Renjin: The Road to Compatibility and Performance

Alexander Bertram
BeDataDriven

DALI 2013 – Indianapolis

What is Renjin?

- A new interpreter for the R language running on the (vanilla) JVM
- Core bits written in Java
- R Language libraries (base, stats, etc) reused whole

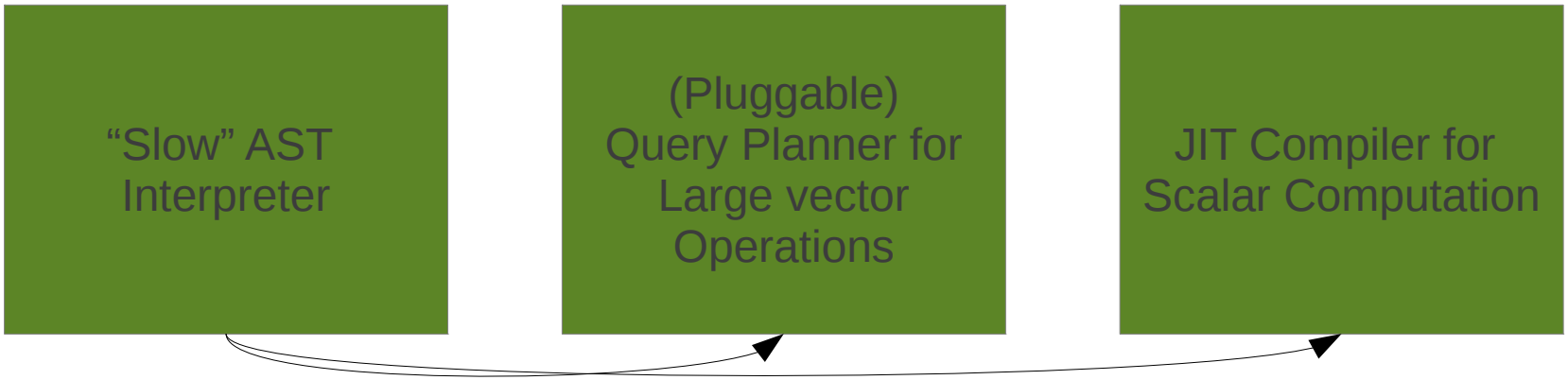
Why?

- R package ecosystem indispensible to our work @BeDataDriven
- But wanted to move faster from model prototype to “production”:
  - Seamless integration with the rest of the technology stack: databases, web servers, PaaS
  - Better performance, out of memory datasets

Highlights

- Runs 20-50% of CRAN packages (depending on measure of completeness)
- Implicit parallelization
- C/Fortran tool-chain to compile native parts of packages to JVM byte code
- JIT compiler to compile a subset of the language to JVM byte code

