JPL-Caltech Virtual Summer School

Big Data Analytics

September 2 – 12, 2014

Ashish Mahabal
California Institute of Technology

R - I

JPL Jet Propulsion Laboratory Keck Institute for Space Studies Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology Caltech

Linear regression example

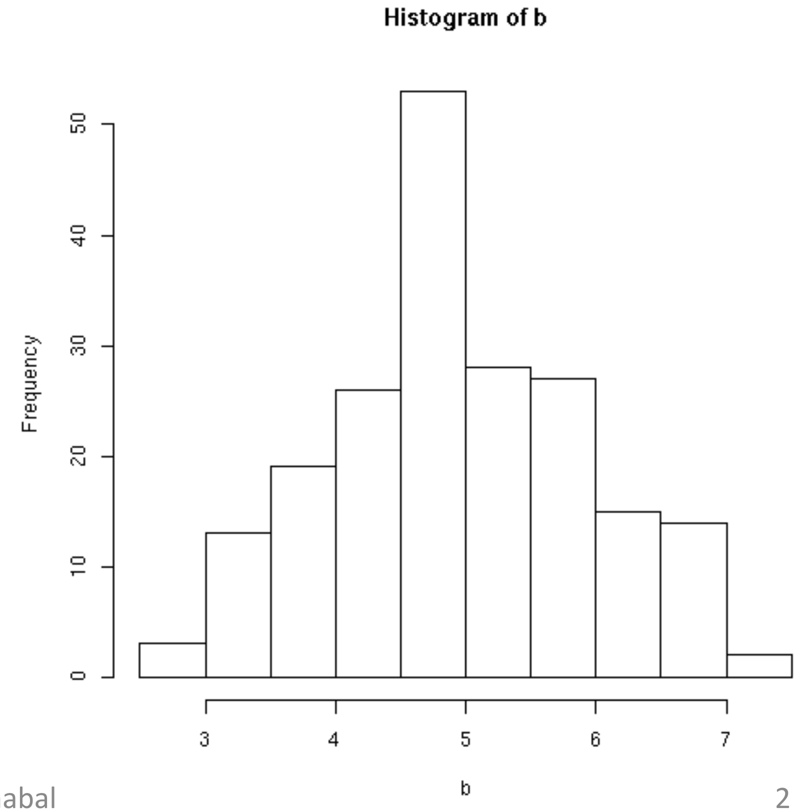
R

```
y <- c(1,2,4)              # assign; combine
x <- c(1,2,3)
foo <- lm(y~x)             # formula object
foo
    Call:
    lm(formula = y ~ x)

    Coefficients:
    (Intercept)           x
    -0.6667         1.5000
```

Example of a simple plot

```
a <- rnorm(100,mean=5,sd=1)
b <- rnorm(200,mean=5,sd=1)
hist(b)
c <- c(100*a,100*b)
length(c)
c
help(c)  # overloading
```



Statistics is extensively used

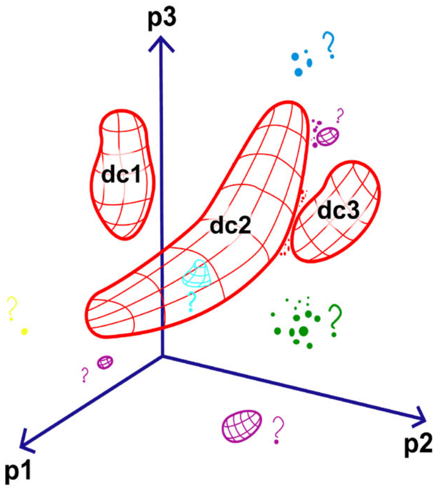
15000 astronomical studies per year

5% have “statistics” in their abstract

20% treat variable objects or multivariate datasets

Circa 2010

A Generic Machine-Assisted Discovery Problem:
Data Mapping and a Search for Outliers



