FRED HUTCH

CURES START HERE®

Easy Update

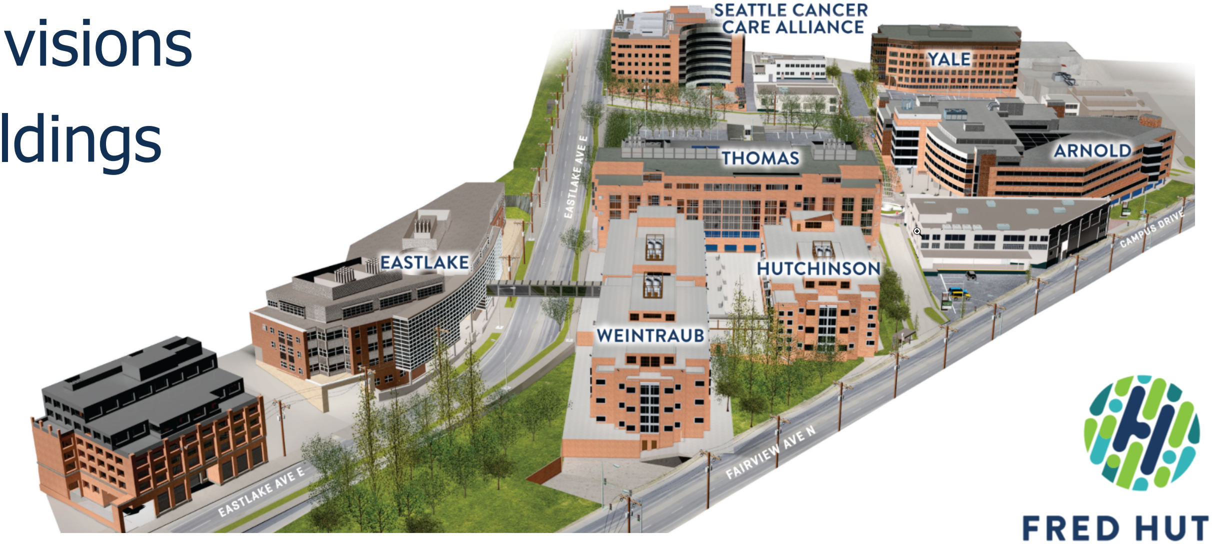
4th EasyBuild Users Conference Jan 30, 2019

John Dey jfdey@fredhutch.org

Fred Hutch Cancer Research Center

About Fred Hutch

- 3 Nobel Laureates
- Cancer & HIV Research
- 3200 Staff
- 240 Faculty
- $500M Budget (71% Grants/Contracts)
- 5 Scientific Divisions
- 1.5M Sqft buildings



FRED HUTCH
CURES START HERE®

Fred Hutch and R

The Fred Hutch has hundreds of R users. Robert Gentleman who is one of the co-developers of R worked at the Hutch while still actively developing the language. Gentleman now works at 23andMe.

Bioconductor originated at the Fred Hutchinson Cancer Research Center in the Fall of 2001. Martin Morgan led Bioconductor development since 2008 joined the Fred Hutch in 2005. The BioConductor group has since moved to Roswell Park Cancer Center.

R is the predominant language used by BioInformatics staff at the Hutch. Our local R build has over 800 modules. Our HPC group (me) has to devote considerable effort to maintain R for the center.

EasyUpdate EasyAnnotate

Tools for Support maintenance of EasyConfigs

Update the version information within EasyConfigs for modules specified with exts_list

Support R, BioConductor, Python. Perl is under development

Easy Annotate document module list for packages

[Fred Hutch Site Software]

Motivation for Easy Update

Major Release of Python annually, multiple updates during the year and 400 packages

Supporting two versions of Python 2.7x and 3.6x

Four Releases of R annually with >800 libraries

Checking, Updating library versions is no longer practical

Dependency ordering adds to the complexity

Easy Update for Perl has not been checked-in

Version Information

R and Python API via Python <requests> library

- R meta data is very reliable
- Python metadata not always correct
- Pypi.Org has improvement in metadata accuracy

BioConductor – No API! Download JSON database

- Metadata is always correct

Perl – No API for Module Metadata

- Require recent full Perl install to access CPAN::Meta
- CPAN in Elastic Search
    - Use MetaCPAN::Client(); - Has method .latest

