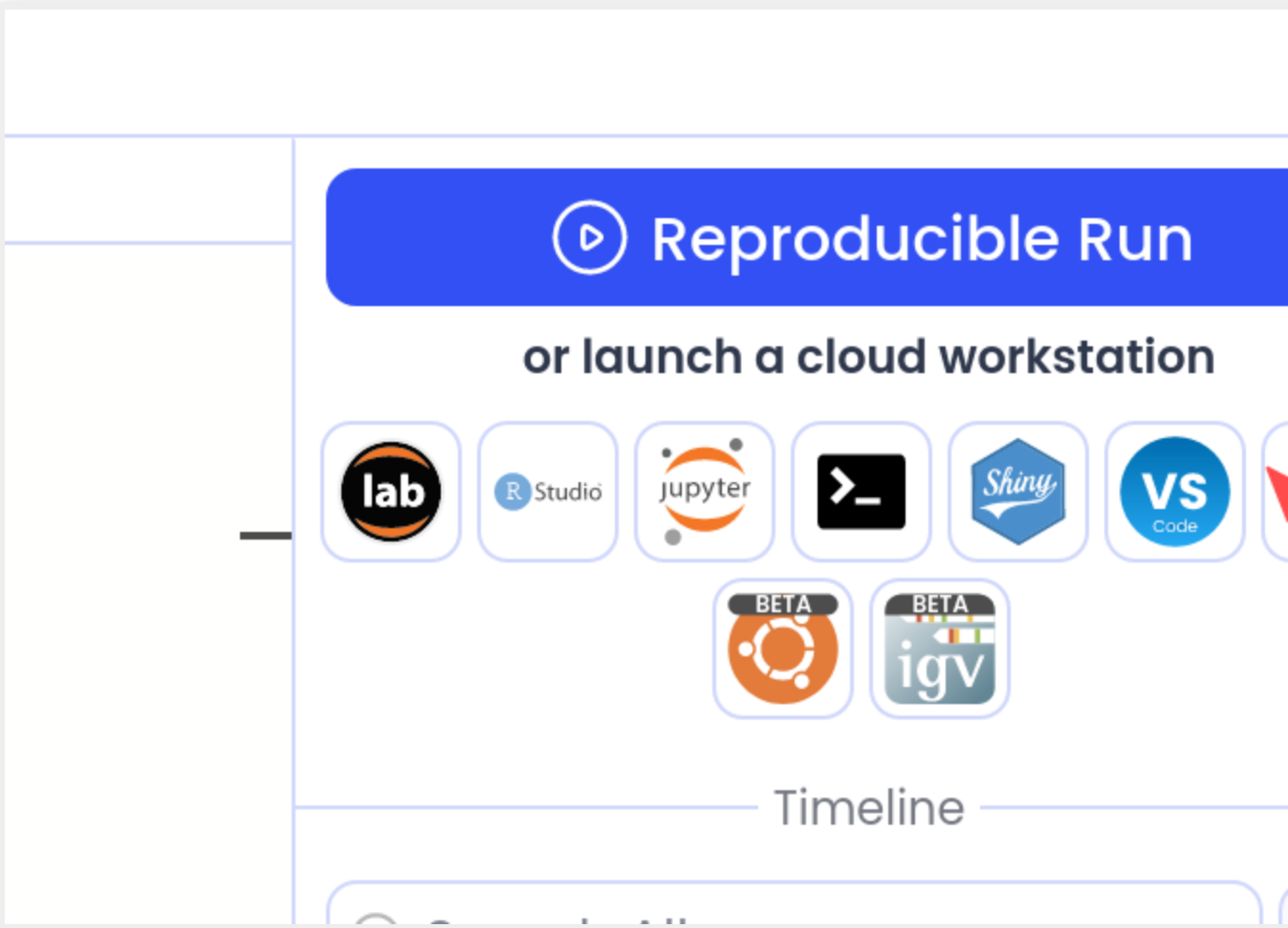
Cost: $0 to low $

Run a container from the comma

```
## read the file run_docker.sh
tail(readLines("run_docker.sh"),n=1)
```

