Deloitte.

ROOT: A Data Mining Tool from CERN

Ravi Kumar and Arun Tripathi

CAS Predictive Modeling Seminar

San Diego

October 2008

Topics

What is ROOT ?

Why ROOT ?

Data Analysis and Modeling tools in ROOT.

What is ROOT ?

- An object oriented data analysis framework in C++, developed at CERN.
  - A particle physics lab located near Geneva, Switzerland
- Particle physics experiments produce vast amounts of very complex data.
  - Thousands of terabytes every year.
  - Handling and analyzing such data volumes ***efficiently*** is a major challenge.
- ROOT project was started at CERN in 1995 to address this challenge.
  - The LHC project at CERN is a $8 Billion, international project.
  - ROOT is the tool being used to analyze data from LHC.
  - Several other scientific projects around the globe use ROOT.
- ***So? What do physics solutions have to do with the "real world"?***

From Physics to “Real World”: An Example

➢ **The World Wide Web (the web) was invented at the same lab, CERN, in 1990. Below are some excerpts from the world’s first website.**

Welcome to info.cern.ch

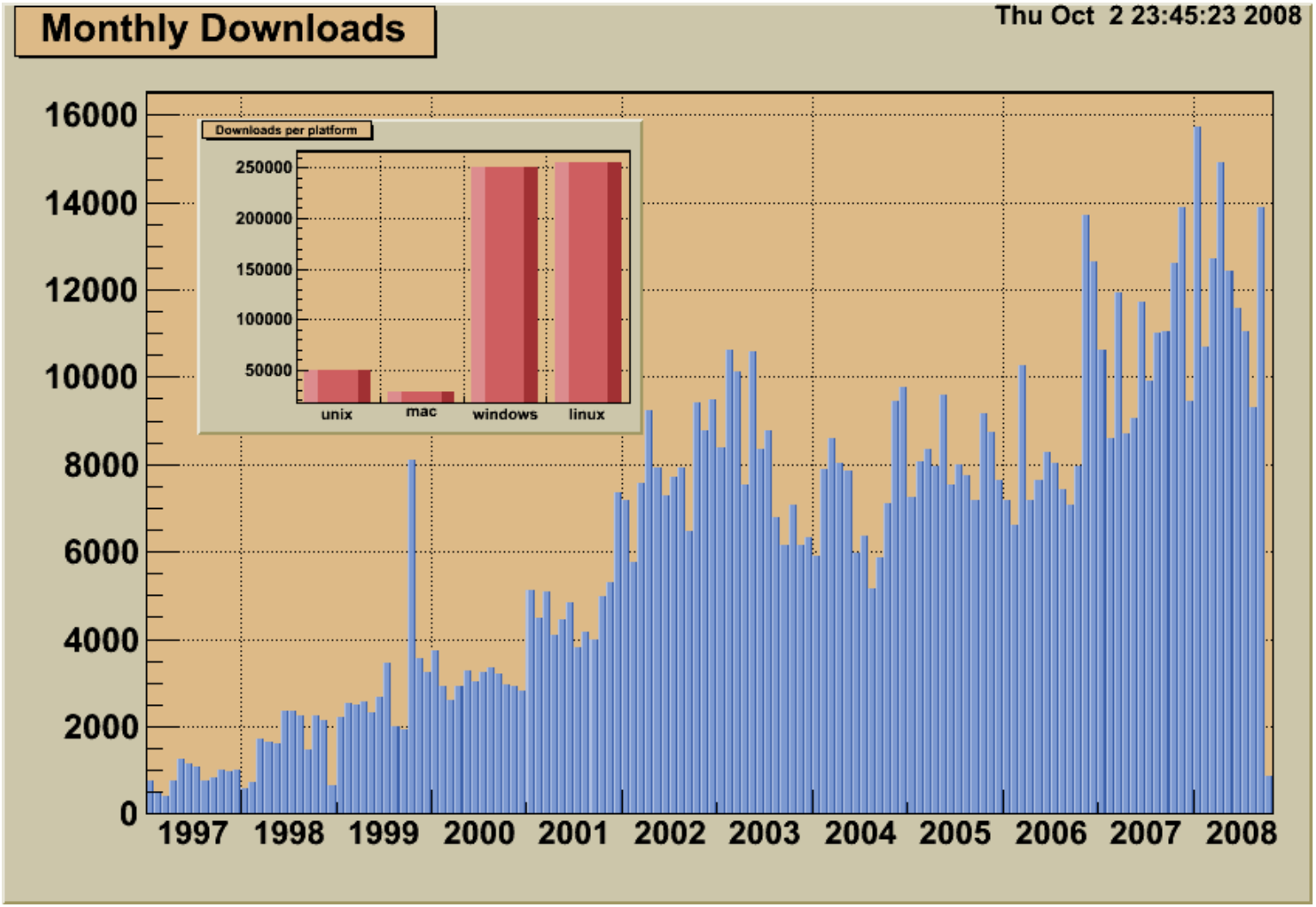
The website of the world's first-ever web server

