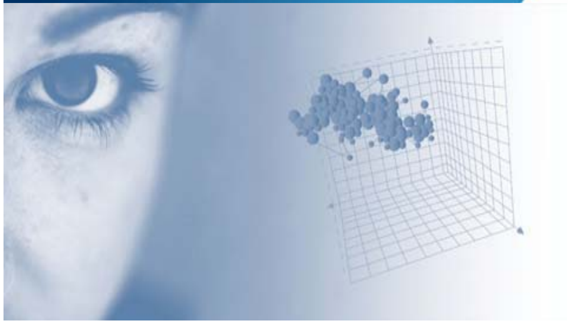
Spotfire®



Configurable & Integrated Life Science Informatics: Putting Data to Work

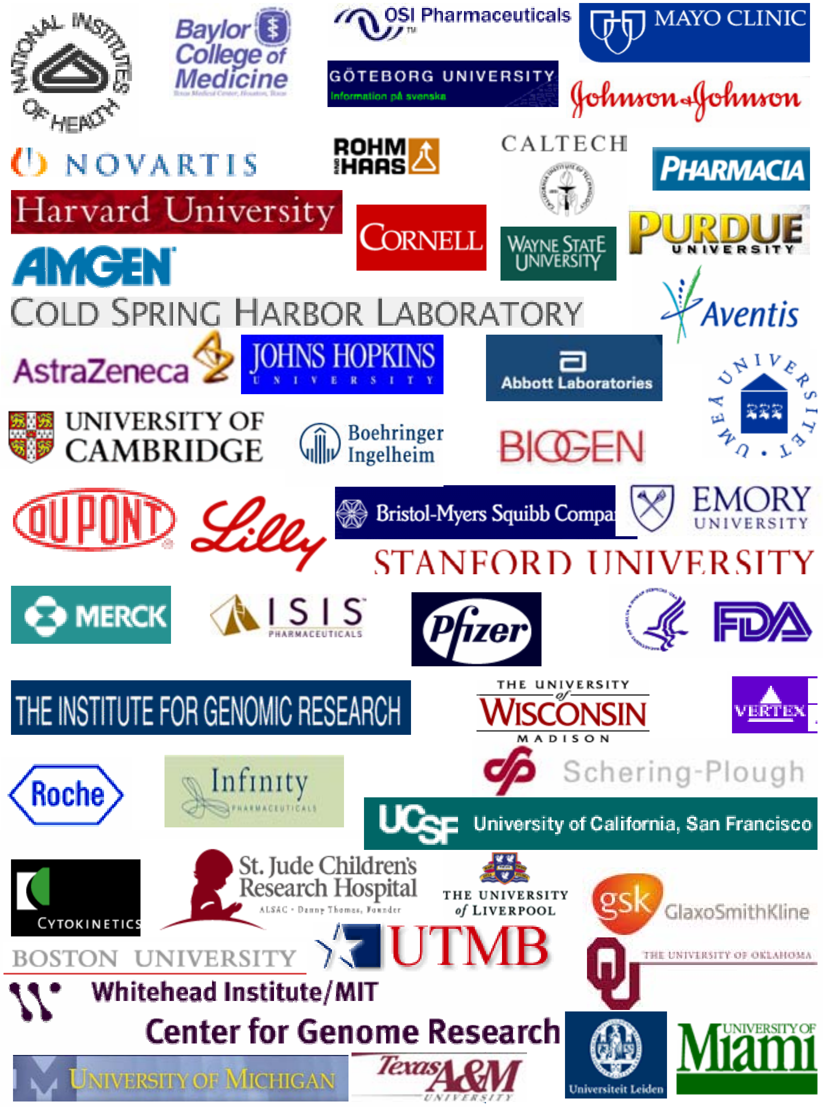
Mark Demesmaeker, PhD

mark@spotfire.com

[BC]² Basel Computational Biology Conference

Spotfire Company Update

- 700+ customers
- 25,000+ users
- 150+ employees
- Headquartered in Somerville, MA & Göteborg, Sweden
- Dominant in life science
  - All major Pharmas & 250 Biotechs
  - Discovery
    - Proteomics
    - Genomics
    - Lead Discovery
    - HTS
    - ADME/Tox
  - Development
    - Clinicals, Pilot Manufacturing Supply-chain, Sales/Marketing