Djorgovski

Limited number of methods still dominate

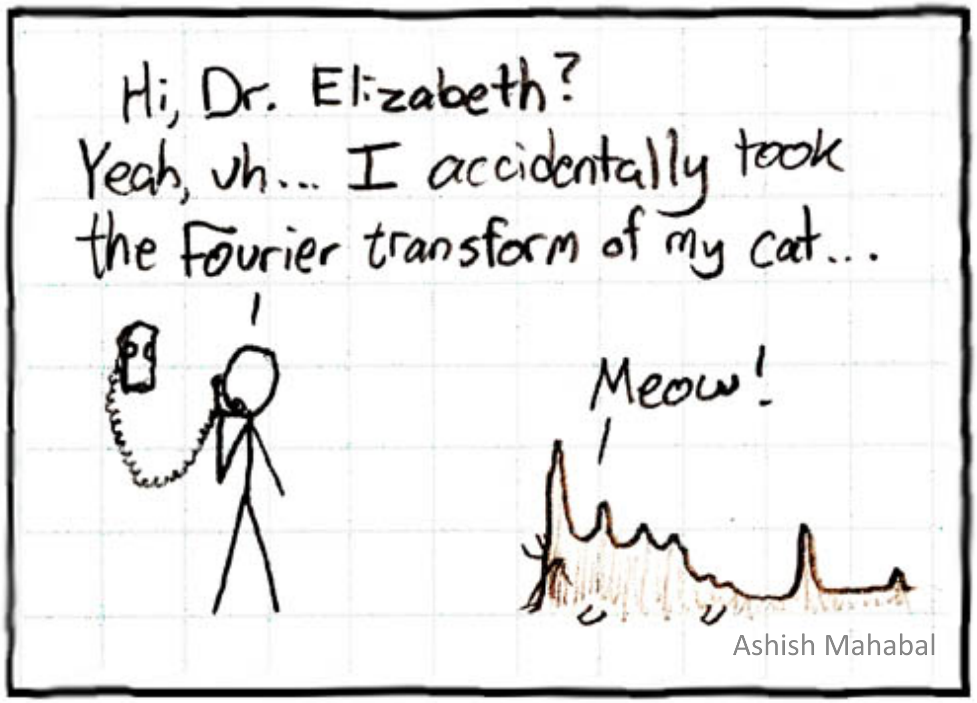
Traditional methods: preWWII

Fourier transform (Fourier 1807)

Least sq. and chisq (Legendre 1805, Pearson 1901)

Kolmogorov-Smirnov test (Kolomogrov 1933)

Principal Component Analysis (Hotelling 1936)


Hi, Dr. Elizabeth?
Yeah, uh... I accidentally took the Fourier transform of my cat...
Meow!

Xkcd/26

Advanced statistical methods are available in most systems

Matlab

Mathematica

IDL

Octave

NumPy

PDL

Why R

- Excellent for statistics (lots of modules)
  - 47% data-miners use R (Rexer's Annual Data Miner Survey in 2011; 1319 participants from 60 countries)
- Great layered graphics using ggplot
  - Generic graphics are (somewhat) clumsy
- Variety of GUIs and interfaces available
- Free

S and R

S: John Chambers (Bell Labs)

S-plus: 1988: Douglas Martin (UWash)

R: 1993: Ross Ihaka, Robert Gentleman

- Current version 3.1.1 (Jul 2014)
- Lexical scoping (ala Scheme)
- Procedural/functions
- Object Oriented
- Command line

R follows S

Linear and nonlinear modeling

Statistical tests

Time series analysis

Classification

Clustering

…

http://www.r-project.org/

(15 standard/recommended packages)