1990 was a momentous year in world events. In February, Nelson Mandela was freed after 27 years in prison. In April, the space shuttle Discovery carried the Hubble Space Telescope into orbit. And in October, Germany was reunified.
Then at the end of 1990, a revolution took place that changed the way we live today.

Tim Berners-Lee followed his dream of a better, easier way to communicate via computers on a global scale, which led him to create the World Wide Web.

... and the rest is Web history.

Although the Web's conception began as a tool to aid physicists answer tough questions about the Universe, today its usage applies to various aspects of the global community and affects our daily lives.



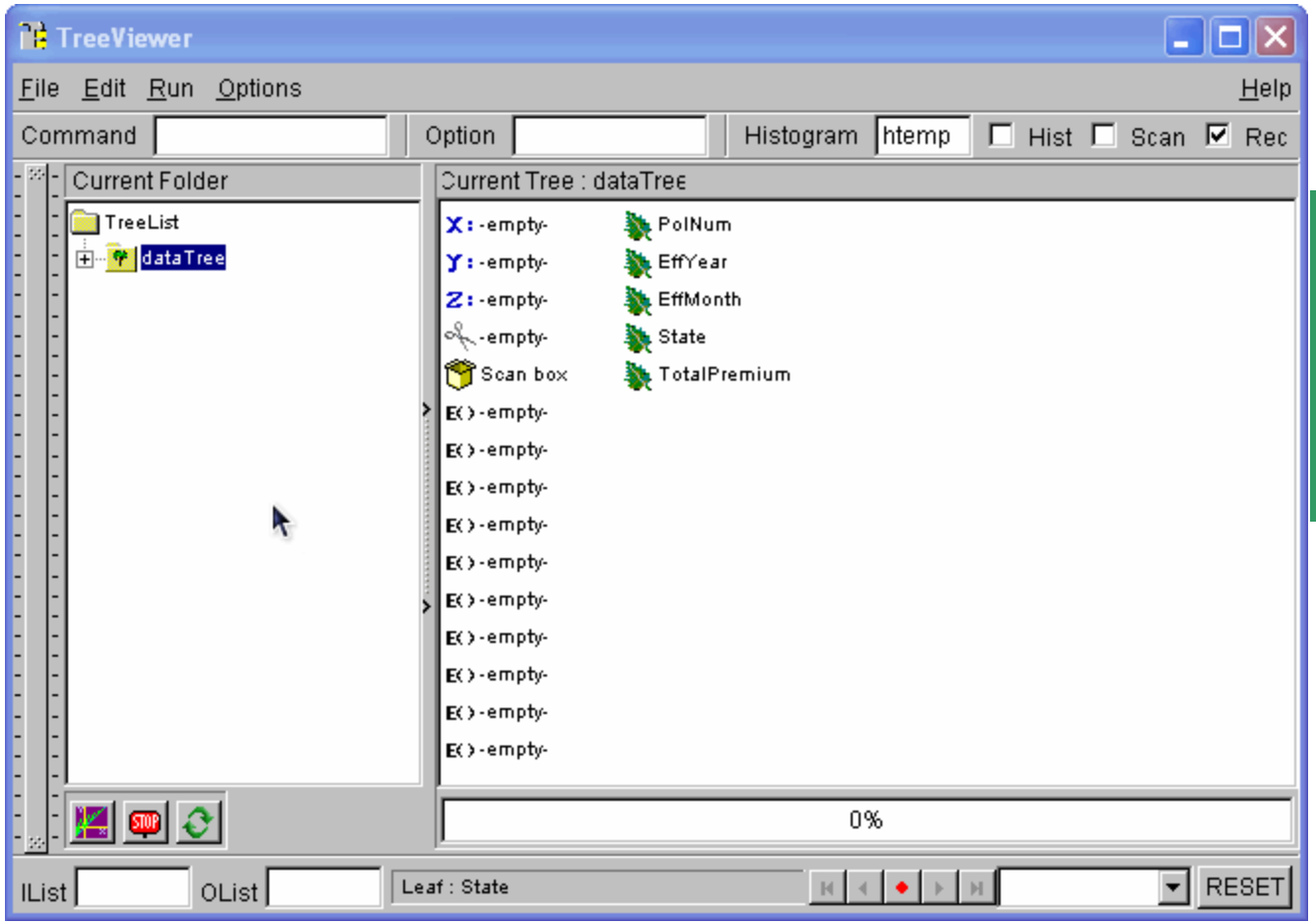
From: http://root.cern.ch/root/images/ftpstats2.gif

