



# Tessera

Open source environment for deep analysis of large complex data

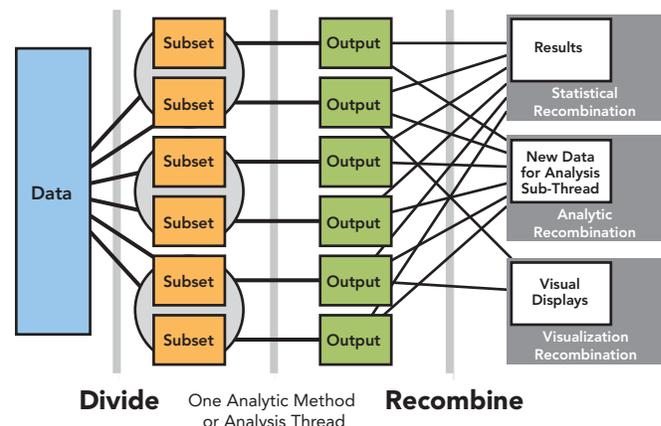
*Tessera extends the capability of the R environment by enabling rapid exploration of large, complex data sets. With Tessera, you can interactively perform exploratory data analysis, create large-scale, customized visualizations, and develop and test statistical models and machine learning algorithms. Tessera automatically manages the complicated tasks of distributed storage and computation, empowering data scientists to do what they do best: achieve research and mission objectives by deriving insight from data.*

*With Tessera, you can apply the thousands of statistical and visualization methods in the R language with simple commands over a back end like Hadoop – without being an expert in distributed computing. At the front end, you program in R. At the back end is a distributed parallel computational environment such as Hadoop that runs R code. In between are the Tessera packages **datadr** and **trelliscope**, which make it easy to communicate with the back end with simple R commands.*

*Tessera can be installed in a variety of configurations: on your local workstation, as a virtual machine, on the cloud, or on your own cluster. For all of these configurations, the user interface remains the same.*

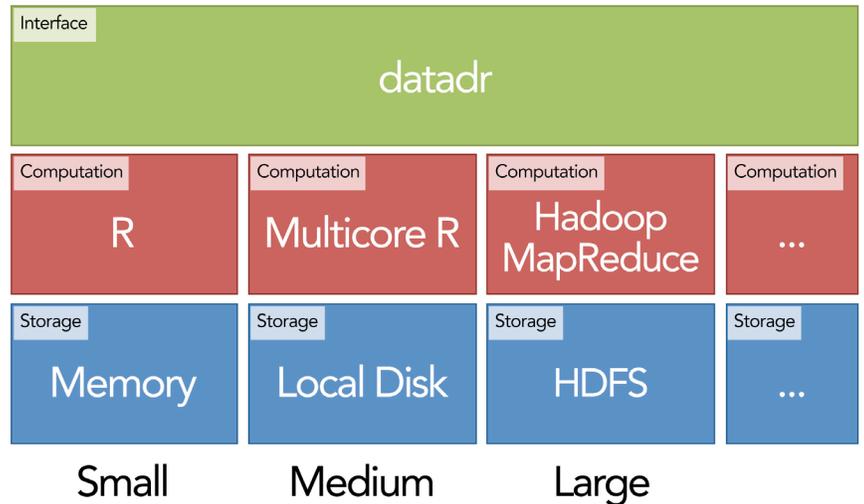
## Divide and Recombine (D&R)

Tessera is powered by Divide and Recombine. In D&R, we seek meaningful ways to divide the data into subsets, apply statistical methods to each subset independently, and recombine the results of those computations in a statistically valid way. This enables us to use the existing vast library of methods available in R – no need to write scalable versions of the code.



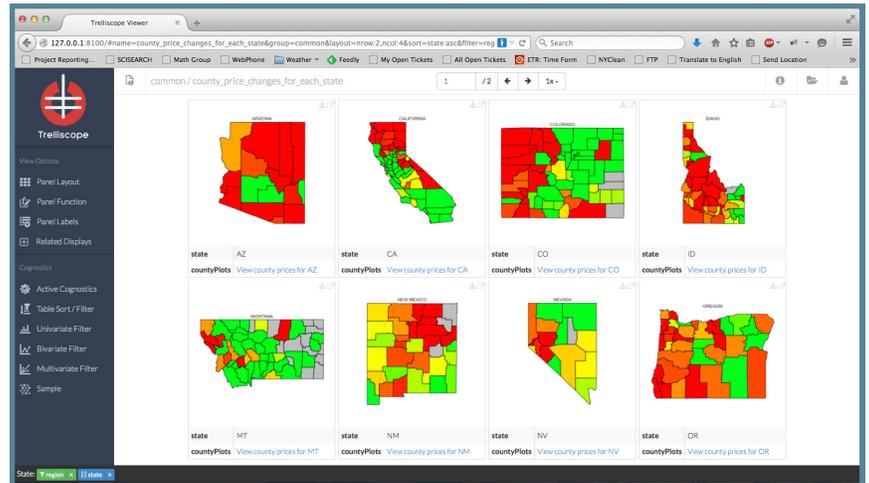
## datadr

The **datadr** package provides a simple interface to D&R operations. The interface is back end agnostic, so that as new distributed computing technology comes along, **datadr** will be able to harness it. The **datadr** software currently supports in-memory, local disk/multicore, and Hadoop back ends, with experimental support for Apache™ Spark®. Regardless of the back end, coding is done entirely in R and data are represented as R objects.



## trelliscope

The **trelliscope** package is a D&R visualization tool based on Trellis display techniques that enables scalable, flexible, and detailed visualization of data. Trellis display has repeatedly proven itself as an effective approach to visualizing complex data. Backed by **datadr**, **trelliscope** scales Trellis display, allowing the analyst to break potentially very large data sets into many subsets, apply a visualization method to each subset, and then interactively sample, sort, and filter the panels of the display using customizable quantities of interest. The **trelliscope** tool is also useful for exploratory data analysis of small datasets, and it provides an intuitive web-based interface that allows anyone to visually explore the data, without requiring expertise in R.



### For More Information, Contact:

Landon Sego, Ph.D.  
Pacific Northwest National Laboratory  
landon.sego@pnnl.gov  
(509) 375-2753

Ryan Hafen, Ph.D.  
Hafen Consulting, LLC  
rhafen@gmail.com

<http://tessera.io>

*The Tessera team consists of researchers at Hafen Consulting, Pacific Northwest National Laboratory, the Purdue University Department of Statistics, Mozilla Corporation, and independent contributors.*

*Apache Spark, Spark, and Apache are trademarks of The Apache Software Foundation.*

  
**Pacific Northwest**  
NATIONAL LABORATORY

Proudly Operated by **Battelle** Since 1965