Analytics Proposition

- Drug discovery is also an information business
- Analytics is one of the few remaining areas of competitive differentiation
- Dozens of critical R&D problems where data analysis is key factor for success
- Knowledge management systems should be architected around the people not the data
- Analytical applications require a platform and strategy to go beyond individual productivity gains

Creating Competitive Advantage in R&D

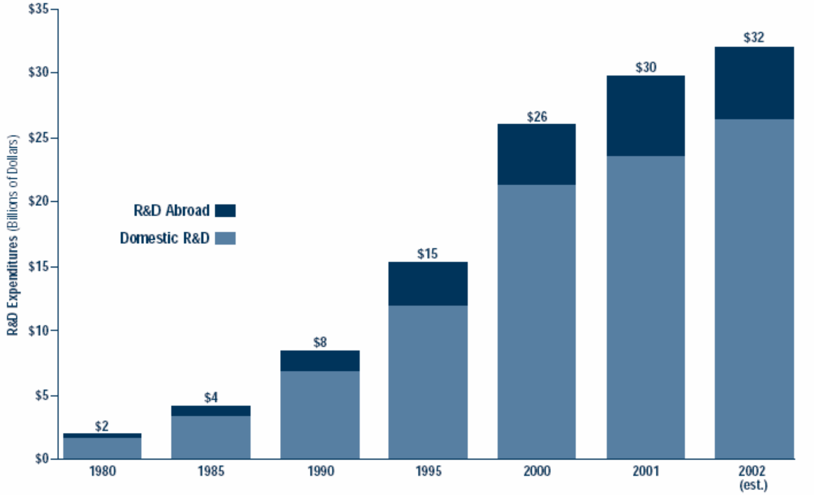
"First Generation Approach"

- High throughput research process investments
  - High throughput screening technology
  - Gene expression technology
  - Combichem technology
  - Genomics databases
- All companies invested, productivity improvement realized?

"Second Generation Approach"

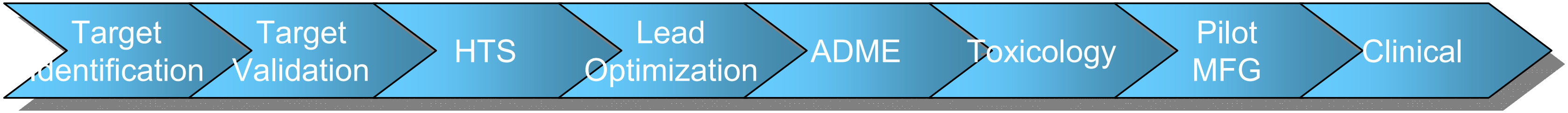
- High quality decision-making investments
  - Focus on taking advantage of data
  - Spread best-practices for choosing compounds and targets
  - Create proprietary decision making capabilities
- Towards effective decision making

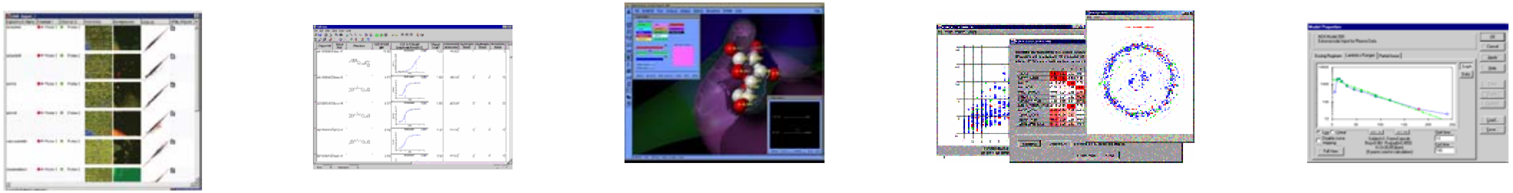
Figure 2.1 Research and Development Continues to Grow