15 recommended packages

http://cran.r-project.org/src/contrib/3.0.2/Recommended/

- [ ] KernSmooth_2.23-10.tar.gz 19-Mar-2013 13:18 33K
- [ ] MASS_7.3-29.tar.gz 31-Aug-2013 19:29 474K
- [ ] Matrix_1.0-14.tar.gz 13-Sep-2013 13:16 1.6M
- [ ] boot_1.3-9.tar.gz 20-Mar-2013 08:28 221K
- [ ] class_7.3-9.tar.gz 21-Aug-2013 14:10 19K
- [ ] cluster_1.14.4.tar.gz 26-Mar-2013 16:50 247K
- [ ] codetools_0.2-8.tar.gz 15-Feb-2011 10:56 12K
- [ ] foreign_0.8-55.tar.gz 02-Sep-2013 14:33 321K
- [ ] lattice_0.20-23.tar.gz 21-Aug-2013 18:16 338K
- [ ] mgcv_1.7-26.tar.gz 06-Sep-2013 11:31 540K
- [ ] nlme_3.1-111.tar.gz 08-Sep-2013 12:40 736K
- [ ] nnet_7.3-7.tar.gz 01-Jul-2013 13:34 29K
- [ ] rpart_4.1-3.tar.gz 02-Sep-2013 07:19 798K
- [ ] spatial_7.3-7.tar.gz 01-Jul-2013 13:34 43K
- [ ] survival_2.37-4.tar.gz 27-Mar-2013 07:14 1.5M

Comprehensive R Archive Network

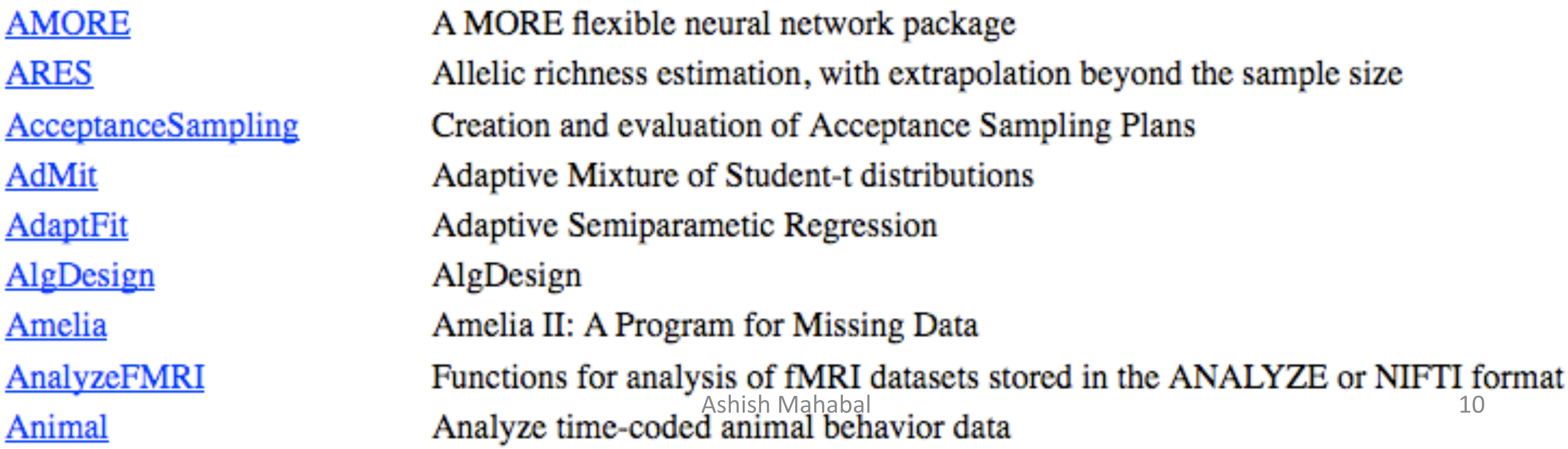
http://cran.r-project.org/,

http://www.bioconductor.org/

Over 5795 (8/2014) user contributed packages

Strength: people contributed

Weakness: organic growth – uniformity lost (e.g. plots)



AMORE	A MORE flexible neural network package
ARES	Allelic richness estimation, with extrapolation beyond the sample size
AcceptanceSampling	Creation and evaluation of Acceptance Sampling Plans
AdMit	Adaptive Mixture of Student-t distributions
AdaptFit	Adaptive Semiparametic Regression
AlgDesign	AlgDesign
Amelia	Amelia II: A Program for Missing Data
AnalyzeFMRI	Functions for analysis of fMRI datasets stored in the ANALYZE or NIFTI format
Animal	Analyze time-coded animal behavior data

