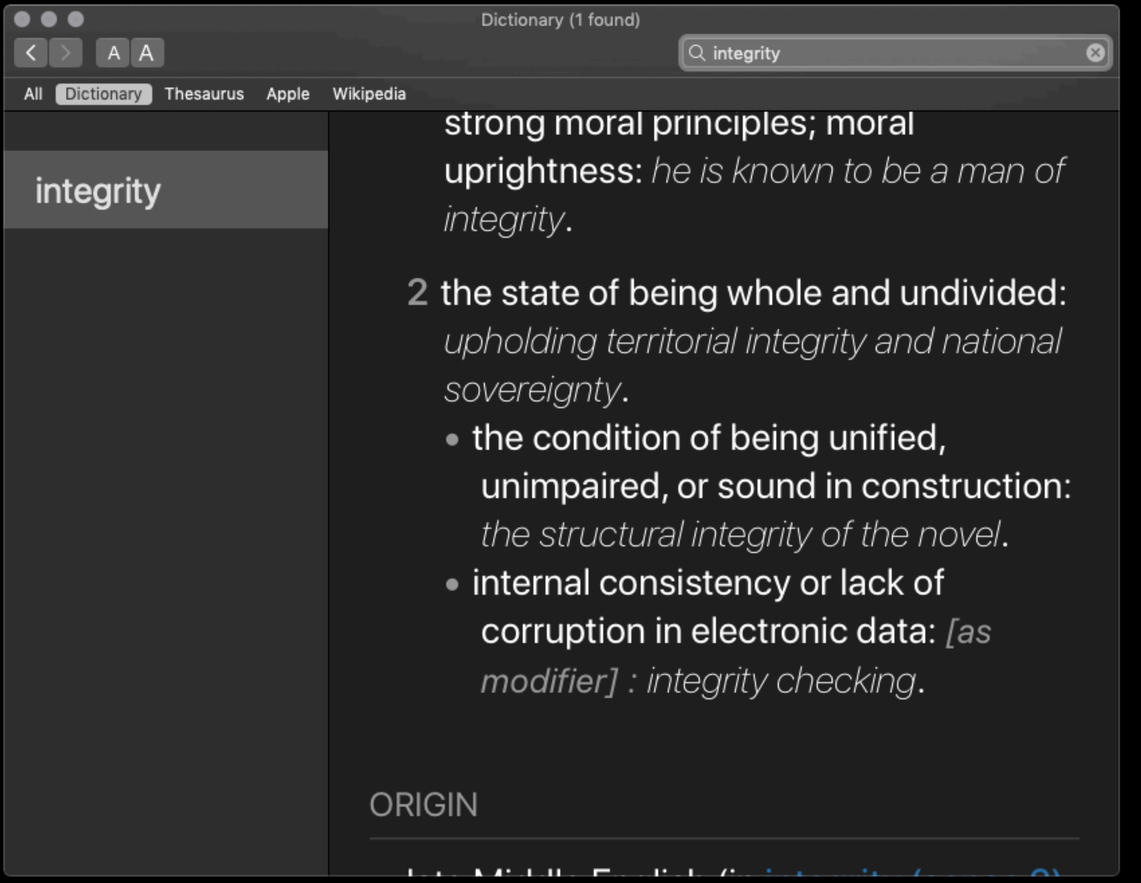
Zen & the art of ensuring confidentiality & integrity in analytics workflows

- Chief Data Scientist @ Rapid7
- Co-author **Data-Driven Security** (Wiley, 2014)
- ~23 mostly really weird/niche CRAN package
- Lots of other ones (mostly web/cyber/weird-data formats focused) on SourceHut (git.sr.ht/~hrbrmstr), GitLab (gitlab.com/hrbrmstr) or GitUgh (github.com/hrbrmstr)
- Blog (mostly R): https://rud.is/b/
- @hrbrmstr • bob@rud.is • bob_rudis@rapid7.com


Dictionary (1 found)
confidentiality
All
Dictionary
Thesaurus
Apple
Wikipedia
confidentiality
confidentiality
| känfədenSHē'ələdē |
noun
the state of keeping or being kept secret or private: the lead bank's duty of confidentiality to the borrower | [as modifier] : visitors have to sign confidentiality agreements.


Dictionary (1 found)
confidentiality
All Dictionary Thesaurus Apple Wikipedia
confidentiality
confidentiality
| känfədenSHē'aləd ē |
noun
the state of keeping or being kept secret or private: the lead bank's duty of confidentiality to the borrower | [as modifier] : visitors have to sign confidentiality agreements.


Dictionary (1 found)
integrity
All Dictionary Thesaurus Apple Wikipedia
integrity
strong moral principles; moral uprightness: he is known to be a man of integrity.
2 the state of being whole and undivided: upholding territorial integrity and national sovereignty.
• the condition of being unified, unimpaired, or sound in construction: the structural integrity of the novel.
• internal consistency or lack of corruption in electronic data: [as modifier] : integrity checking.
ORIGIN

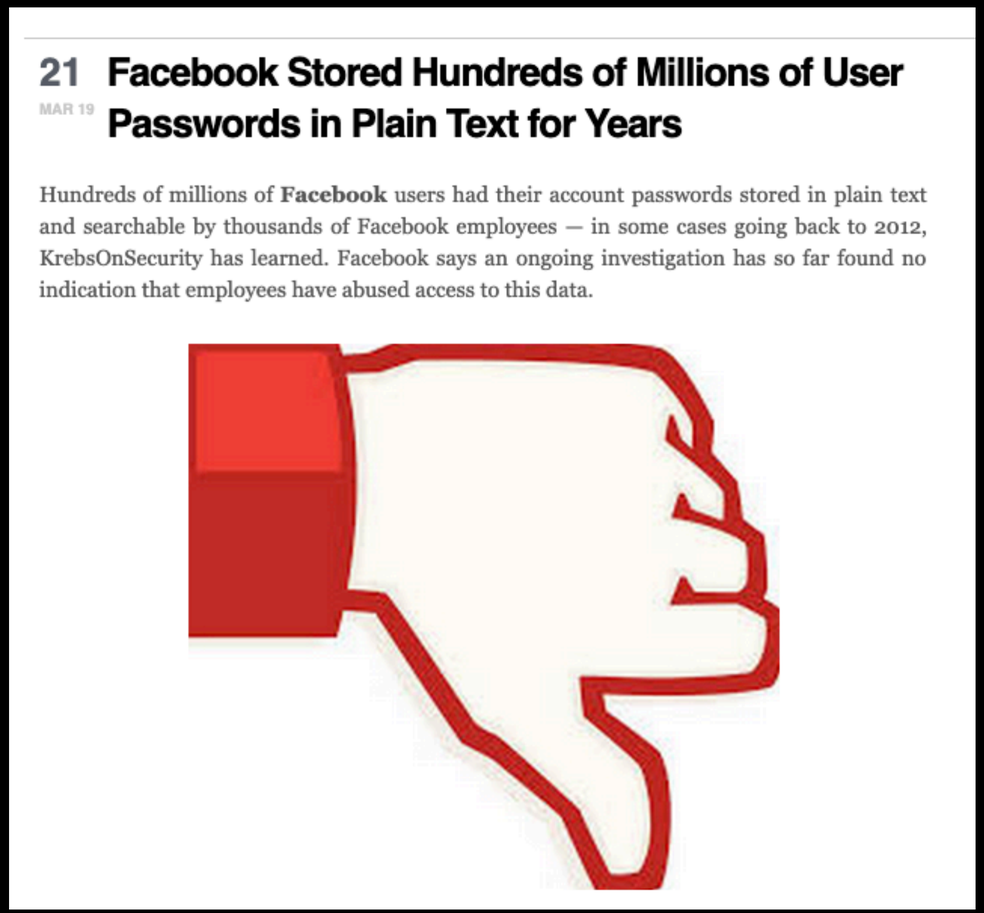
"*…both functional & emergent properties of reproducible analytics workflows with* **integrity** *being a required property & confidentiality being a conditional property.*"

ROpenSci

CONFIDENTIAL(ity)

- Credentials/Passwords (hrbrmstr:myPetsname1!, Pa$$w0rd123!)