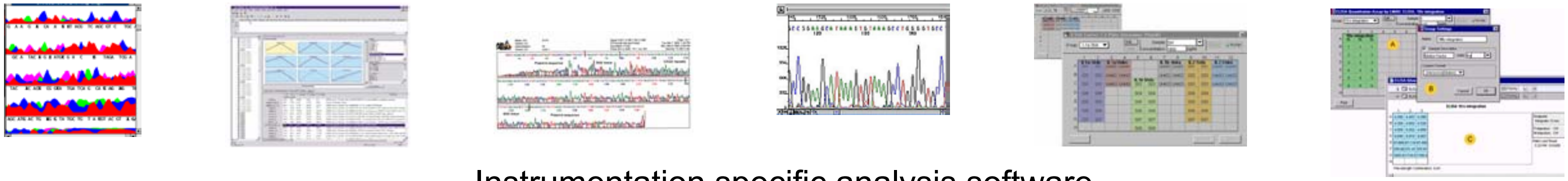
Source: Pharmaceutical Research and Manufacturers of America, PhRMA Annual Membership Survey, 2003.

First Steps of Data Analysis

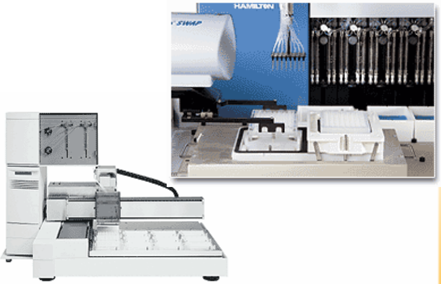




Domain specific analysis software

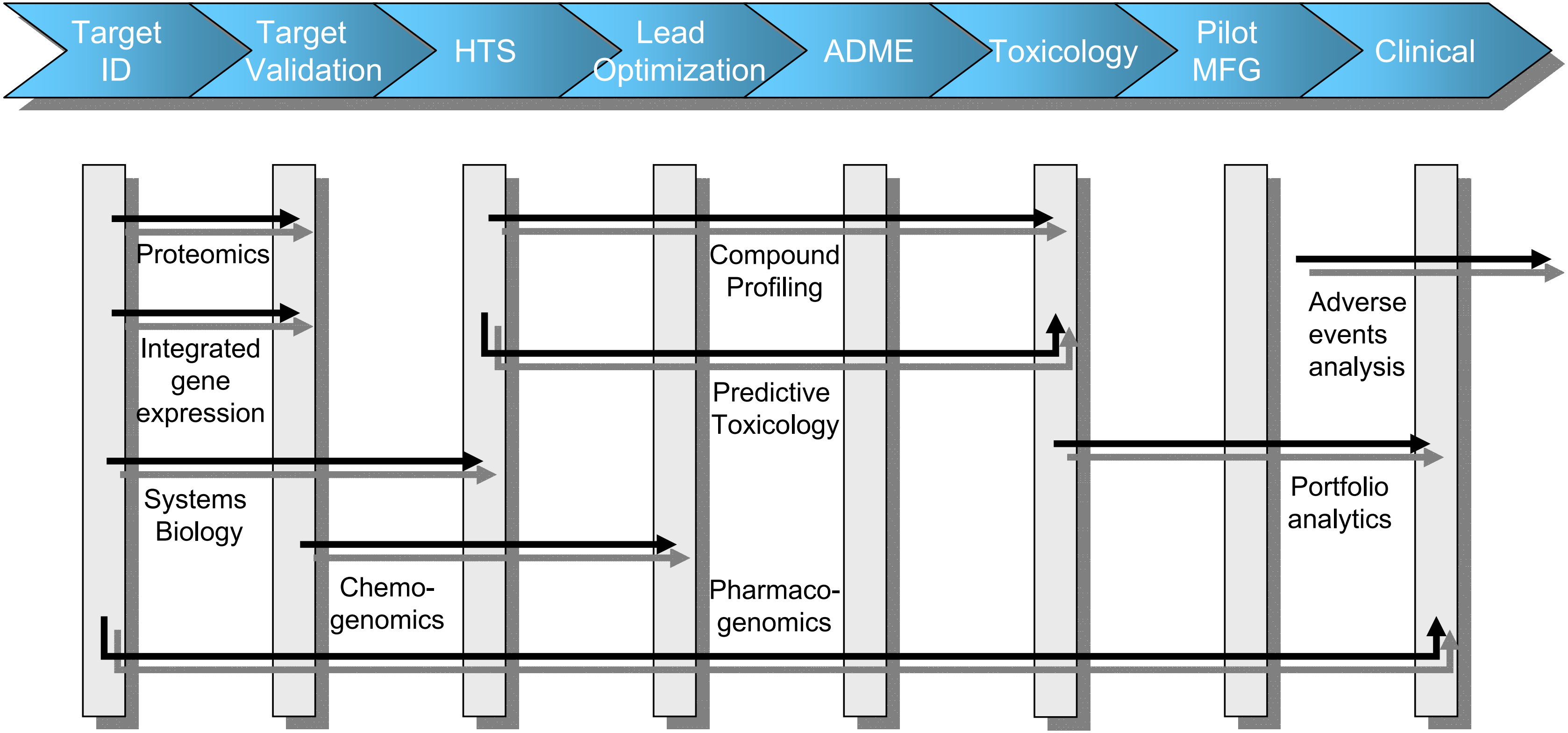


Instrumentation specific analysis software



Spotfire®

R&D Wide Analytics



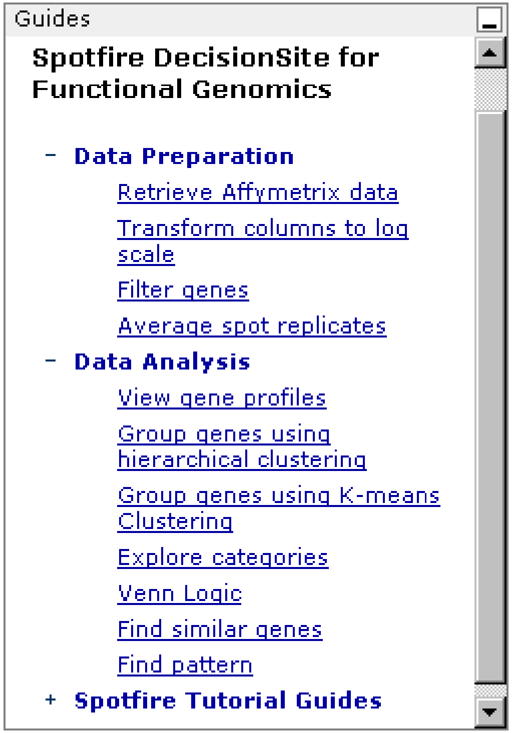
- Faster, high quality R&D requires integrated drug discovery approaches

- Integrated project teams need
  - Integrated data access
  - Integrated data analysis