So, What is ROOT ?

- An object-oriented data analysis framework in C++, developed at CERN.
- Specifically designed to handle, store, and analyze large amounts of data *very efficiently*.
- Allows both highly interactive and batch mode analysis of data.
  - Data visualization and exploration tools.
  - Statistical modeling tools.
  - High quality graphics
  - Networking
  - Parallel processing.
- It is free, open-source and dynamically extensible.
  - Users can write and use their own classes with ROOT.
- Several operating systems are supported, including Windows, Linux, Unix, and Mac OS X.

Data Storage and Access in ROOT

- **ROOT stores data in a Tree format, with branches and leaves.**
- **Can store complex objects, arrays, images, in addition to simple data types.**
- **Optimized for reduced disk usage and increased access speed.**



An Example:

Data File Type	Size on Disk
ASCII	160 MB
SAS	108 MB
ROOT	28 MB

- **In this case, ROOT file takes only 18% space compared to the corresponding ASCII text file.**

Importance of Data Visualization

Some Data:

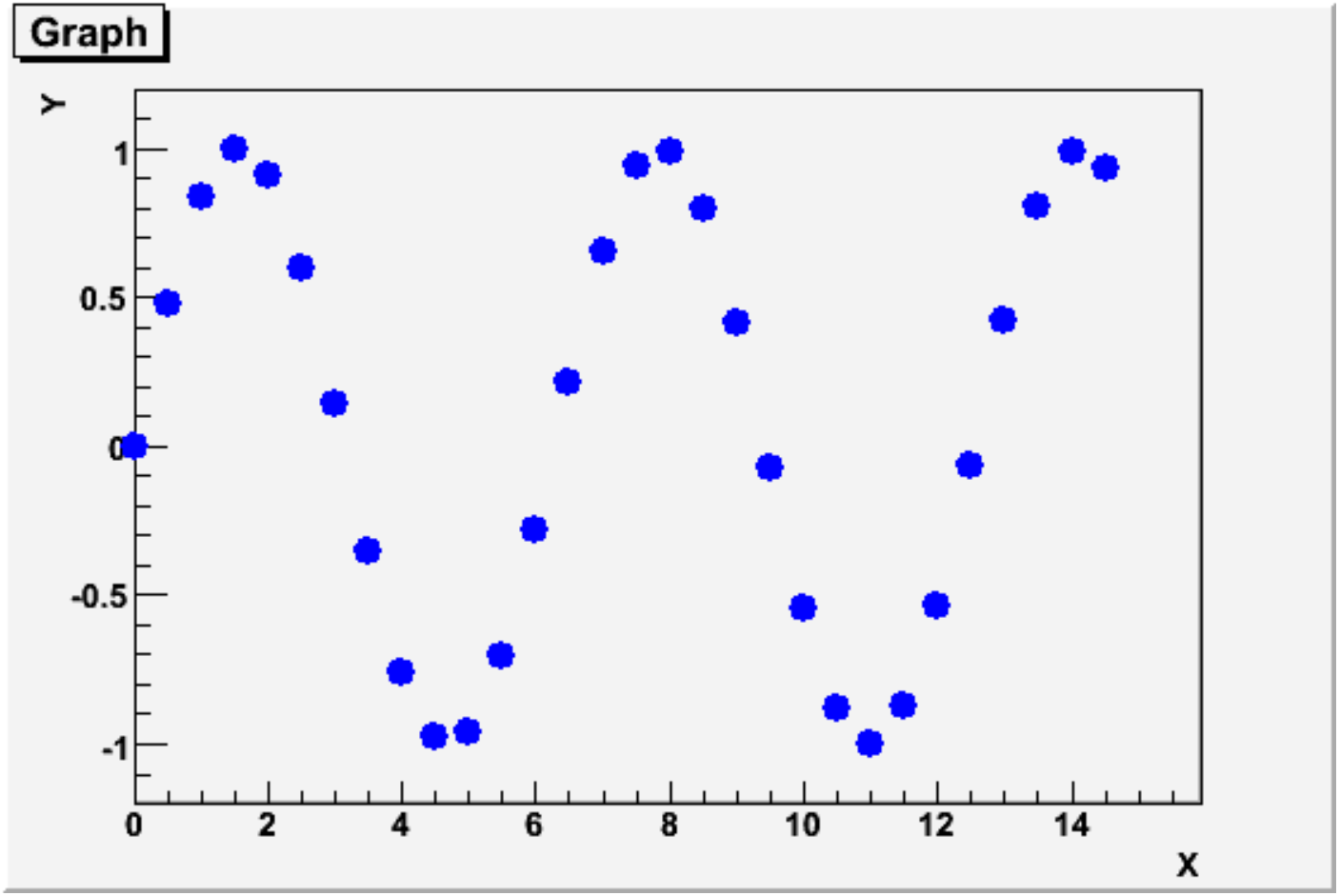
X	Y	X	Y
3	0.141	12.5	-0.066
5.5	-0.706	6.5	0.215
14.5	0.935	1.5	0.997
11.5	-0.875	7	0.657
9	0.412	8	0.989
13.5	0.804	0	0.000
8.5	0.798	4	-0.757
12	-0.537	10.5	-0.880
10	-0.544	14	0.991
7.5	0.938	3.5	-0.351
2	0.909	9.5	-0.075
11	-1.000	13	0.420
5	-0.959	6	-0.279
2.5	0.598	1	0.841
0.5	0.479	4.5	-0.978