Interfaces, editors etc.

http://rgl.neoscientists.org/about.shtml

(3D visualization with interface to R)

RapidMiner http://rapid-i.com/content/view/181/190/)

Weka (http://www.cs.waikato.ac.nz/ml/weka/)

has an R interface (RWeka)

http://www.sciviews.org/Tinn-R/

http://www.rforge.net/JGR/

http://www.rstudio.com/

Downloading and installing R (Current version: 3.1.1 – 7/2014)

- Download: http://cran.cnr.berkeley.edu/
- Do one of the following based on your OS
  - **Install on Mac**:

    http://cran.r-project.org/doc/manuals/R-admin.html#Installing-R-under-_0028Mac_0029-OS-X
  - **Install on Windows**:

    http://cran.r-project.org/doc/manuals/R-admin.html#Installing-R-under-Windows
  - **Install on other Unix-alikes**:

    http://cran.r-project.org/doc/manuals/R-admin.html#Installing-R-under-Unix_002dalikes

Running R

Create a subdir “R_work”

```
	PROMPT> mkdir R_work
	PROMPT> cd R_work
```

Start R

```
	PROMPT> R
```

Now you are in R

```
	R_PROMPT>
```

Do stuff

Quit

```
	R_PROMPT> q()
```

- Windows has GUI
- Create dir
- Start R
- Change dir
- Exit
- Save
- .Rdata created
- Double click

A few years ago …

VOStat

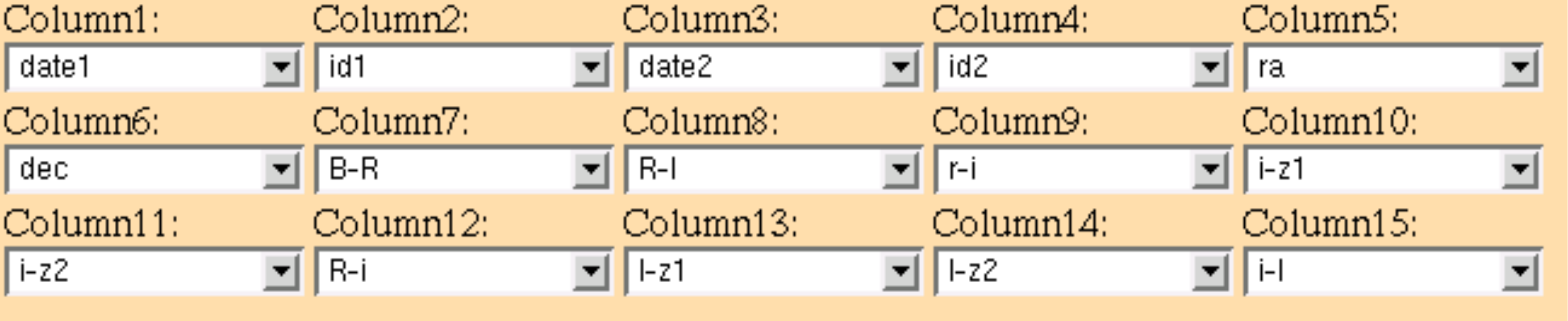
Columns are autoselected (and can be deselected)