Guided Analytics

Guided Analytics supports processes, analysis workflows, and common end-user tasks required to analyze data from multiple sources

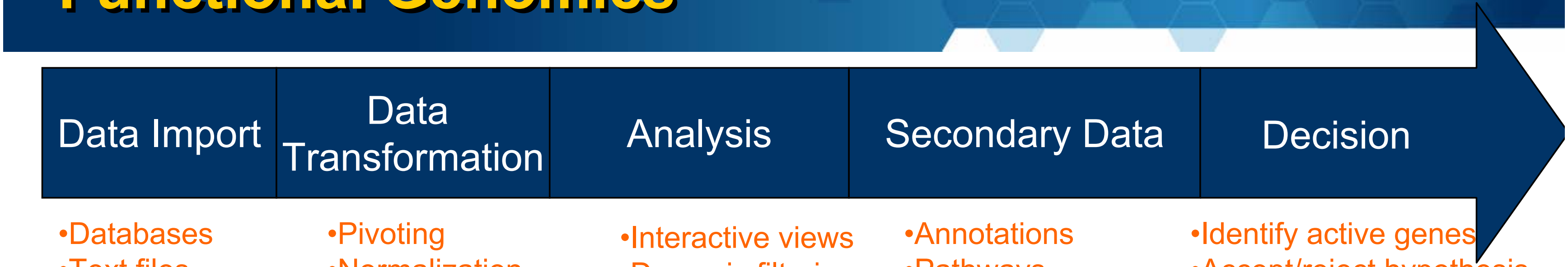
- Capture decision-making processes
  - Expert knowledge capture – spread the knowledge
  - Best practices – higher accuracy and efficiency
- Tie together data access, analysis, visualization and reporting tools suitable for a particular task
  - Get the right information
  - Apply the right calculations
  - Create the right visualizations
  - Create the appropriate reports