Some Statistics About this Data:

	X	Y
Number of Data Points	30	30
Mean	7.25	0.104
Standard Deviation	4.33	0.705
Skewness	0.00205	-0.199
Kurtosis	-1.196	-1.5

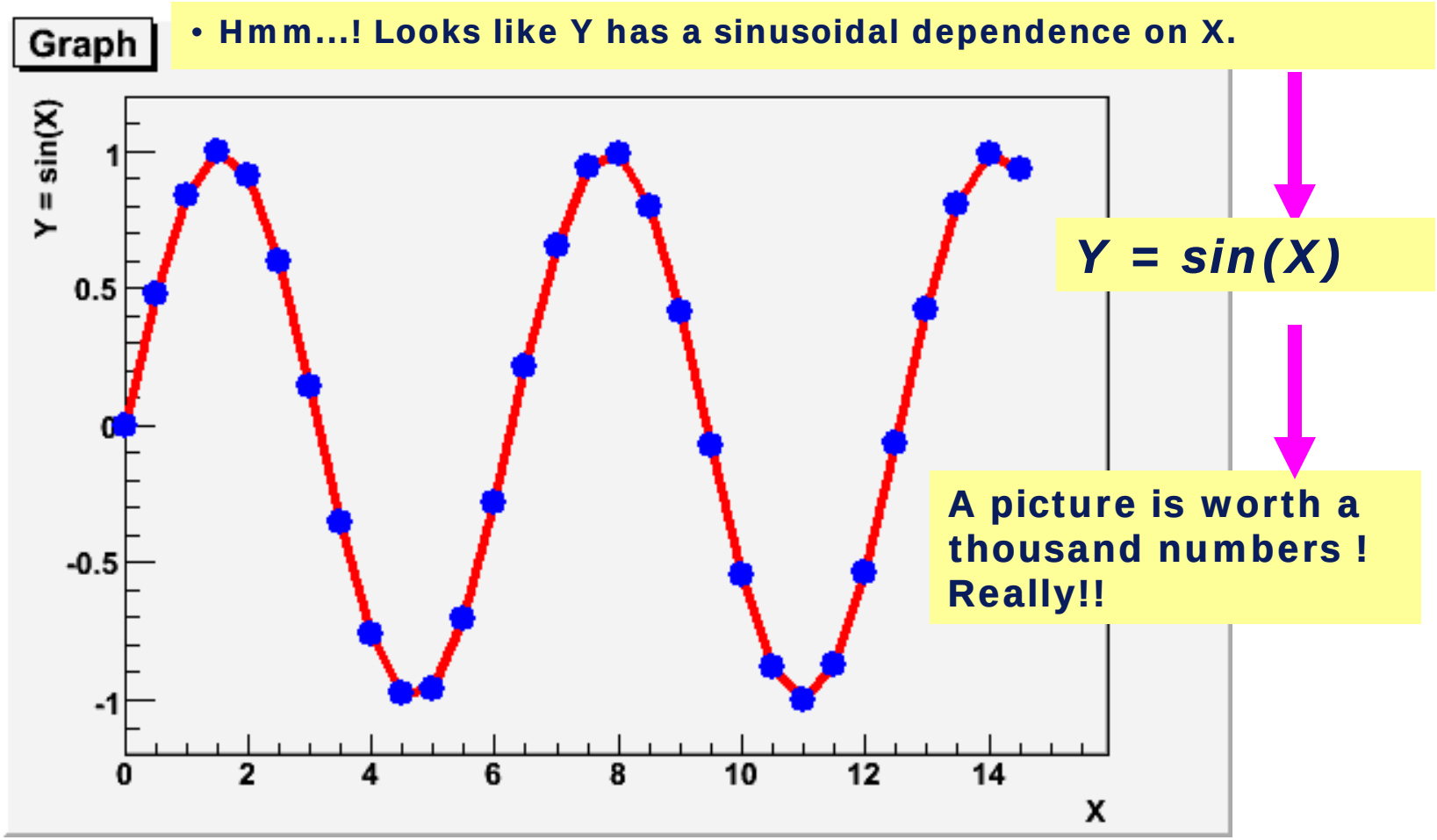
- **What do these numbers tell us about the data ?**
- **What is the relationship between Y and X ??**
- ***Maybe I should do a regression analysis ?***