Parameter choices for functions are conveniently placed

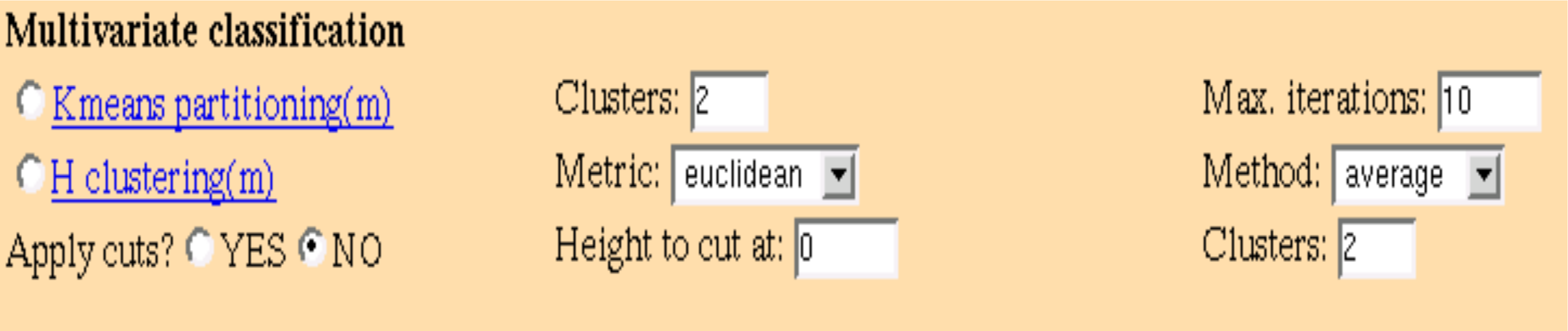
Can be used from your own webpages on tables residing elsewhere

Java/perl

ASCII/fits



Column1:	Column2:	Column3:	Column4:	Column5:
date1	id1	date2	id2	ra
Column6:	Column7:	Column8:	Column9:	Column10:
dec	B-R	R-I	r-i	i-z1
Column11:	Column12:	Column13:	Column14:	Column15:
i-z2	R-i	I-z1	I-z2	i-I