Spotfire DecisionSite for Functional Genomics

- **Data Preparation**
  - Retrieve Affymetrix data
  - Transform columns to log scale
  - Filter genes
  - Average spot replicates
- **Data Analysis**
  - View gene profiles
  - Group genes using hierarchical clustering
  - Group genes using K-means Clustering
  - Explore categories
  - Venn Logic
  - Find similar genes
  - Find pattern
- **Spotfire Tutorial Guides**

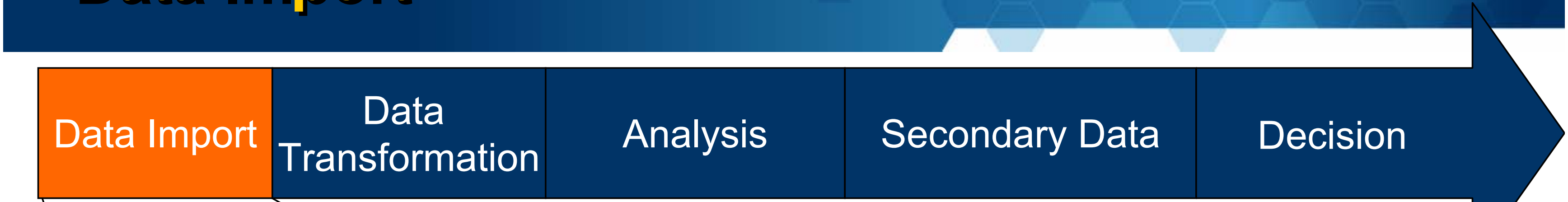
Functional Genomics

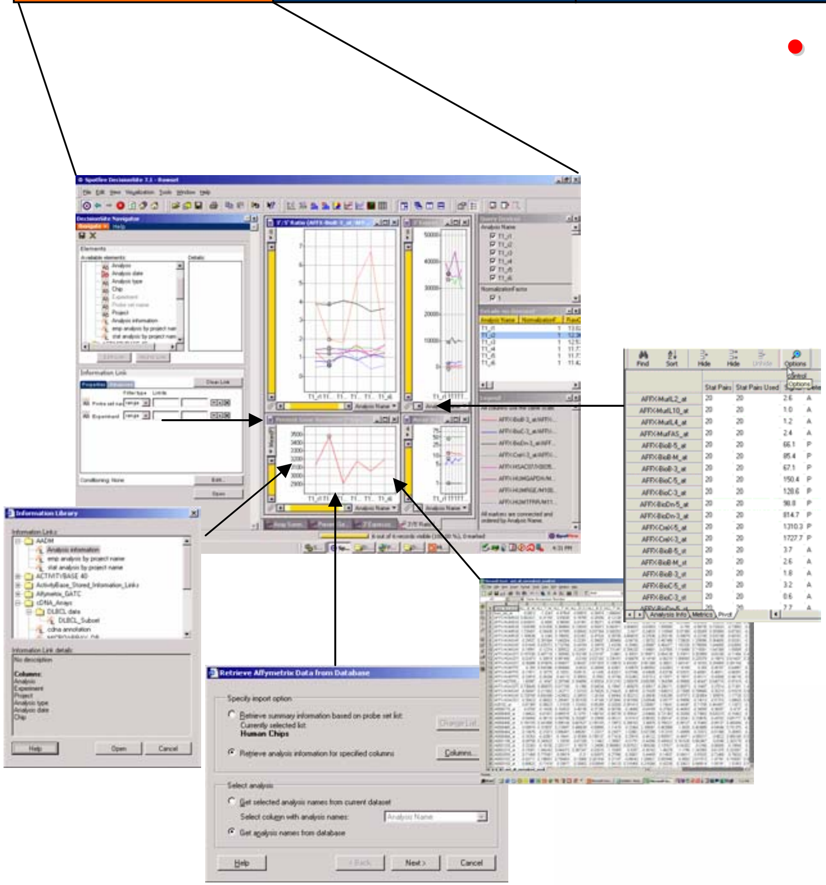


Data Import	Data Transformation	Analysis	Secondary Data	Decision
•Databases	•Pivoting	•Interactive views	•Annotations	•Identify active genes
•Text files	•Normalization	•Dynamic filtering	•Pathways	•Accept/reject hypothesis
•Clipboard	•Scaling	•Clustering	•Public websites	•Share thought process
•Affymetrix AADM	•Combine replicates	•PCA	•Internal websites	•Create a report
•Affymetrix MAS		•Profile searching		
•GenePix files		•ANOVA/t-test		
•Rosetta Resolver		•List comparison		

- Steps in analyzing gene expression data
  - Identify the data sources
  - Select methods of data transformation
  - Incorporate appropriate analysis tools to test hypothesis