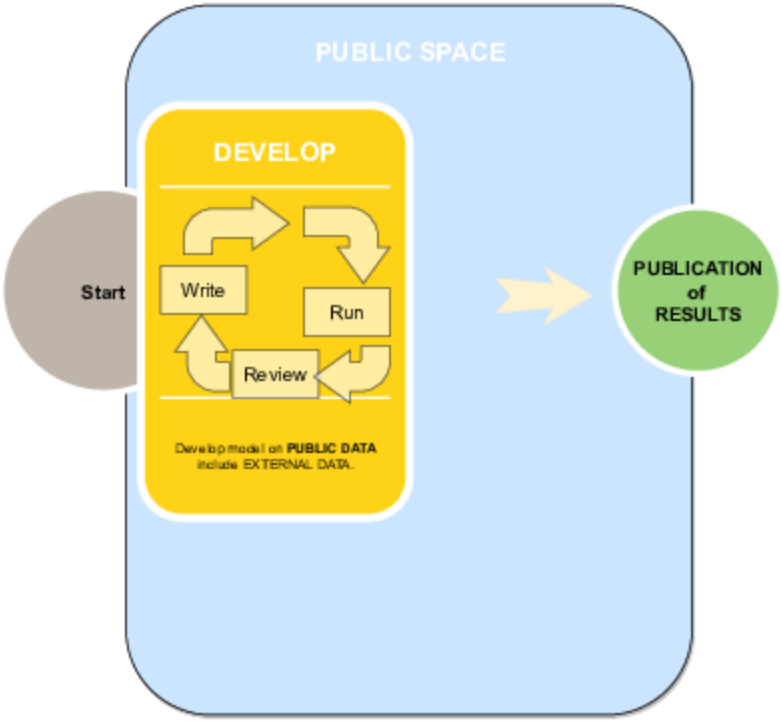
Run a container from Codeoce




Reproducible Run
or launch a cloud workstation
lab
Studio
jupyter
Shiny
VS
Code
BETA
BETA
igv
Timeline

Develop at will

- Arbitrary Stata, R, Python, etc. code


PUBLIC SPACE
DEVELOP
Start
Write
Run
Review
Develop model on PUBLIC DATA
include EXTERNAL DATA.
PUBLICATION
of
RESULTS

Provider perspective: Se
build

First impressions

. 3

Internal build

- Prepare an internal container, compliant with IT secu
  - secure configuration of container running system (
  - add layer of common software (Stata, R, Python, va combinations) for **analysis system**
  - test suite (scripted) for updates

Ability to leverage existing exper

- Can leverage existing container recipes for well-know
  packages (`rocker` for R containers, datascience conta
- Can leverage existing containers and harden the OS (
- Already has process in place to securely vet imported
  packages - can be **reused**

Public build

- Public “recipe” is the same as for internal
  - possibly up to secure base container - close enoug
    enough
  - built by StatAgency itself

Example: Build internal analysis sy

```
FROM registry.internal.statagency.gov/os/ubuntu-24.04-secured

# Install Stata from internal sources (simple tar file), no license
...
# Install R from internal sources
...
USER rstudio
```

Example: Build public analysis sy

```
FROM ubuntu-24.04

# Install Stata from internal sources (simple tar file), no license
...
# Install R from internal sources
...
USER rstudio
```

Optional elements

While not strictly necessary, containers might co

- development environments (Stata GUI, Jupyter noteb
Rstudio)
- standard set of libraries (Stata ado files, R libraries, Py
packages)

Public posting

Prepared containers and recipes can be posted on pub

- post container on public registry (Docker Hub, Google Registry, etc.)
- post recipe on public repository (GitHub, GitLab, etc.)

Posted on Docker Hub

docker hub

Search Docker Hub

Explore / dataeditors/stata17

dataeditors/stata17 ☆9

By dataeditors • Updated 2 months ago

Docker image for Stata, to be used in automation and reproducibility.

IMAGE

Overview Tags

Docker image basic Stata image

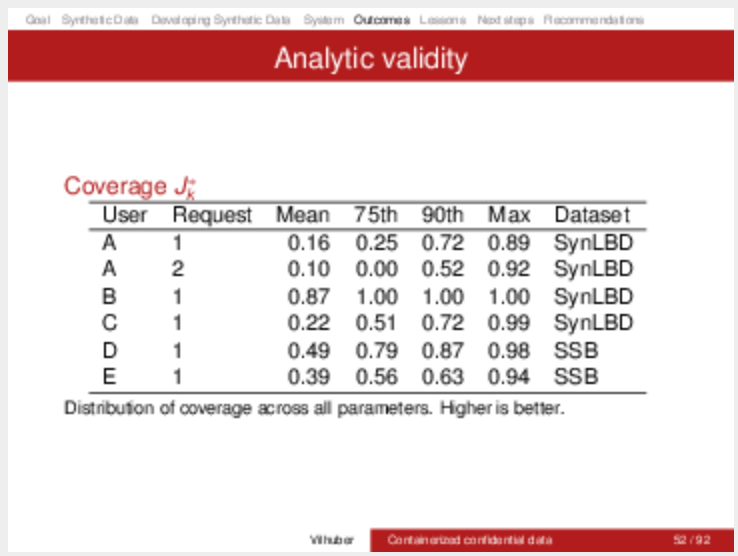
Purpose

This Docker image is meant to isolate and stabilize that environment, and should be portable across operating system, as long as Docker is available.