- Credentials/Passwords (hrbrmstr:myPetsname1!, Pa$$w0rd123!)



21 MAR 19

Facebook Stored Hundreds of Millions of User Passwords in Plain Text for Years

Hundreds of millions of **Facebook** users had their account passwords stored in plain text and searchable by thousands of Facebook employees — in some cases going back to 2012, KrebsOnSecurity has learned. Facebook says an ongoing investigation has so far found no indication that employees have abused access to this data.

- Credentials/Passwords (hrbrmstr:myPetsname1!, Pa$$w0rd123!)

21 MAR 19

Facebook Stored Hundreds of Millions of User Passwords in Plain Text for Years

Hundreds of millions of **Facebook** users had their account passwords stored in plain text and searchable by thousands of Facebook employees — in some cases going back to 2012, KrebsOnSecurity has learned. Facebook says an ongoing investigation has so far found no indication that employees have abused access to this data.

10 most hacked passwords revealed in 2019 report

By Jennifer Earl | Published April 25, 2019 | Cyber Security | FOXBusiness

After sifting through the top 100,000 hacked passwords, the agency pointed to these 10 as the most hacked.

1. 123456
2. 123456789
3. qwerty
4. password
5. 111111
6. 12345678
7. abc123
8. 1234567
9. password1
10. 12345

- API Tokens (GITHUB_TOKEN)

- API Tokens (GITHUB_TOKEN)

Over 100,000 GitHub repos have leaked API or cryptographic keys

Thousands of new API or cryptographic keys leak via GitHub projects every day.

By Catalin Cimpanu for Zero Day | March 21, 2019 -- 23:21 GMT (16:21 PDT) | Topic: Security

A scan of billions of files from 13 percent of all GitHub public repositories over a period of six months has revealed that over 100,000 repos have leaked API tokens and cryptographic keys, with thousands of new repositories leaking new secrets on a daily basis.

The scan was the object of academic research carried out by a team from the North Carolina State University (NCSU), and the study's results have been shared with GitHub.

- Cryptographic keys (used for encrypting/decrypting plaintext)

- Cryptographic keys (used for encrypting/decrypting plaintext)

PGP/GPG "pretty good privacy"
Ref: gnupg.org

- Cryptographic keys (used for encrypting/decrypting plaintext)

PGP/GPG "pretty good privacy"
Ref: gnupg.org

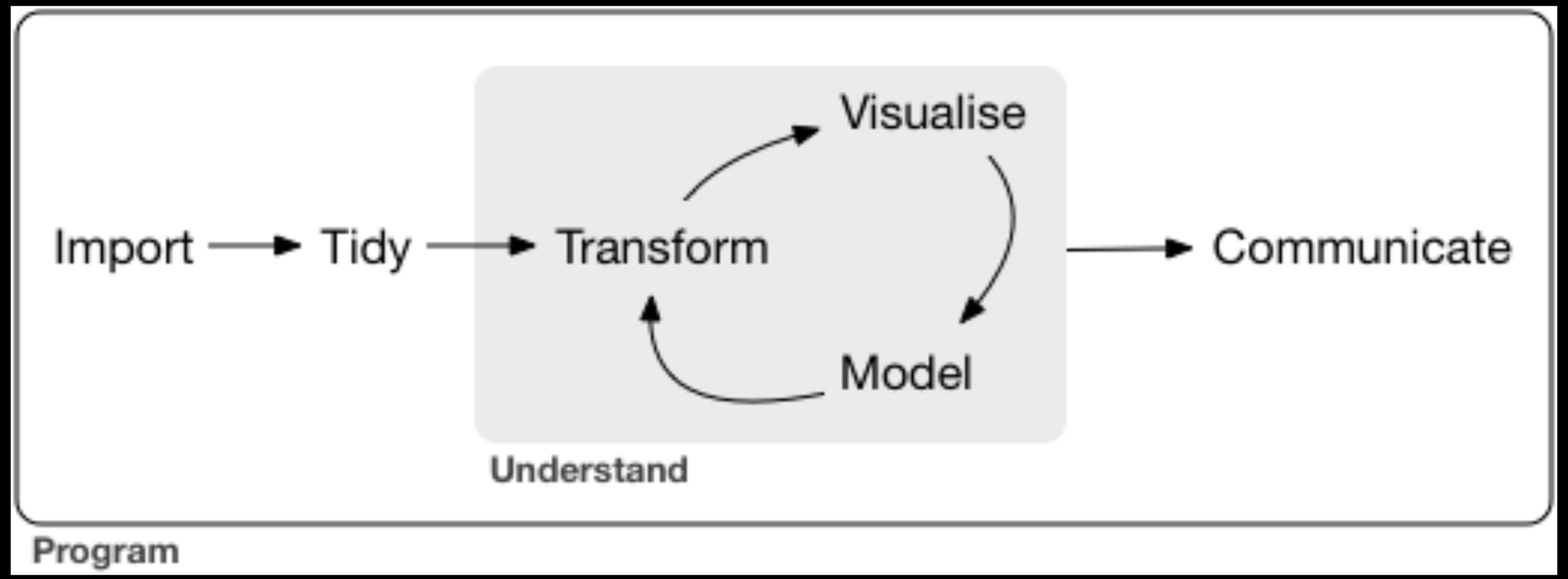
SSH "secure shell"
jumpcloud.com/blog/what-are-ssh-keys