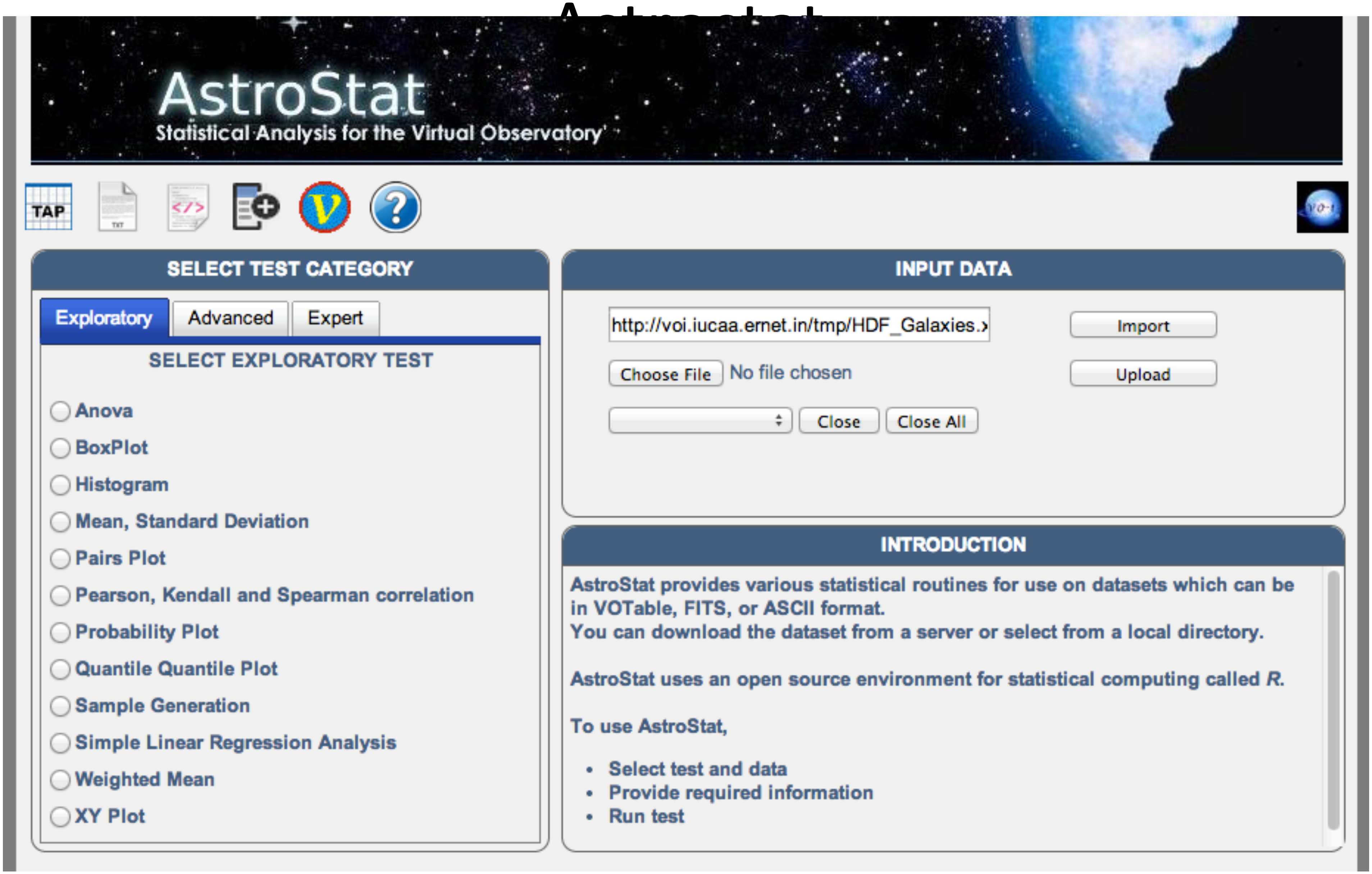
Multivariate classification

○ Kmeans partitioning(m) Clusters: 2 Max. iterations: 10

○ H clustering(m) Metric: euclidean Method: average

Apply cuts? ○ YES ● NO Height to cut at: 0 Clusters: 2

Astrostat



AstroStat
Statistical Analysis for the Virtual Observatory
TAP
TXT
</>

SELECT TEST CATEGORY

Exploratory | Advanced | Expert

SELECT EXPLORATORY TEST

- ○ Anova
- ○ BoxPlot
- ○ Histogram
- ○ Mean, Standard Deviation
- ○ Pairs Plot
- ○ Pearson, Kendall and Spearman correlation
- ○ Probability Plot
- ○ Quantile Quantile Plot
- ○ Sample Generation
- ○ Simple Linear Regression Analysis
- ○ Weighted Mean
- ○ XY Plot