To learn more about the use of containers for research reproducibility, see Carpentries' docker

Also required: data

But if validation and verification are a key part of it, the
can be lower (**plausible, not analytically va**



Goal Synthetic Data Developing Synthetic Data System **Outcomes** Lessons Next steps Recommendations

Analytic validity

Coverage J_k^*

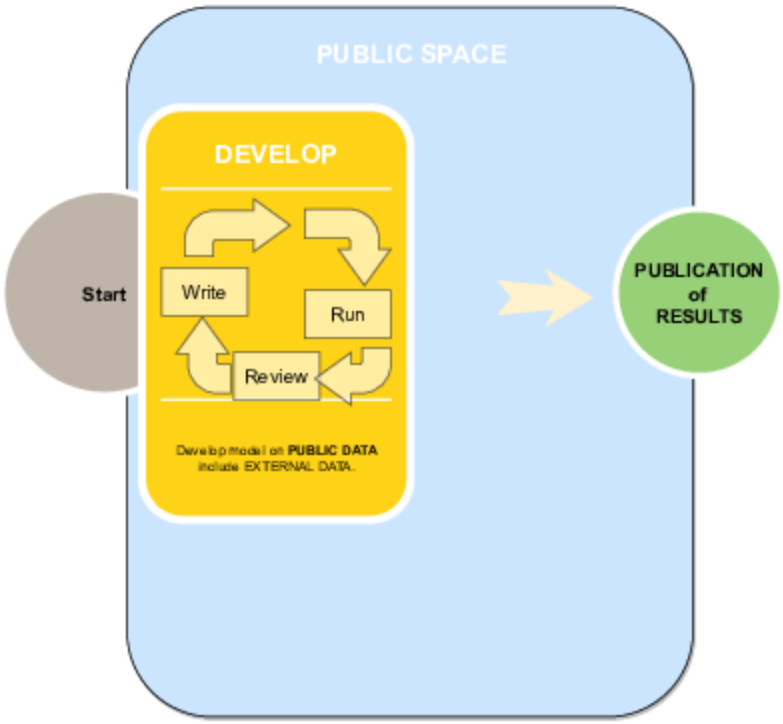
User	Request	Mean	75th	90th	Max	Dataset
A	1	0.16	0.25	0.72	0.89	SynLBD
A	2	0.10	0.00	0.52	0.92	SynLBD
B	1	0.87	1.00	1.00	1.00	SynLBD
C	1	0.22	0.51	0.72	0.99	SynLBD
D	1	0.49	0.79	0.87	0.98	SSB
E	1	0.39	0.56	0.63	0.94	SSB

Distribution of coverage across all parameters. Higher is better.

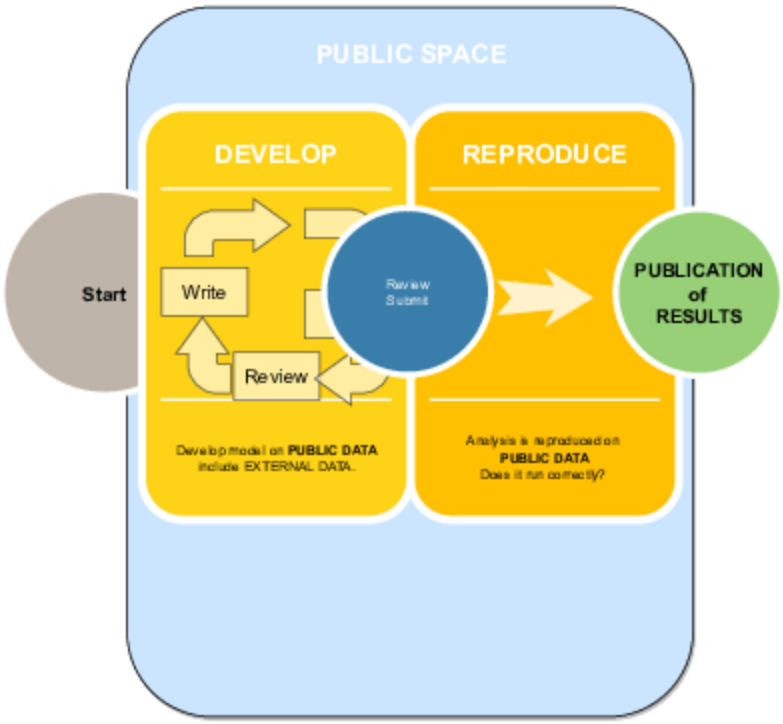
Vilhuber Containerized confidential data 52 / 92