Module Dependencies

Recursively follow dependences – Re Order exts_list

R – check metadata for: Depends, Imports, LinkingTo

- Check if ‘Part of R’

Python – “requires field”

- MetaData not consistent – package name, project name, import name can differ
- Requires library name does not match the Project name
- Underscore vs. Dash - Mismatch between Python Syntax and URLs
- MetaData can be incomplete, dependency data is messy

Easy Update Dependency Checking

Dependency Checking Single Library

```
./easy_update.py --rver 3.5 --search rstan
BioCondutor verserion: biocver not set
    ('Rcpp', '1.0.0', ext_options),
    ('BH', '1.69.0-1', ext_options),
    ('lattice', '0.20-38', ext_options),
    ('Matrix', '1.2-15', ext_options),
    ('RcppEigen', '0.3.3.5.0', ext_options),
    ('StanHeaders', '2.18.0-1', ext_options),
    ('gtable', '0.2.0', ext_options),
    ('withr', '2.1.2', ext_options),
    ('R6', '2.3.0', ext_options),
    ('viridisLite', '0.3.0', ext_options),
    ('labeling', '0.3', ext_options),
    ('RColorBrewer', '1.1-2', ext_options),
    ('colorspace', '1.4-0', ext_options),
    ('munsell', '0.5.0', ext_options),
    ('scales', '1.0.0', ext_options),
    ('rlang', '0.3.1', ext_options),
    ('lazyeval', '0.2.1', ext_options),
    ('plyr', '1.8.4', ext_options),
    ('magrittr', '1.5', ext_options),
    ('glue', '1.3.0', ext_options),
    ('stringi', '1.2.4', ext_options),
    ('stringr', '1.3.1', ext_options),
    ('reshape2', '1.4.3', ext_options),
    ('MASS', '7.3-51.1', ext_options),
    ('assertthat', '0.2.0', ext_options),
    ('crayon', '1.3.4', ext_options),
    ('cli', '1.0.1', ext_options),
    ('utf8', '1.1.4', ext_options),
    ('fansi', '0.4.0', ext_options),
    ('pillar', '1.3.1', ext_options),
    ('pkgconfig', '2.0.2', ext_options),
    ('tibble', '2.0.1', ext_options),
    ('nlme', '3.1-137', ext_options),
    ('mgcv', '1.8-26', ext_options),
```

```
('digest', '0.6.18', ext_options),
('ggplot2', '3.1.0', ext_options),
('prettyunits', '1.0.2', ext_options),
('ps', '1.3.0', ext_options),
('processx', '3.2.1', ext_options),
('callr', '3.1.1', ext_options),
('backports', '1.1.3', ext_options),
('rprojroot', '1.3-2', ext_options),
('desc', '1.2.0', ext_options),
('pkgbuild', '1.0.2', ext_options),
('matrixStats', '0.54.0', ext_options),
('loo', '2.0.0', ext_options),
('inline', '0.3.15', ext_options),
('gridExtra', '2.3', ext_options),
('rstan', '2.18.2', ext_options),
```

How to Increase Automation

How to update 150 Bio EasyConfigs?

Can Automation be used to update EasyConfigs?

Git projects - No consistent way to extract version information

- tag, release, master.zip
- Create Issue if project does not have tags or releases

Linux from scratch

Source Forge

Vendor supported products – Nanopore

Community Supported EasyConfigs

Cross platform support for Centos and Ubuntu

Contribute where possible – Increase community support

- Many commercial users in US do not contribute

Create Generic EasyBuilds for community

Use Bundles to support site customization

Wanted List ?

OS upgrade project – Rebuild 3 years of packages with new toolchain

Hutch Cluster Update Project

Ubuntu 14.04 -> 18.04 Toolchain 2016b -> 2018b

Re-Compile hundreds modules

138 Bio modules currently installed

Log all Lmod load commands to Syslog -> Splunk

Query all modules loads by User and Package

Summary: 137 Module Class ‘bio’ that have been used during 2018

51,337,445 Module loads - 1,362,483 Bio Module

Thank you

FRED HUTCH
CURES START HERE®