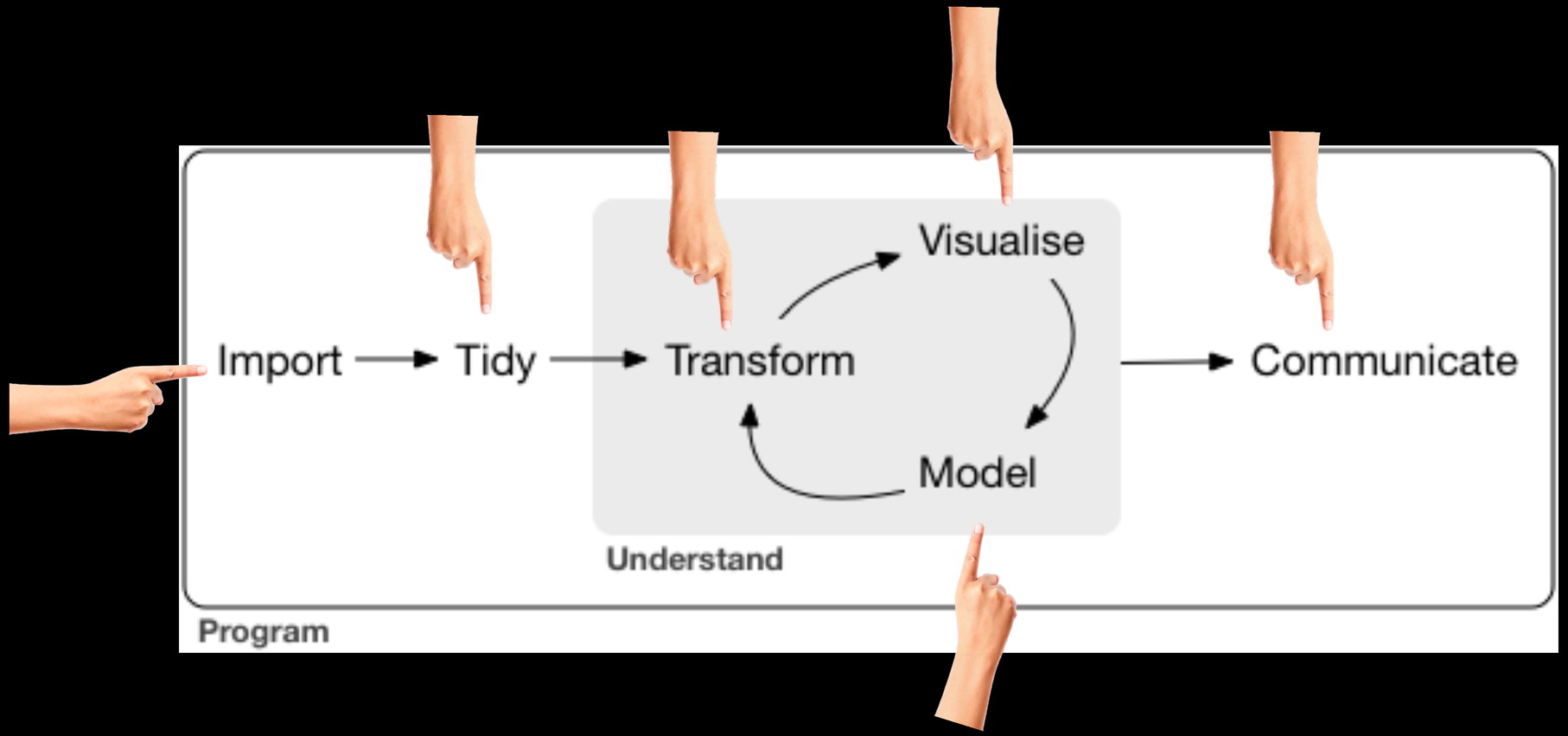
- Customer Data

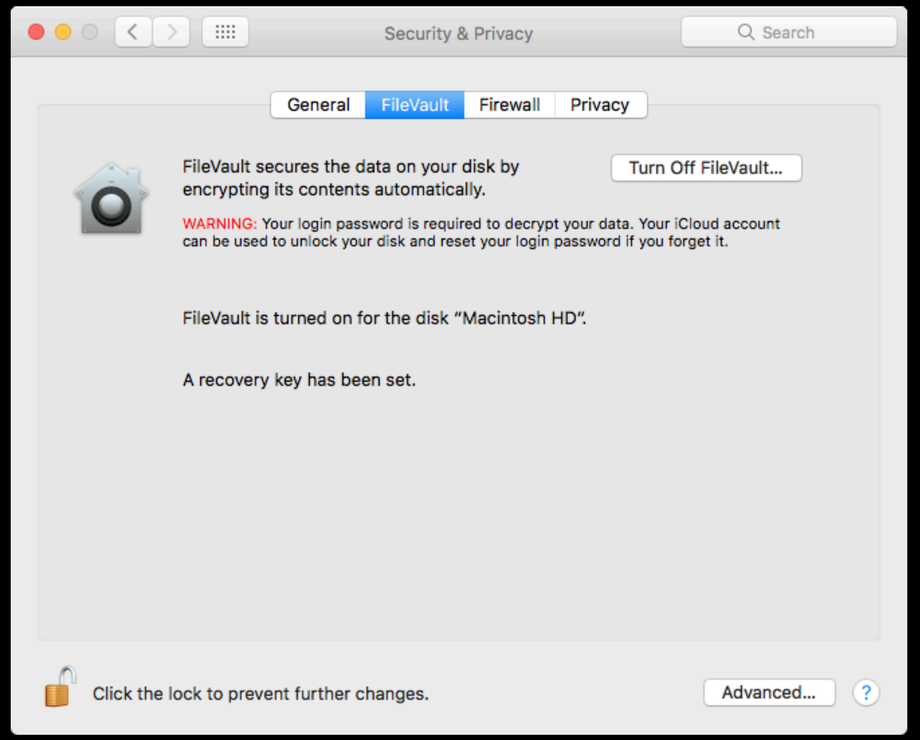
- Customer Data
- Patient Data

- Customer Data
- Patient Data
- Survey Responses

- Customer Data
- Patient Data
- Survey Responses
- Sensitive Research Topics (Victims, Refugees, Exposés)


Visualise
Import
Tidy
Transform
Communicate
Model
Understand
Program


Import
Tidy
Transform
Visualise
Model
Communicate
Understand
Program


Security & Privacy
Search
General
FileVault
Firewall
Privacy
FileVault secures the data on your disk by encrypting its contents automatically.
Turn Off FileVault...
WARNING: Your login password is required to decrypt your data. Your iCloud account can be used to unlock your disk and reset your login password if you forget it.
FileVault is turned on for the disk "Macintosh HD".
A recovery key has been set.
Click the lock to prevent further changes.
Advanced...
?


KeePass
PASSWORD SAFE

okta

okta

okta

okta

Amazon S3

Leaky Buckets: 10 Worst Amazon S3 Breaches

By Ericka Chickowski on Jan 24, 2018 | 0 Comments

The last year has proved out about security naysayers' warnings about the undisciplined use of cloud architectures. While many organizations work hard to secure data stored on cloud stores, the truth is that there's a lot of work to go. That fact is made abundantly clear by the growing number of incidents caused by extremely poor security hygiene within Amazon Simple Storage Service (S3) storage buckets that are holding very sensitive information.

According to recent statistics, as many as 7% of all S3 servers are completely publicly accessible without any authentication and 35% are unencrypted. And if the incidents of the past six months or so are any indication, these aren't low-value data stores. Here are some of the worst recent leaks caused by poorly configured Amazon S3 resources.

Keybase

{keybase} • https://github.com/ropenscilabs/keybase

"This is me" / "I Own This"

GPG Key Management

Encrypted & User-access Controlled Git Repos

Keybase

Encrypted & User-access Controlled Folders

Encrypted Comms

{keybase} • https://github.com/ropenscilabs/keybase

- **{openssl}** • https://cran.r-project.org/package=openssl

  Tools for working with certificates (like those used in your browser) and generating signatures & hashes.

https://cran.r-project.org/web/packages/openssl/vignettes/crypto_hashing.html

- **{sodium}** • https://cran.r-project.org/package=sodium

  Generic tools for symmetric (single, seekrit key) or asymmetric (public/private keys) encryption/decryption

https://cran.r-project.org/web/packages/sodium/vignettes/crypto101.html

- **{gpg}** • https://cran.r-project.org/package=gpg

  Tools to interface with GnuPG and supports encryption, decryption, digital signatures & local GPG management.

https://cran.r-project.org/web/packages/gpg/vignettes/intro.html

- **{ssh}** • https://cran.r-project.org/package=ssh

  Connect to a remote server over SSH to transfer files via SCP, setup a secure tunnel, or run a command or script on the host while streaming stdout and stderr directly to the client.

https://cran.r-project.org/web/packages/ssh/vignettes/intro.html

- **{cyphr}** • https://cran.r-project.org/package=cyphr

  Higher-level/friendly wrappers to cryptographic functions in **{sodium}** & **{openssl}** designed to make it really, really easy

https://ropensci.github.io/cyphr/

- **{encryptr}** • https://cran.r-project.org/package=encryptr

  Uses **{openssl}** and is focused primarily on data frame (tibble) columns but also works with files.

https://encrypt-r.org

- **{digest}** • https://cran.r-project.org/package=digest

  Create hashes of arbitrary R objects.

https://encrypt-r.org

"Homomorphic encryption is a form of encryption that allows computation on ciphertexts, generating an encrypted result which, when decrypted, matches the result of the operations as if they had been performed on the plaintext."