Validation

User Develops



User tests


PUBLIC SPACE
DEVELOP
REPRODUCE
Start
Write
Review
Review
Submit
PUBLICATION
of
RESULTS
Develop model on PUBLIC DATA
include EXTERNAL DATA.
Analysis is reproduced on
PUBLIC DATA
Does it run correctly?

User Submits for Validation

- Submit container *recipe* (`Dockerfile`) and code for v
  StatAgency.gov

```
./Dockerfile
./code/01_prepare_data.R
./code/02_run_analysis.R
./code/03_create_figures.R
```

Important security aspect

No binary code is transmitted

Any external data may need to be vetted.

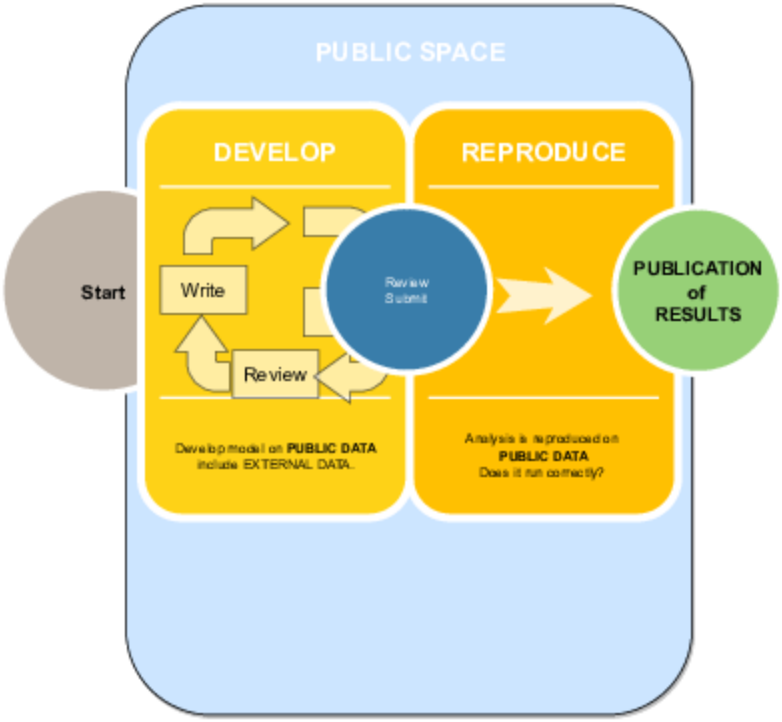
StatAgency Upon receipt of subm

(**Automated**) system receives and processe

```
./Dockerfile
./code/01_prepare_data.R
./code/02_run_analysis.R
./code/03_create_figures.R
```

StatAgency validates reproducil

Just to check that user actually did test...


PUBLIC SPACE
DEVELOP
REPRODUCE
Start
Write
Review
Review
Submit
PUBLICATION
of
RESULTS
Develop model on PUBLIC DATA
include EXTERNAL DATA.
Analysis is reproduced on
PUBLIC DATA
Does it run correctly?

Provider validates reproducibi

If rejected, **automated system** returns to user without

If accepted, proceed to validation step

Provider rebuilds container using
base image

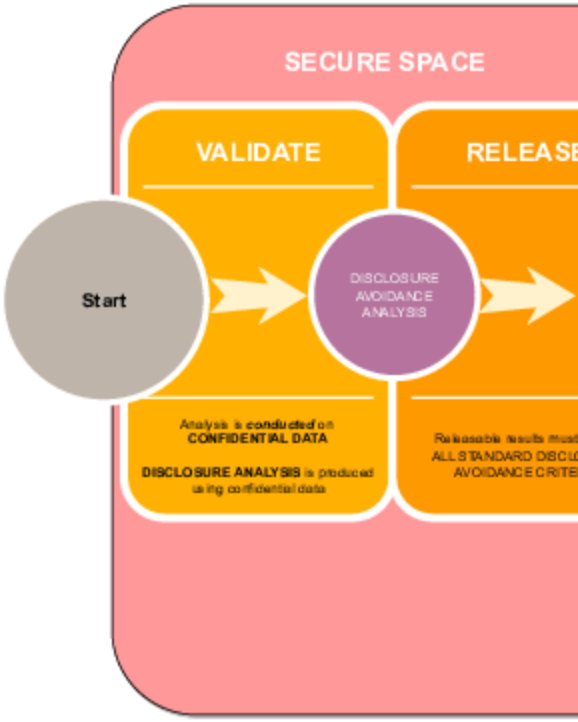
- Input is only the `Dockerfile` recipe
- Security scanning of (plaintext) scripts and of resul
- Build can occur in a sandboxed environment