Design



Abstracts away vector storage, file system, threading, etc

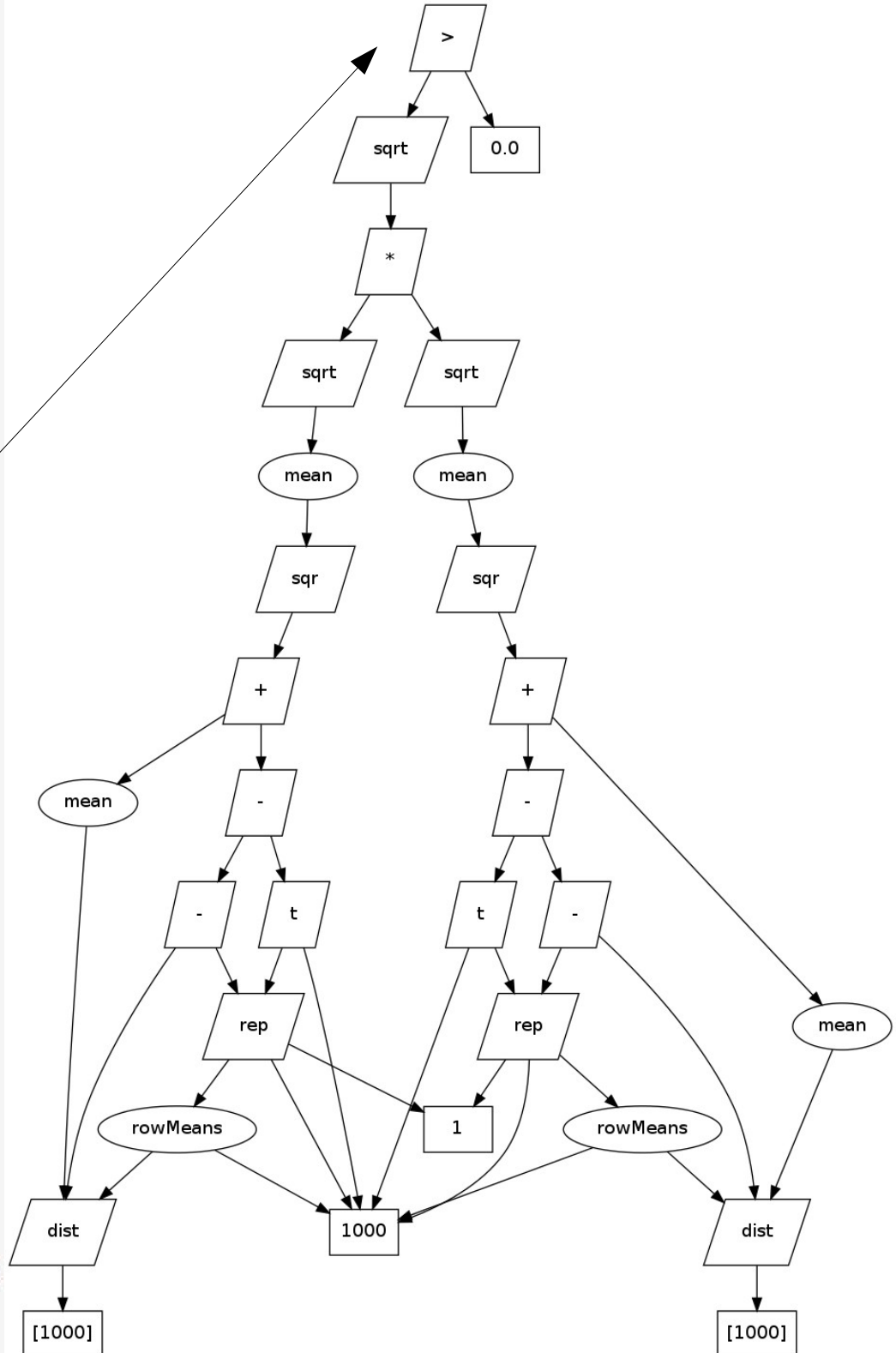
We wrote a C/Fortran compiler along the way to deal with native code in packages

```
function (x, y, index = 1)
{
 x <- dist(x)
 y <- dist(y)
  x <- as.matrix(x)
  y <- as.matrix(y)
  n <- nrow(x)
  m <- nrow(y)
  dims <- c(n, ncol(x), ncol(y))
  Akl <- function(x) {
    d <- as.matrix(x)^index
    m <- rowMeans(d)
    M <- mean(d)
    a <- sweep(d, 1, m)
    b <- sweep(a, 2, m)
    return(b + M)
  }
  A <- Akl(x)
  B <- Akl(y)
  dCov <- sqrt(mean(A * B))
  dVarX <- sqrt(mean(A * A))
  dVarY <- sqrt(mean(B * B))
  V <- sqrt(dVarX * dVarY)
  if (V > 0)
    dCor <- dCov/V
  else dCor <- 0
    return(list(dCov = dCov, dCor = dCor, dVarX = dVarX, dVarY = dVarY))
}
```