— en.wikipedia.org/wiki/Homomorphic_encryption

- **{homomorpheR}** • https://cran.r-project.org/package=homomorpheR

  Based on **{sodium}** & enables performing statistical and machine learning operations on encrypted data.

https://cran.r-project.org/web/packages/homomorpheR/vignettes/introduction.html

- **{credentials}** • https://cran.r-project.org/package=credentials

  Setup and retrieve HTTPS and SSH credentials for use with git and other services.

https://encrypt-r.org

- **{keyring}** • https://cran.r-project.org/package=keyring

  An R API to access credential stores such as macOS Keychain, Windows Credential Store, Linux Secret Service along with the ability for others to create additional back-ends.

https://github.com/r-lib/keyring/blob/master/inst/development-notes.md

"KEEP IT SEEKRIT"

"KEEP IT SEEKRIT"

"KEEP IT SAFE"

- Where do you source your packages from?

- Where do you source your packages from?

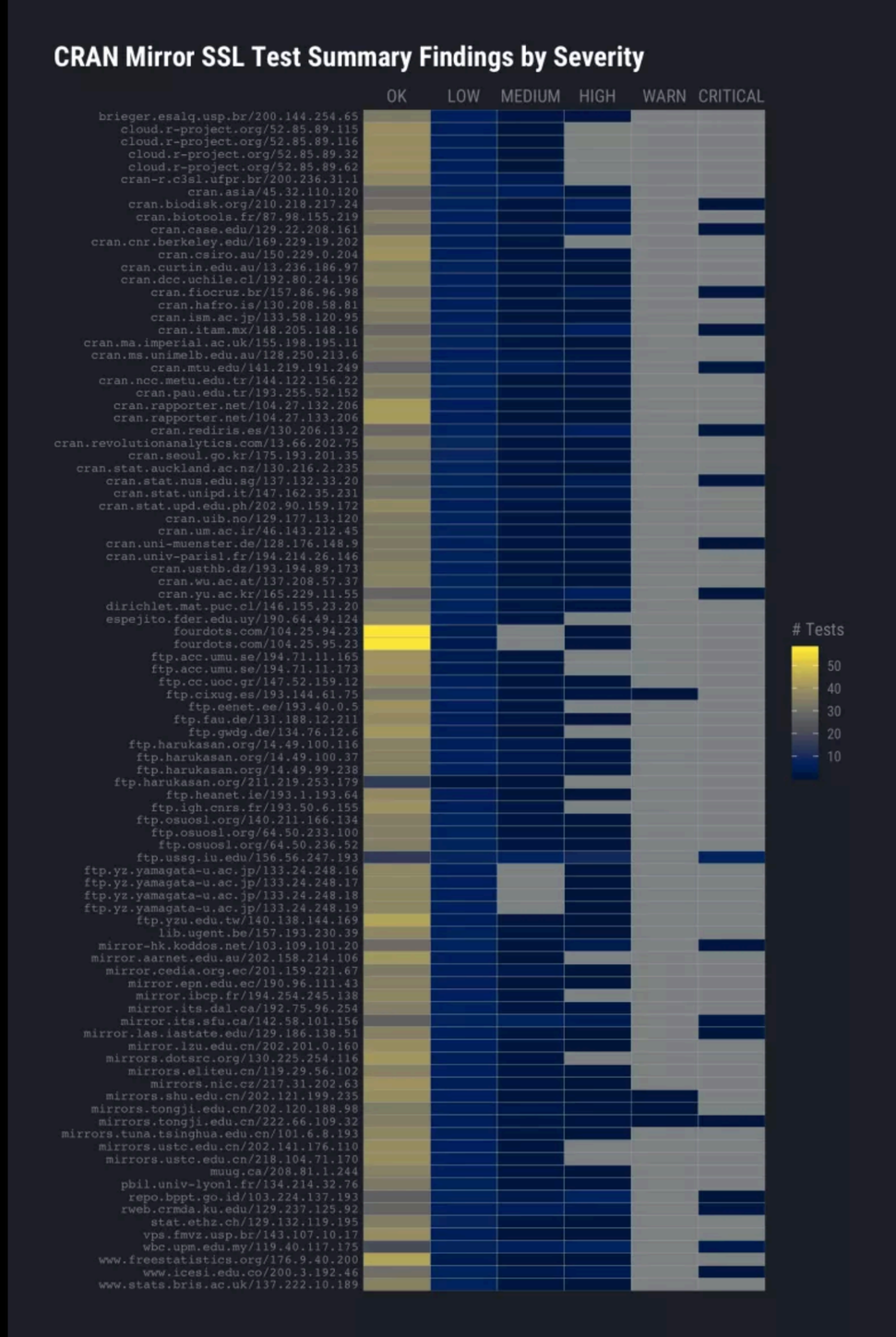
Twelve malicious Python libraries found and removed from PyPI

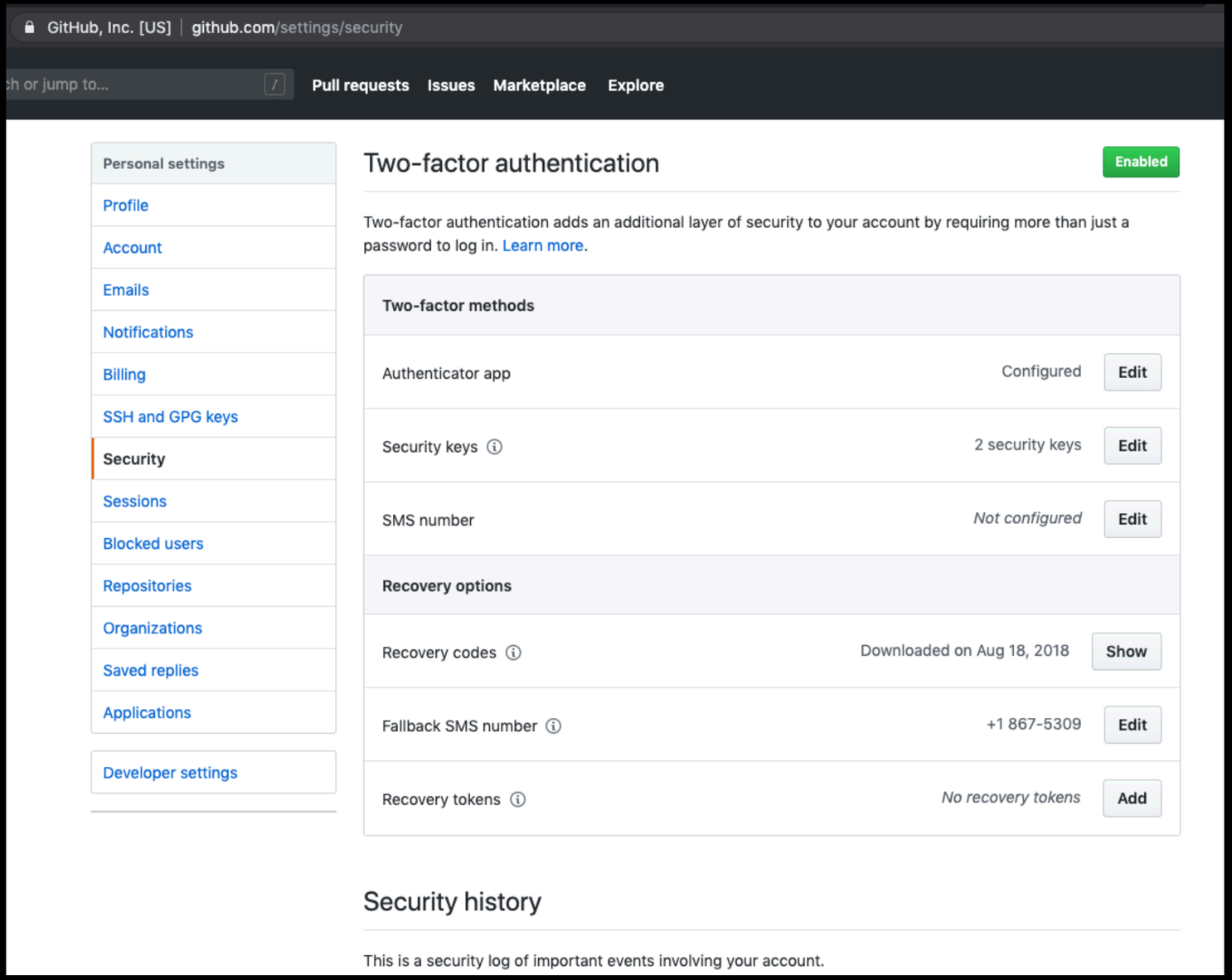
One package contained a clipboard hijacker that replaced victims' Bitcoin addresses in an attempt to hijack funds from users.