Necessary restrictions

While useful in the public space, when running intern
pre-vetting,

- containers would be restricted in terms of internet ac
- containers may be built against only known safe sour
packages (e.g. internal mirrors)

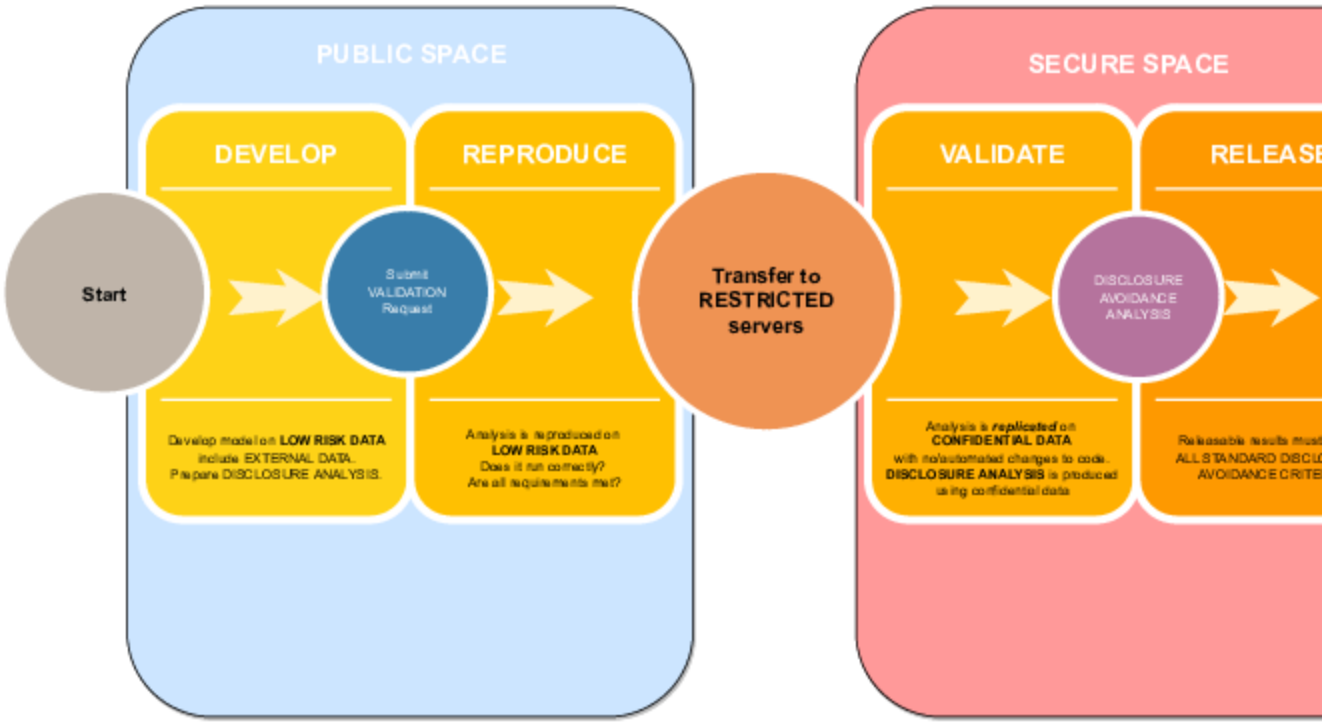
Once image is built



Validate against confidential d

- Same image is used for confidential data
- Only difference: swap out public (test) data for confid
- Processing may involve more complex processing, for **bootstrapping errors** or **obtaining multiple estima** various partially protected datasets
- Disclosure avoidance may involve transparently **mod** **certain functions**, or **post-processing of results**

Return results to user



Challenges

Automation or streamlining of disc
avoidance

Scalability of a system **hinges critically on streamlin**
vetting.