  - Add known information to extend the analysis
  - Identify subsets of genes for further study
  - Accept or reject hypothesis and share conclusions

Data Import




Spotfire DecisionSite 7.1 - Rosset
DecisionSite Navigator
Information Library
Retrieve Affymetrix Data from Database

Retrieve Affymetrix Data from Database

Specify import option

○ Retrieve summary information based on probe set list
Currently selected list:
Human Chips

● Retrieve analysis information for specified columns

Select analysis

○ Get selected analysis names from current dataset
Select column with analysis names: Analysis Name

● Get analysis names from database

Help | < Back | Next > | Cancel

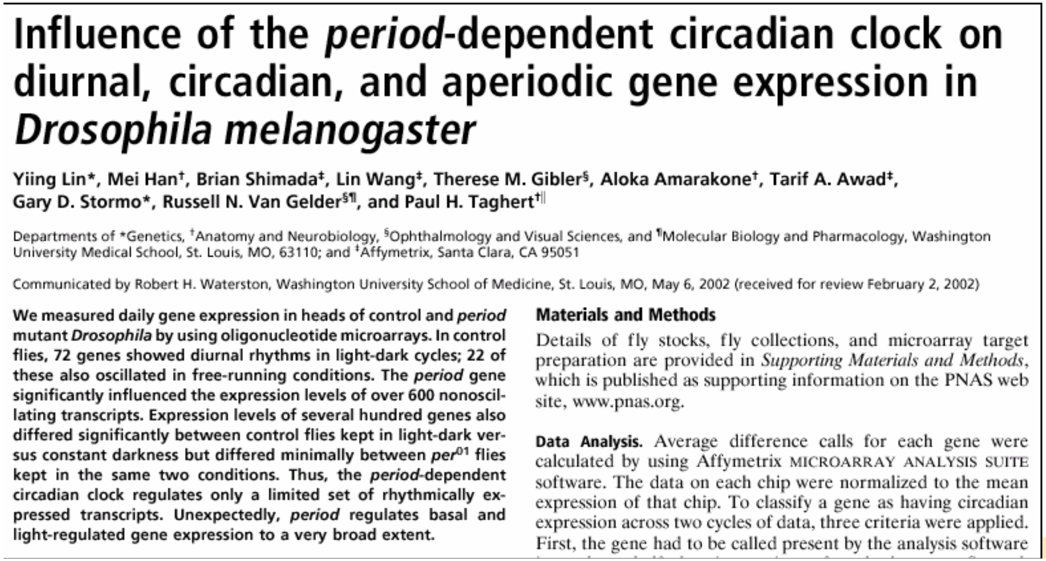
- Flexible, easy data import
  - Spreadsheet, text files, clipboard
  - Relational databases
  - Define columns on the fly
  - Merge additional columns of data
  - Support multiple types of gene expression data: cDNA arrays, GeneChip, A/P, experimental parameters, average and raw data together, p-values
  - data access from standard tools
    - Affymetrix MAS5, databases, QA/QC
    - Rosetta Resolver