Same Data in a “picture”



- **A linear regression of Y vs X would certainly give the wrong answer.**
- **Looking at this graphs gives us a good idea of the functional relationship between Y and X.**

Same Data in a “picture”

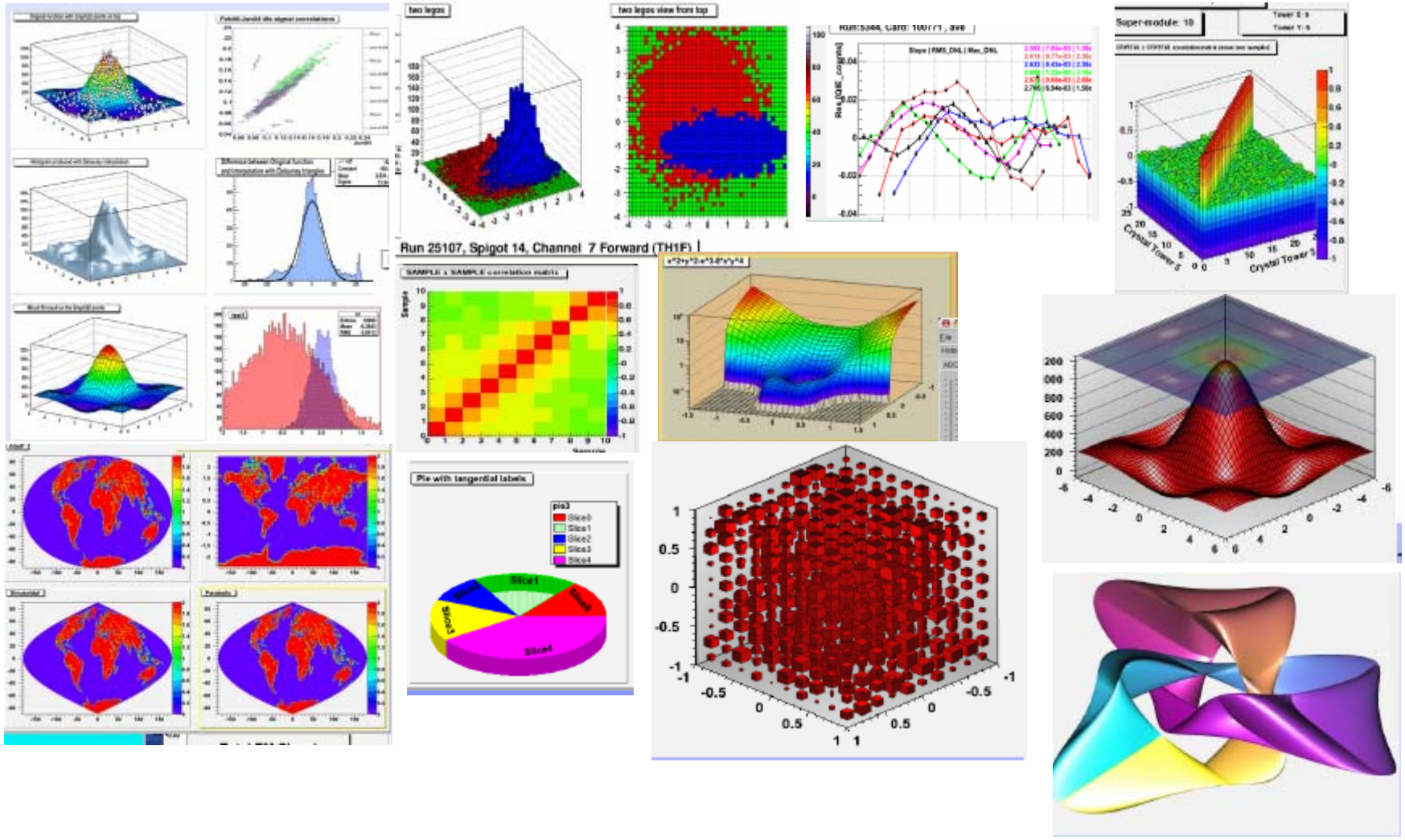


- **Data visualization is a *crucial* step in *optimal* modeling of data.**
- **ROOT provides *convenient* graphic tools to visualize and explore even large data sets.**

Some Visualization Tools in ROOT

- Histograms
  - 1, 2 and 3 dimensional histograms.
  - Profile histograms.
  - Graphic interface for fitting the distributions.
- Easy visualization of any arbitrary function in 1, 2 and 3 dimensions.
- Graphs
- Pie charts.
- etc....

Some Examples of Data Visualization in ROOT



From ROOT Workshop 2007

Some Statistical Tools in ROOT

- Multiple regression.
- Maximum likelihood fitting.
  - ROOT comes with general purpose function minimisers.
  - They can be used to fit **any user defined function** to the data.
  - One can use either least-squares, or maximum likelihood method.
- Neural Networks.
- Function approximation using basis functions.
- Linear Algebra
- Principal components analysis.
- Numerous mathematical functions.
- Fourier transforms.
- Random number generators and simulation tools.
- RooFit: A toolkit for Data Modeling
- And much more.....

More Machine Learning: TMVA

- TMVA: Toolkit for Multivariate Analysis
  - TMVA comes as a part of the standard ROOT installation.
- TMVA can be used to identify the predictive power of variables.