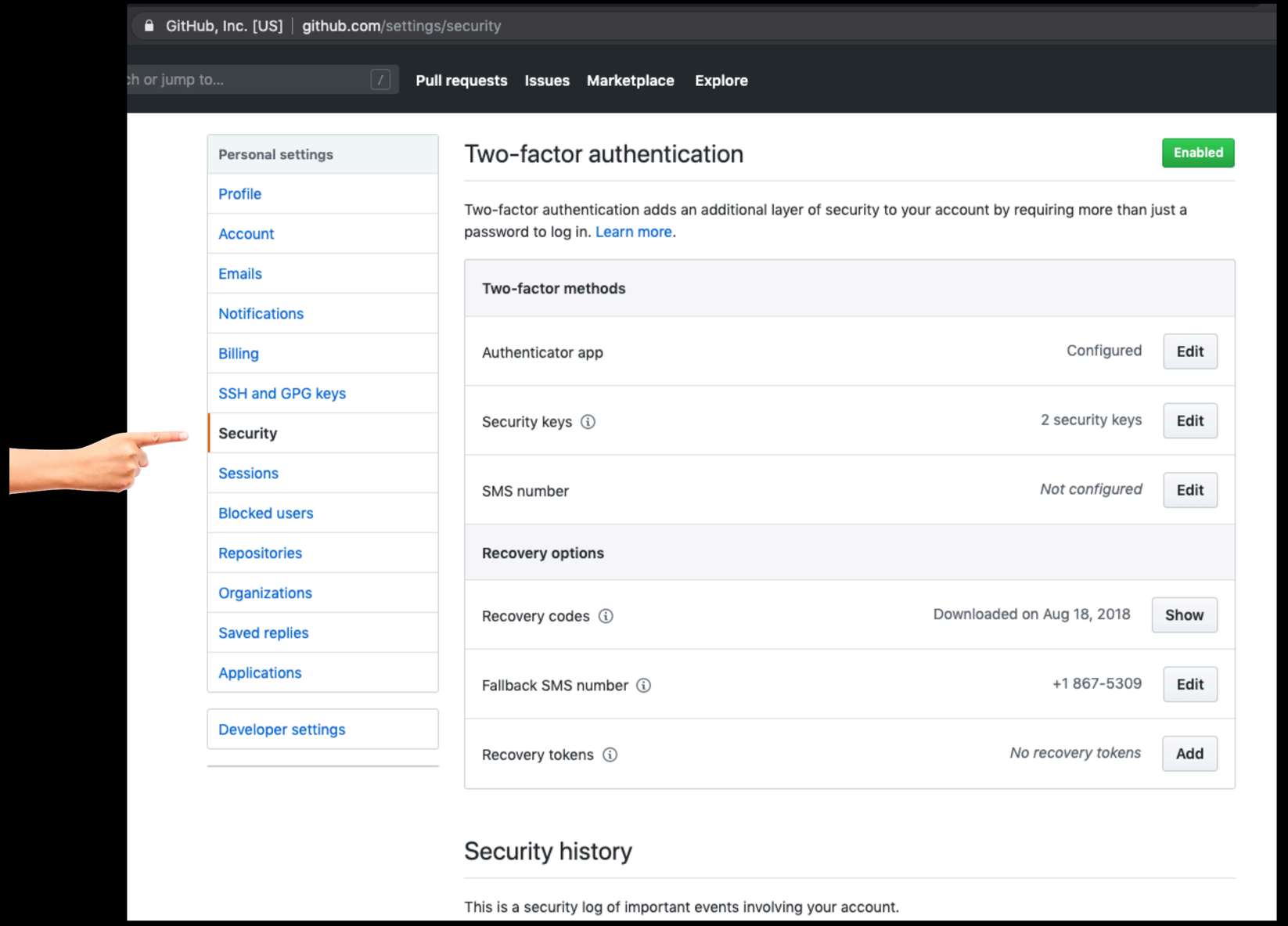
By Catalin Cimpanu for Zero Day | October 27, 2018 -- 08:00 GMT (01:00 PDT) | Topic: Security