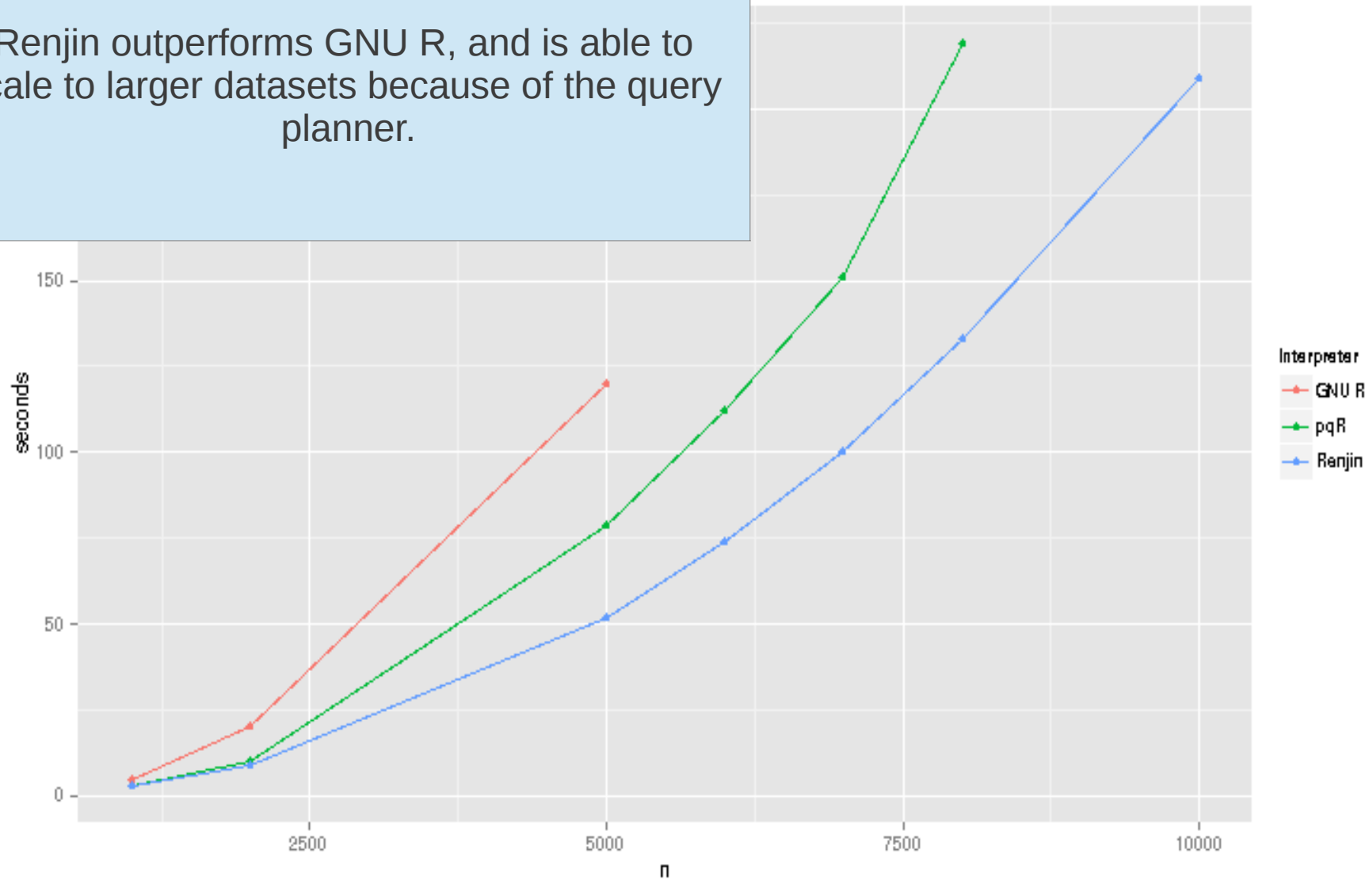
DCOR() from Energy Package
Real world code with O(n^2) memory requirements in GNU R because of distance matrix.