- Several classification algorithms available in TMVA, including:
  - Multidimensional k-NN classifier.
  - Linear discriminant analysis.
  - Function discriminant analysis.
  - Neural Networks.
  - Boosted/Bagged decision trees.
  - Rules based predictive learning (RuleFit)
  - Support Vector Machine.

Some other relevant features

- Connectivity to external databases via ODBC.
- CINT – the C interpreter
  - CINT interprets C/C++ code.
  - Allows fast prototyping of the analysis.
- High quality graphics, including:
  - Animation
  - Image-processing.
  - Three dimensional graphics.
- Extensive self-contained GUI toolkit
  - Can create Customized user interfaces.
- Networking and parallel processing.

Summary

- ROOT is a free, open-source data analysis and data mining tool from CERN (where the Web was invented).
- Specifically designed to handle, store and analyze large amounts of data ***very efficiently***.
  - Allows for serious data analysis even on a laptop PC.
- Comes with:
  - Numerous statistical, mathematical and data mining tools.
  - Excellent data exploration/visualization tools.
  - Interfaces for convenient interaction with data.
  - Complete GUI development toolkit.
  - CINT – the C-interpreter.
  - Connectivity with external databases.
- Suitable for any environment requiring serious and **Scalable** data analysis, modeling and reporting.

Useful Links

- ROOT download, documentation, tutorials etc.:

  http://root.cern.ch/

- TMVA Home Page:

  http://tmva.sourceforge.net/

- RooFit Home Page:

  http://roofit.sourceforge.net/