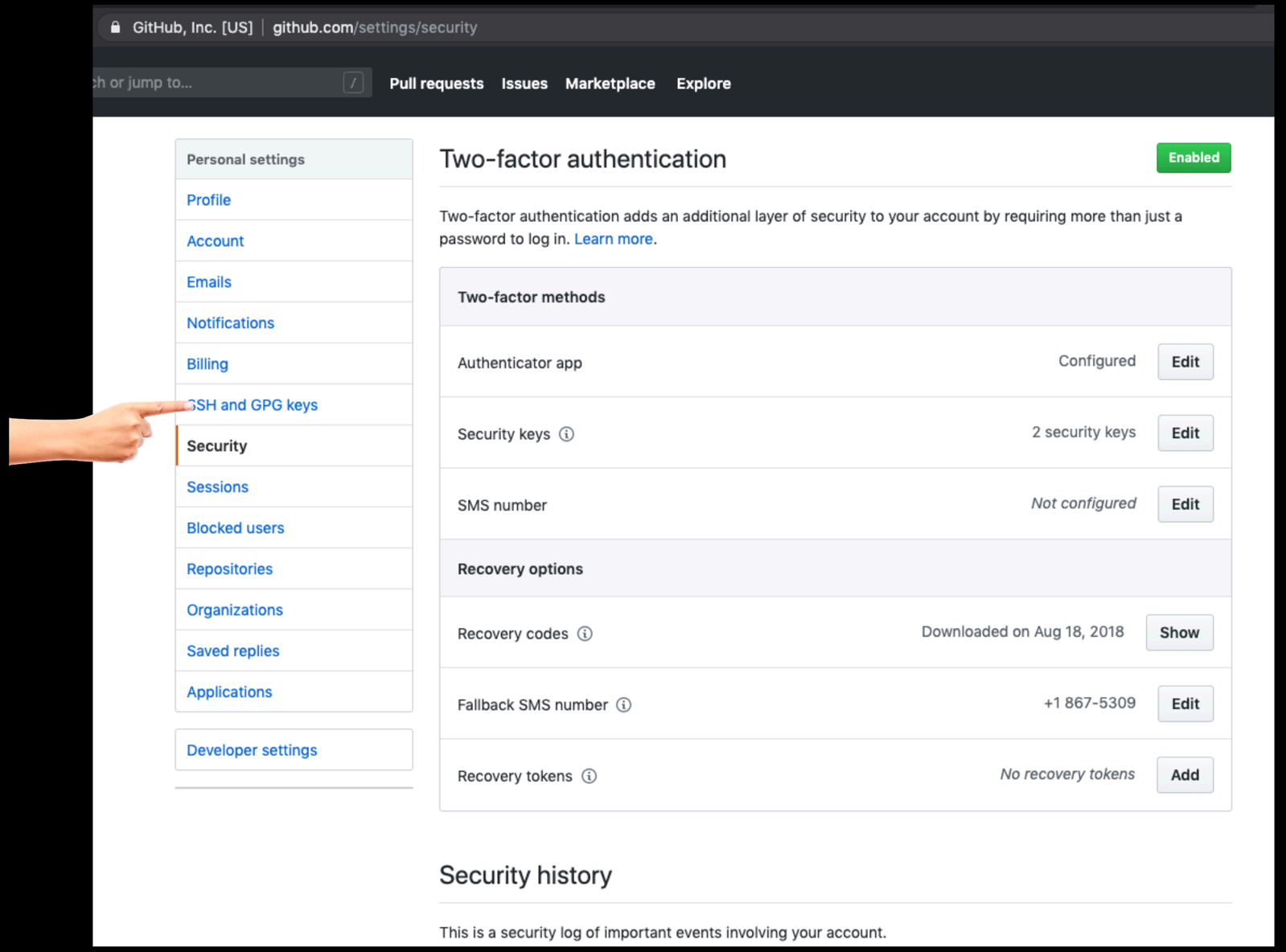
- Where do you source your packages from?

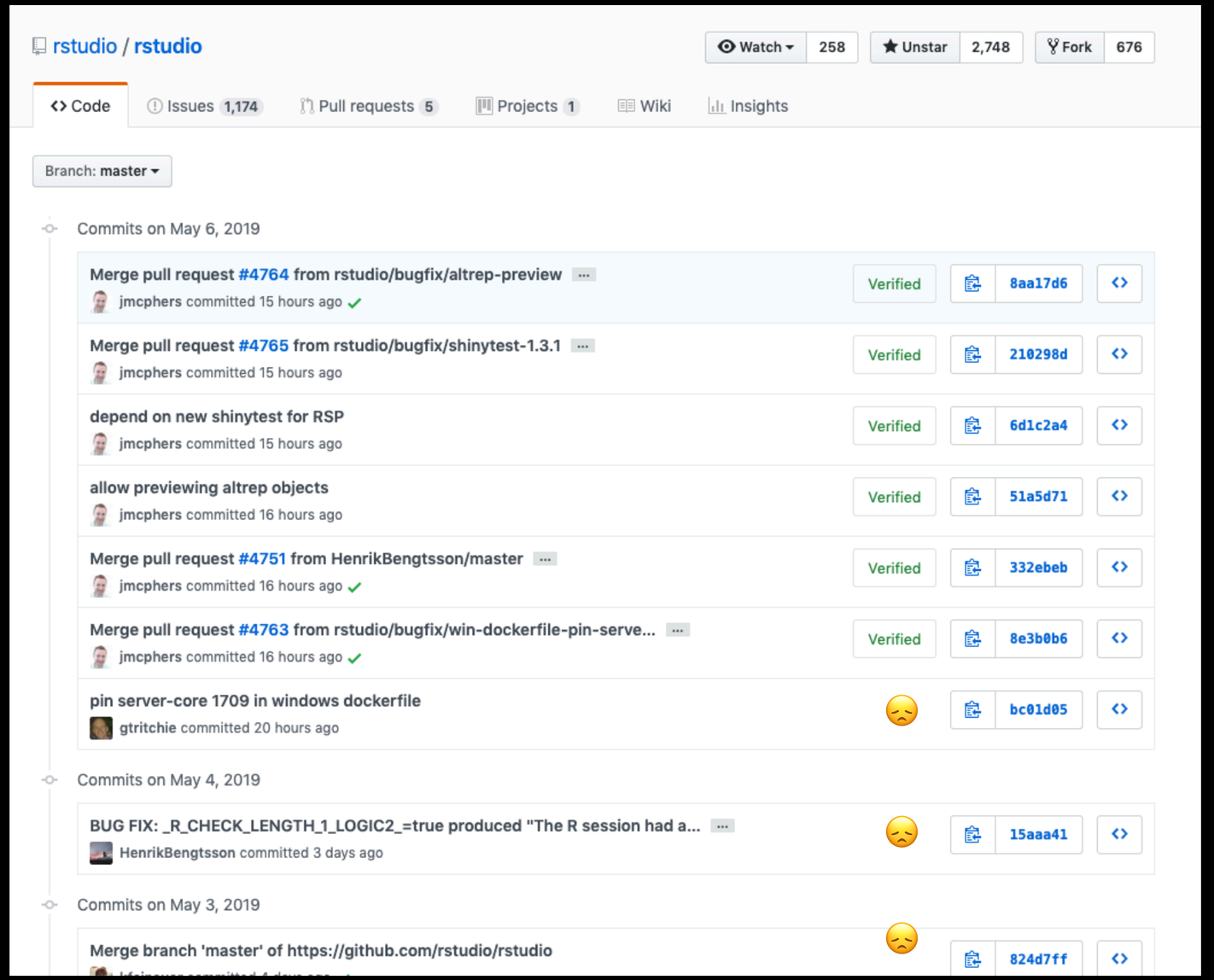
CRAN Mirror "Security"
https://rud.is/b/2019/03/03/cran-mirror-security/




GitHub, Inc. [US] | github.com/settings/security
ch or jump to...
/
Pull requests
Issues
Marketplace
Explore
Personal settings
Profile
Account
Emails
Notifications
Billing
SSH and GPG keys
Security
Sessions
Blocked users
Repositories
Organizations
Saved replies
Applications
Developer settings
Two-factor authentication
Enabled
Two-factor authentication adds an additional layer of security to your account by requiring more than just a password to log in. Learn more.
Two-factor methods
Authenticator app
Configured
Edit
Security keys
2 security keys
Edit
SMS number
Not configured
Edit
Recovery options
Recovery codes
Downloaded on Aug 18, 2018
Show
Fallback SMS number
+1 867-5309
Edit
Recovery tokens
No recovery tokens
Add
Security history
This is a security log of important events involving your account.