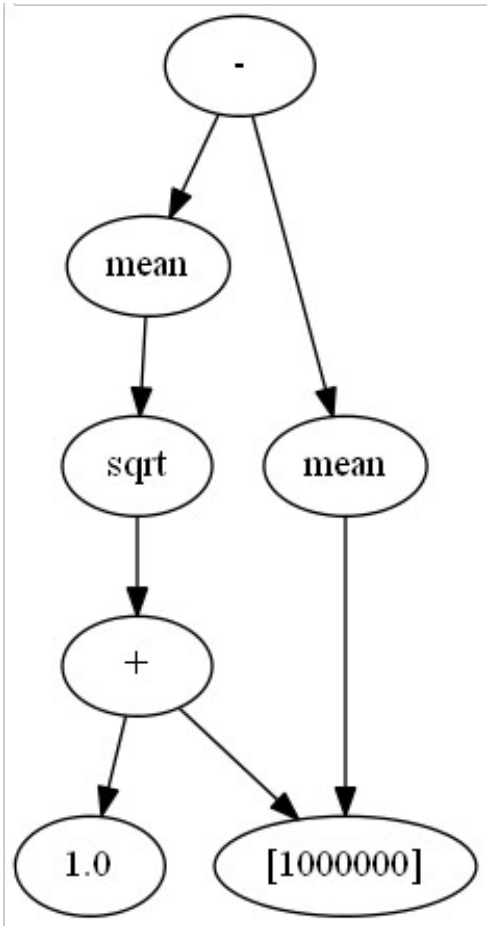
Renjin defers computation as long as possible, using deferred “views” for large data structures to avoid memory allocation

```r
function (x, y, index = 1)
{
 x <- dist(x)
 y <- dist(y)
  x <- as.matrix(x)
  y <- as.matrix(y)
  n <- nrow(x)
  m <- nrow(y)
  dims <- c(n, ncol(x), ncol(y))
  Akl <- function(x) {
    d <- as.matrix(x)^index
    m <- rowMeans(d)
    M <- mean(d)
    a <- sweep(d, 1, m)
    b <- sweep(a, 2, m)
    return(b + M)
  }
  A <- Akl(x)
  B <- Akl(y)
  dCov <- sqrt(mean(A * B))
  dVarX <- sqrt(mean(A * A))
  dVarY <- sqrt(mean(B * B))
  V <- sqrt(dVarX * dVarY)
  if (V > 0)
    dCor <- dCov/V
  else dCor <- 0
    return(list(dCov = dCov, dCor = dCor, dVarX =
}
```



Renjin outperforms GNU R, and is able to scale to larger datasets because of the query planner.





Query Planner

- JVM-based JIT w/ loop fusion and implicit parallelism
- OpenGL JITTer
- SQL Translator?
- Hadoop Job Launcher?