INPUT DATA

http://voi.iucaa.ernet.in/tmp/HDF_Galaxies.x

No file chosen

Import
Choose File
Upload
Close
Close All

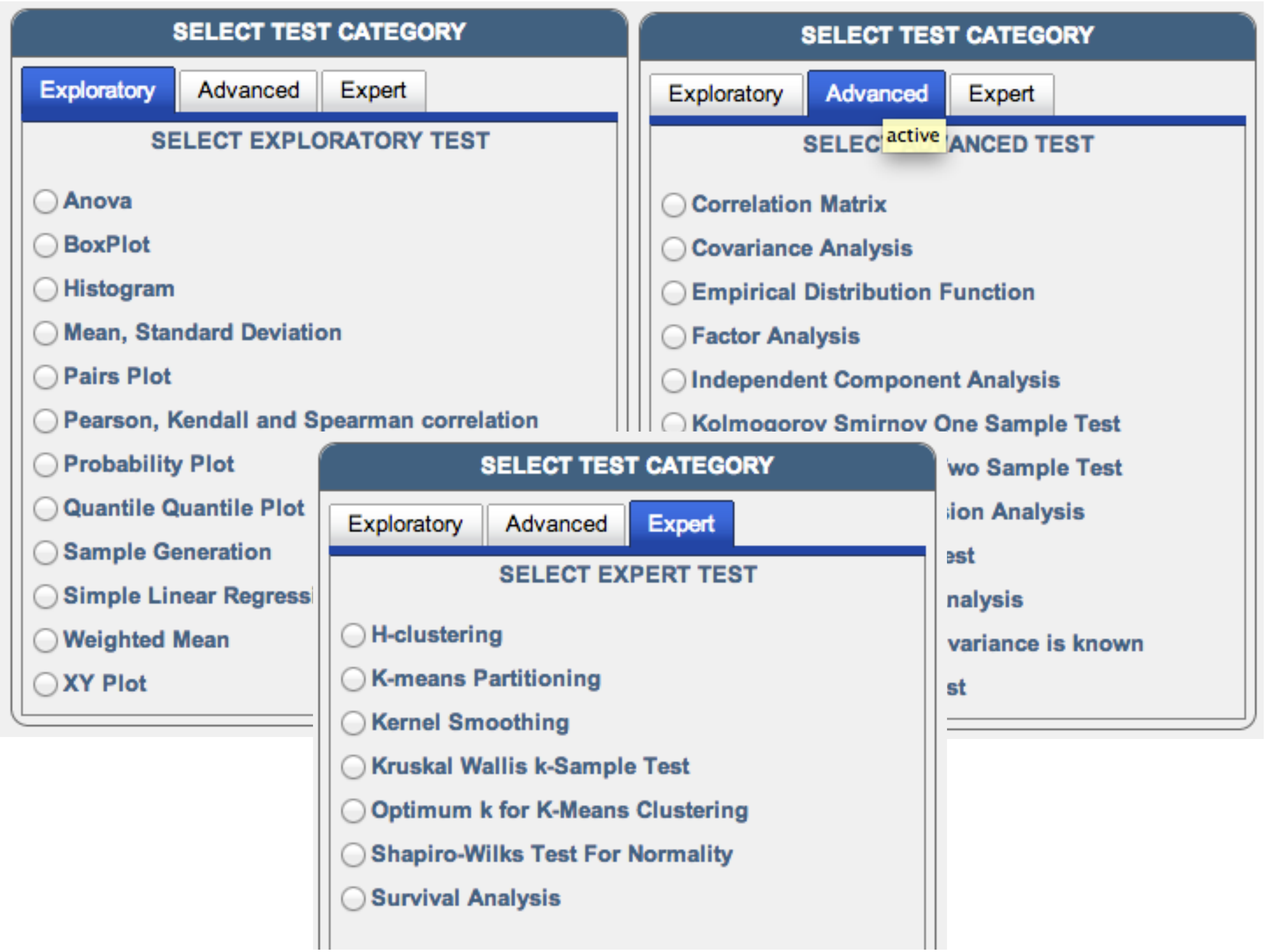
INTRODUCTION

AstroStat provides various statistical routines for use on datasets which can be in VOTable, FITS, or ASCII format.
You can download the dataset from a server or select from a local directory.

AstroStat uses an open source environment for statistical computing called *R*.

To use AstroStat,

- Select test and data
- Provide required information
- Run test



SELECT EXPLORATORY TEST

○ Anova
○ BoxPlot
○ Histogram
○ Mean, Standard Deviation
○ Pairs Plot
○ Pearson, Kendall and Spearman correlation
○ Probability Plot
○ Quantile Quantile Plot
○ Sample Generation
○ Simple Linear Regressi
○ Weighted Mean
○ XY Plot

SELEC ANCED TEST

○ Correlation Matrix
○ Covariance Analysis
○ Empirical Distribution Function
○ Factor Analysis
○ Independent Component Analysis
○ Kolmogorov Smirnov One Sample Test
wo Sample Test
ion Analysis
est
nalysis
variance is known
st

SELECT EXPERT TEST

○ H-clustering
○ K-means Partitioning
○ Kernel Smoothing
○ Kruskal Wallis k-Sample Test
○ Optimum k for K-Means Clustering
○ Shapiro-Wilks Test For Normality
○ Survival Analysis

Getting help

```
help(solve)
?search
help(“[[“)
help.start()          # this is for html help
??matrix
Sys.getenv(“R_HOME”)   # Case sensitive
Sys.getenv(c(“OS”,”R_HOME”))
summary(a)
```

Next time …

Assignments

Objects

Dataframes