GitHub, Inc. [US] | github.com/settings/security
ch or jump to...
/
Pull requests
Issues
Marketplace
Explore
Personal settings
Profile
Account
Emails
Notifications
Billing
SSH and GPG keys
Security
Sessions
Blocked users
Repositories
Organizations
Saved replies
Applications
Developer settings
Two-factor authentication
Enabled
Two-factor authentication adds an additional layer of security to your account by requiring more than just a password to log in. Learn more.
Two-factor methods
Authenticator app
Configured
Edit
Security keys
2 security keys
Edit
SMS number
Not configured
Edit
Recovery options
Recovery codes
Downloaded on Aug 18, 2018
Show
Fallback SMS number
+1 867-5309
Edit
Recovery tokens
No recovery tokens
Add
Security history
This is a security log of important events involving your account.


GitHub, Inc. [US] | github.com/settings/security
ch or jump to...
/
Pull requests
Issues
Marketplace
Explore
Personal settings
Profile
Account
Emails
Notifications
Billing
SSH and GPG keys
Security
Sessions
Blocked users
Repositories
Organizations
Saved replies
Applications
Developer settings
Two-factor authentication
Enabled
Two-factor authentication adds an additional layer of security to your account by requiring more than just a password to log in. Learn more.
Two-factor methods
Authenticator app
Configured
Edit
Security keys
2 security keys
Edit
SMS number
Not configured
Edit
Recovery options
Recovery codes
Downloaded on Aug 18, 2018
Show
Fallback SMS number
+1 867-5309
Edit
Recovery tokens
No recovery tokens
Add
Security history
This is a security log of important events involving your account.


rstudio / rstudio
Watch 258
Unstar 2,748
Fork 676
Code
Issues 1,174
Pull requests 5
Projects 1
Wiki
Insights
Branch: master
Commits on May 6, 2019
Merge pull request #4764 from rstudio/bugfix/altrep-preview
jmcphers committed 15 hours ago
Verified
8aa17d6
Merge pull request #4765 from rstudio/bugfix/shinytest-1.3.1
jmcphers committed 15 hours ago
Verified
210298d
depend on new shinytest for RSP
jmcphers committed 15 hours ago
Verified
6d1c2a4
allow previewing altrep objects
jmcphers committed 16 hours ago
Verified
51a5d71
Merge pull request #4751 from HenrikBengtsson/master
jmcphers committed 16 hours ago
Verified
332ebeb
Merge pull request #4763 from rstudio/bugfix/win-dockerfile-pin-serve...
jmcphers committed 16 hours ago
Verified
8e3b0b6
pin server-core 1709 in windows dockerfile
gtritchie committed 20 hours ago
bc01d05
Commits on May 4, 2019
BUG FIX: _R_CHECK_LENGTH_1_LOGIC2_=true produced "The R session had a...
HenrikBengtsson committed 3 days ago
15aaa41
Commits on May 3, 2019
Merge branch 'master' of https://github.com/rstudio/rstudio
824d7ff

While we're on the subject of 2FA…

- Travis-CI does not appear to support 2FA 😞
- Codecov does not appear to support 2FA 😞
- Coveralls does not appear to support 2FA 😞
- AppVeyor does support 2FA 👍
- GitUgh does support 2FA 👍
- GitLab does support 2FA 👍
- SourceHut does support 2FA 👍
- Netifly does support 2FA 👍

- Where do you source your packages from?

  Use only SSL/TLS CRAN mirrors externally to prevent

    - PITM attacks
    - organizational snooping
    - nation-state snooping
    - ISP snooping.

https://www.r-consortium.org/blog/2019/03/27/cii-best-practices-r-package-leaderboard

- Don't believe the badge

CII Best Practices – R Package Leaderboard

By Mark Hornick | March 27, 2019 | Blog

Since my last post on the Core Infrastructure Initiative CII Best Practices Badge for R Packages – responding to concerns, there have been many R language projects started – and completed – on the CII Best Practices site. In this post, we recognize the R projects that have achieved the CII Best Practices – Passing level, and note that several are well on their way to achieving *silver* level. In all, there are more than 50 CII projects related to R packages, with the popular **ggplot2** package at the cusp of joining the group below with 97% completion as of this post.

Please congratulate these package owners for their achievement. If you're a package developer, consider adding your package to the CII Best Practices ranks, and work your way through the levels of *passing*, *silver*, and *gold*.

Id	Name	Description	Owner
265	madrid.air	Parse air quality data published by http://datos.madrid.es/	Ramón Novoa
1882	DBI	A database interface (DBI) definition for communication between R and RDBMSs	Kirill Müller
2011	Delaporte	Provides the probability mass, distribution, quantile, random variate generation, and method of moments parameter estimation	Avraham Adler
2022	lamw	Calculates the real-valued branches of the Lambert-W function	Avraham Adler
2033	pade	Returns the numerator and denominator when given a vector of Taylor series coefficients of sufficient length as input	Avraham Adler
2041	fixedWidth	Save fixed width files	Jeston

DBI

Expand panels | Show all details | Hide met & N/A

Projects that follow the best practices below can voluntarily self-certify and show that they've achieved a Core Infrastructure Initiative (CII) badge. Show details

If this is your project, please show your badge status on your project page! The badge status looks like this: cii best practices passing Here is how to embed it: Show details

These are the passing level criteria. You can also view the silver or gold level criteria.

- Basics 12/12
- Change Control 9/9
- Reporting 8/8
- Quality 13/13
- Security 16/16

Secure development knowledge

Met
Unmet
?

The project MUST have at least one primary developer who knows how to design secure software. (See 'details' for the exact requirements.) [know_secure_design] Show details

Quoting identifiers and literals for SQL queries to avoid SQL injection is a key concept of DBI.

DBI

Expand panels | Show all details | Hide met & N/A

Projects that follow the best practices below can voluntarily self-certify and show that they've achieved a Core Infrastructure Initiative (CII) badge. Show details

If this is your project, please show your badge status on your project page! The badge status looks like this: cii best practices passing Here is how to embed it: Show details

These are the passing level criteria. You can also view the silver or gold level criteria.