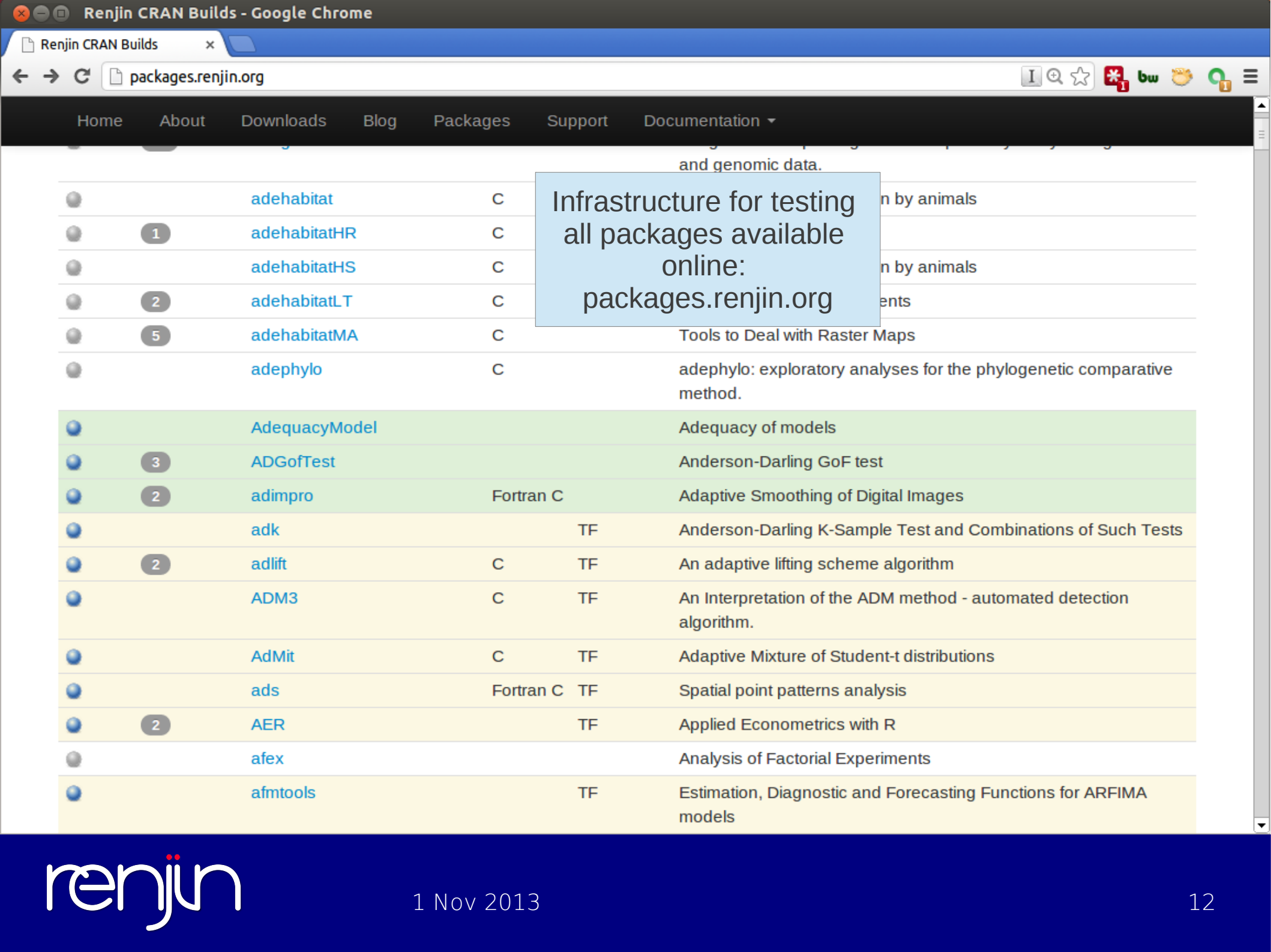
“Scalar” JIT for Eligible Code

- **GOOD:**
```
y <- sapply(x, function(x) x+1)
```
- **GOOD:**
```
s <- 0
for(i in 1:1e6) {
  s <- s + sqrt(i^2+(i-1)^2)
}
```
- **FORGET-ABOUT-IT: (defer to slow interperter)**
```
for(i in 1:1e6) {
  paste(“s <- s “, opName, “ i”)
}
```

Compatibility

	0.7.0-RC6	0.7.0-RC7
Total packages in CRAN	4602	?
Packages built	2456	?
Packages with some passing tests	831	?
Packages with all tests passing	236	?
Packages built, no tests	196	?

Huge number of fixes, implementation due in RC7 (Nov 2013)


Renjin CRAN Builds - Google Chrome
Renjin CRAN Builds
packages.renjin.org
Home
About
Downloads
Blog
Packages
Support
Documentation
and genomic data.
adehabitat C n by animals
1 adehabitatHR C
adehabitatHS C n by animals
2 adehabitatLT C ents
5 adehabitatMA C Tools to Deal with Raster Maps
adephylo C adephylo: exploratory analyses for the phylogenetic comparative method.
AdequacyModel Adequacy of models
3 ADGofTest Anderson-Darling GoF test
2 adimpro Fortran C Adaptive Smoothing of Digital Images
adk TF Anderson-Darling K-Sample Test and Combinations of Such Tests
2 adlift C TF An adaptive lifting scheme algorithm
ADM3 C TF An Interpretation of the ADM method - automated detection algorithm.
AdMit C TF Adaptive Mixture of Student-t distributions
ads Fortran C TF Spatial point patterns analysis
2 AER TF Applied Econometrics with R
afex Analysis of Factorial Experiments
afmtools TF Estimation, Diagnostic and Forecasting Functions for ARFIMA models
Infrastructure for testing all packages available online: packages.renjin.org

Where does Renjin sit?

- **Compared to FastR**: Renjin is prioritizing completeness first, performance next
- **Compared to dlyr**: similar ideas, but trying to bring transparently to all R functionality/packages
- **Compared to StochSS**: AppEngine/AppScale is also a big focus in terms of deployment for us!

Where to find information

- Main website at http://www.renjin.org for downloads, documentation and blog
- GitHub at https://github.com/bedatadriven/renjin for the source code and the issue tracker
- Google group (i.e. public mailing list) for developers (and users) at http://groups.google.com/group/renjin-dev