However, the challenge of creating automated and
disclosure avoidance procedures *is not unique* to the
process described here.

Security of containers

In general, bad idea to blindly run untrusted containe
this is a **solved problem** in the industry, facilitated by t
sparsity of the build process.

User acceptance

As a reminder, most social scientists are **not famil**
containers.

- **Mitigation**:
  - Off-the-shelf solutions (Codeocean, Whole Tale
  - IT support at universities and research institut

Advantages

Existing technology

- Containers are well-known technology, including in ot
- Used by online services (Codeocean, Onyxia, but also
  etc.)

Scalability

- Easy to scale to large number of users
- Easy to scale to technologies that allow for sophistica
computing intensive disclosure avoidance

Cheap

- for users
  - most of the core enabling technology is free to use
  - support by university IT is generally available
- for providers (*StatAgency*)
  - no need to provision scaled infrastructure for user
  - can leverage existing on-site software stacks (e.g., that anything used internally is already security-ve

Additional benefit

- **StatAgency** can accumulate a library of confirmed re
containers and models, and can test out new data, di
methods, etc. at scale against prior scientific findings

Consider the new disclosure avoidance method fo

- Can be tested against every submitted model that us
as long as database schema is the same.

Thank you

Quick links for the curio

- https://www.datacamp.com/tutorial/docker-for-data-
introduction
- CodeOcean
- Whole Tale
- Onyxia
- Docker Hub
- Stata on Docker

Published **Compute Capsule for: Robust Predictions for DSGE Models with Incomplete Information** *(Ryan Chahrour & Robert Ulbricht)*

Capsule File Help

Files | Tabs

Core Files

- metadata 1.06 KB
- environment 199 B
- code 523.97 KB
  - auto_generated 64 B
  - helper_functions 21.08 KB
  - Section4 88.46 KB
  - Section5 220.51 KB
  - Section5_estimation 79.99 KB
  - clean_up.m 182 B
  - LICENSE 1.47 KB
  - make_all_outputs.m 979 B
  - read_me.pdf 109.8 KB
  - run 262 B
  - run_all.m 1.17 KB
- data 0 B

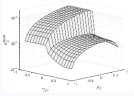
Results

- results 2.52 MB

Other Files

Metadata

Economics



Compute Capsule for: Robust Predictions for DSGE Models with Incomplete Information

Ryan Chahrour, Robert Ulbricht

We provide predictions for DSGE models with incomplete information that are robust across information structures. Our approach maps an incomplete-information model into a full-information economy with time-varying expectation wedges and provides conditions that ensure the wedges are rationalizable by some information structure. Using our approach, we quantify the potential importance of information as a source of business cycle fluctuations in an otherwise frictionless model. Our approach uncovers a central role for firm-specific demand shocks in supporting aggregate confidence fluctuations. Only if firms face unobserved local demand shocks can confidence fluctuations account for a significant portion of the US business cycle.

Capsule

DOI 10.24433/CO.5177698.v2

Citation General

Ryan Chahrour, Robert Ulbricht (2024) Compute Capsule for: Robust Predictions for DSGE Models with Incomplete Information [Source Code]. https://doi.org/10.24433/CO.5177698.v2

Licenses **Code** BSD 3-Clause license

Authors Ryan Chahrour Boston College Robert Ulbricht Boston College

Corresponding Contributor Ryan Chahrour ryan.chahrour@bc.edu

Associated Publication

DOI 10.1257/mac.20200053

Title Robust Predictions for DSGE Models with Incomplete Information

Publication Date January 2023

Journal/Conference American Economic Journal: Macroeconomics

Citation

This presentation

- Github
- Presentation
  - Dockerfile!
  - Container!

References

Barrientos, Andrés F., Alexander Bolton, Tom Balmat
Reiter, John M. de Figueiredo, Ashwin Machanavajjhal
Charley Kneifel, and Mark DeLong. 2018. "Providing
Confidential Research Data Through Synthesis and Ve
Application to Data on Employees of the U.S. Federal G
The Annals of Applied Statistics, June. http://arxiv.org/abs
Boettiger, Carl. 2015. "An Introduction to Docker for R
Research." *ACM SIGOPS Operating Systems Review* 49
https://doi.org/10.1145/2723872.2723882

1. https://www.datacamp.com/tutorial/docker-for-data-
introduction↩

2. An earlier version of this presentation mentioned Gig
not unusual in this space, Gigantum no longer functio
company.↩

3. Image credit Christopher Scholz, under
SA 2.0↩