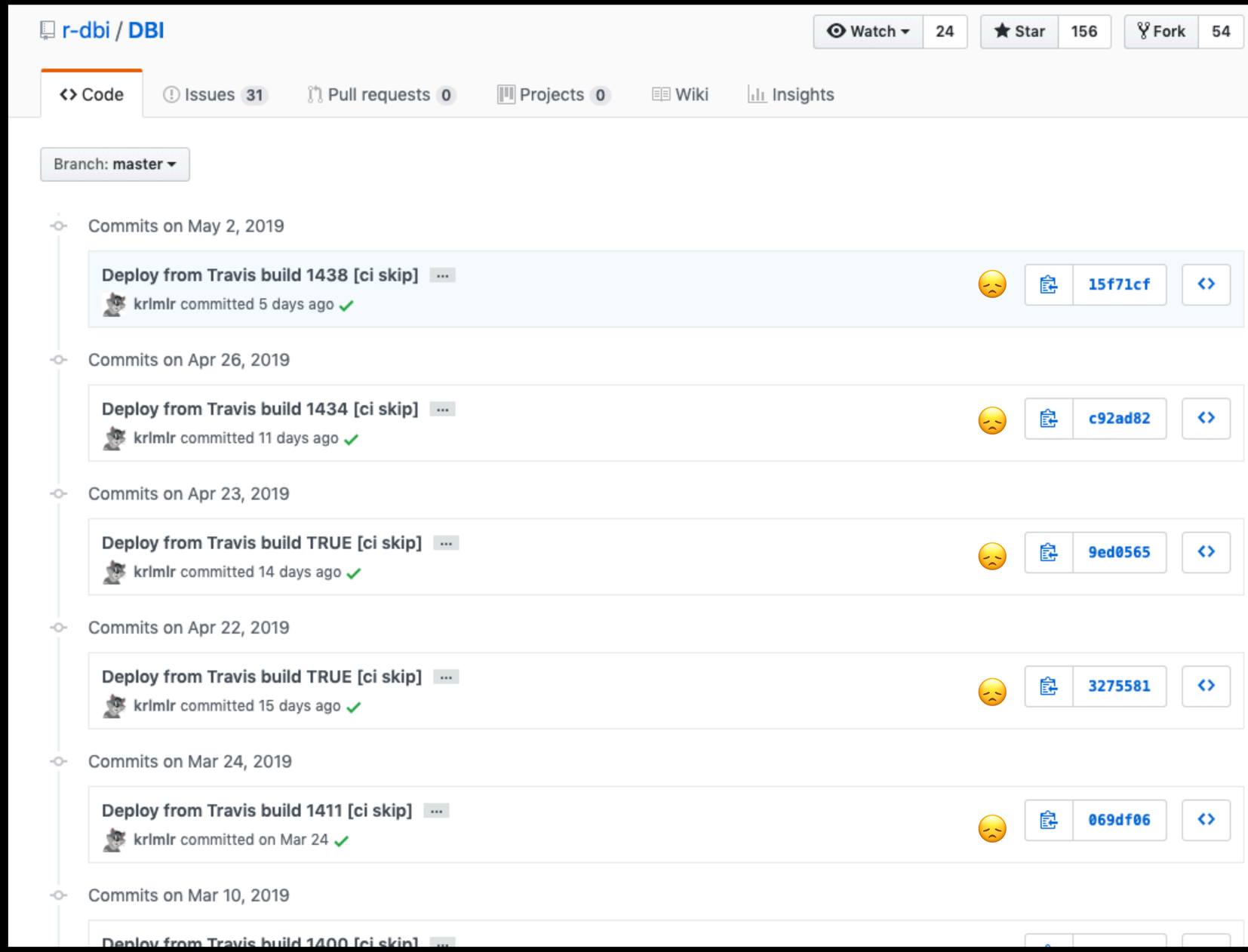
- Basics 12/12
- Change Control 9/9
- Reporting 8/8
- Quality 13/13
- Security 16/16

Secure development knowledge

Met / Unmet / ?

The project MUST have at least one primary developer who knows how to design secure software. (See 'details' for the exact requirements.) [know_secure_design] Show details

Quoting identifiers and literals for SQL queries to avoid SQL injection is a key concept of DBI.


r-dbi / DBI
Watch 24
Star 156
Fork 54
Code
Issues 31
Pull requests 0
Projects 0
Wiki
Insights
Branch: master
Commits on May 2, 2019
Deploy from Travis build 1438 [ci skip]
krlmlr committed 5 days ago
15f71cf
Commits on Apr 26, 2019
Deploy from Travis build 1434 [ci skip]
krlmlr committed 11 days ago
c92ad82
Commits on Apr 23, 2019
Deploy from Travis build TRUE [ci skip]
krlmlr committed 14 days ago
9ed0565
Commits on Apr 22, 2019
Deploy from Travis build TRUE [ci skip]
krlmlr committed 15 days ago
3275581
Commits on Mar 24, 2019
Deploy from Travis build 1411 [ci skip]
krlmlr committed on Mar 24
069df06
Commits on Mar 10, 2019

- Where do you source your packages from?

  Consider using **{drat}** ,RStudio's fancy new pro-pkg manager, or your own tooling around `tools:*_PACKAGES()` to setup your own internal CRAN and internal-org package repository that have verified "OK" versions of any and all packages you depend on (including ones from external source-only repositories).

  Further consider using file integrity monitoring to ensure no unexpected changes have occurred.

- Are you *sure* that's the same data file?

```
tools::md5sum("~/Data/critical-data.csv")
/Users/hrbrmstr/Data/critical-data.csv
"110524e0c5e7ab188a10090e9a7960e5"
```

Monday

```
tools::md5sum("~/Data/critical-data.csv")
/Users/hrbrmstr/Data/critical-data.csv
"c1682d837829a84bc411864dbfec0762"
```

Thursday

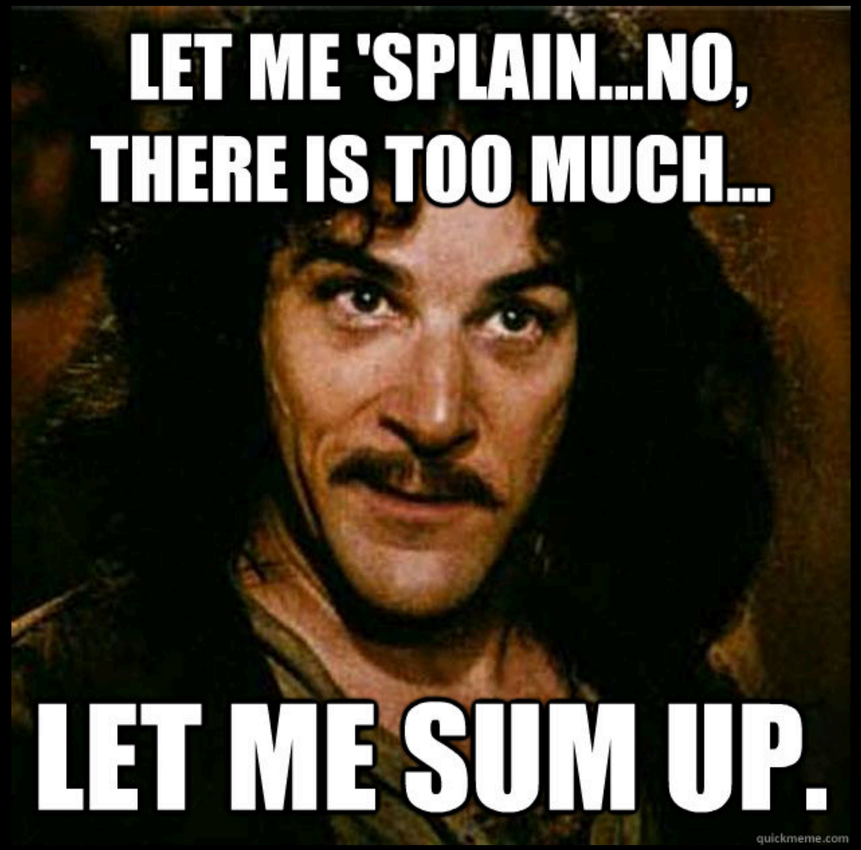
- Are you *sure* that's the same data file?

```
tools::md5sum("~/Data/critical-data.csv")
/Users/hrbrmstr/Data/critical-data.csv
"110524e0c5e7ab188a10090e9a7960e5"
```

Monday

```
tools::md5sum("~/Data/critical-data.csv")
/Users/hrbrmstr/Data/critical-data.csv
"c1682d837829a84bc411864dbfec0762"
```

Thursday


LET ME 'SPLAIN...NO,
THERE IS TOO MUCH...
LET ME SUM UP.
quickmeme.com


STAY
WOKE

- Integrity (and, perhaps, Confidentiality) are functional properties of many, many (many) parts of the data analysis workflow.

- **You must be aware of the status of each functional component** and ensure it meets the requirements of the project you're working on to have Integrity (and/or Confidentiality) be an emergent property of your *entire analytics workflow.*

благодаря

ευχαριστώ

ありがとう

Շնորհակալություն

감사합니다

Дякую

(thank you)

@hrbrmstr • bob@rud.is • bob_rudis@rapid7.com