Spotfire®

Data analysis demonstration

- Dataset
  - **Publication:** "Influence of the period-dependant circadian clock on diurnal, circadian, and aperiodic gene expression in *Drosophila melanogaster*" PNAS, July 9, 2002.
  - **Data available at:**

    http://circadian.wustl.edu/circadian.html



Influence of the *period*-dependent circadian clock on diurnal, circadian, and aperiodic gene expression in *Drosophila melanogaster*

Yiing Lin[*], Mei Han[†], Brian Shimada[‡], Lin Wang[‡], Therese M. Gibler[§], Aloka Amarakone[†], Tarif A. Awad[‡], Gary D. Stormo[*], Russell N. Van Gelder[§¶], and Paul H. Taghert[†||]

Departments of [*]Genetics, [†]Anatomy and Neurobiology, [§]Ophthalmology and Visual Sciences, and [¶]Molecular Biology and Pharmacology, Washington University Medical School, St. Louis, MO, 63110; and [‡]Affymetrix, Santa Clara, CA 95051

Communicated by Robert H. Waterston, Washington University School of Medicine, St. Louis, MO, May 6, 2002 (received for review February 2, 2002)

We measured daily gene expression in heads of control and *period* mutant *Drosophila* by using oligonucleotide microarrays. In control flies, 72 genes showed diurnal rhythms in light-dark cycles; 22 of these also oscillated in free-running conditions. The *period* gene significantly influenced the expression levels of over 600 nonoscillating transcripts. Expression levels of several hundred genes also differed significantly between control flies kept in light-dark versus constant darkness but differed minimally between per^{01} flies kept in the same two conditions. Thus, the *period*-dependent circadian clock regulates only a limited set of rhythmically expressed transcripts. Unexpectedly, *period* regulates basal and light-regulated gene expression to a very broad extent.

Materials and Methods

Details of fly stocks, fly collections, and microarray target preparation are provided in *Supporting Materials and Methods*, which is published as supporting information on the PNAS web site, www.pnas.org.

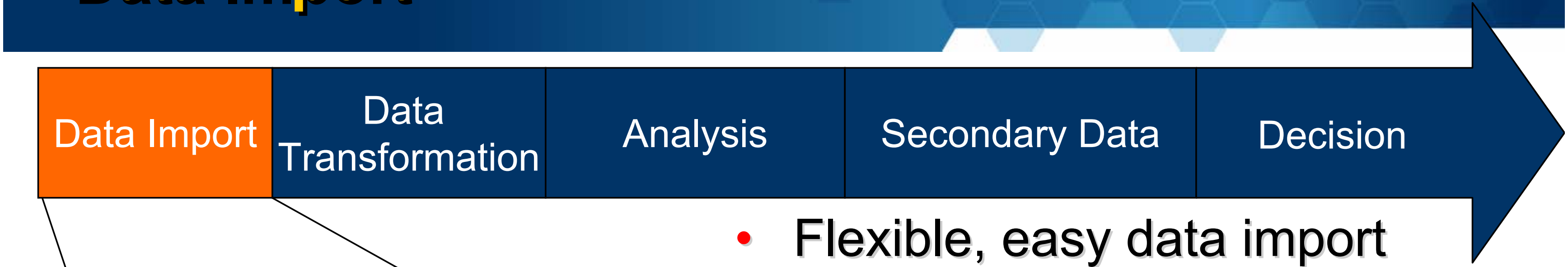
Data Analysis. Average difference calls for each gene were calculated by using Affymetrix MICROARRAY ANALYSIS SUITE software. The data on each chip were normalized to the mean expression of that chip. To classify a gene as having circadian expression across two cycles of data, three criteria were applied. First, the gene had to be called present by the analysis software

Data analysis demonstration

- Research performed at Washington University on *Drosophila melanogaster* (fruit fly)
- Gene expression data from Affymetrix AADM (Affymetrix Advanced Data Model)
- Experiments were run on 48 chips with some extra repeats and time points, most due to noise

• Samples:	per01 (period gene mutants)
	W33 (wild type)
• Condition:	LD= exposed to light and dark cycles
	DD= exposed to continuous dark cycles
• Time:	0,4,8,12,16,20 hour time points
• Replicates:	2 replicates for every experiment
• Repeated:	Chips having “#2” notation repeated due to pixel noise

Data Import



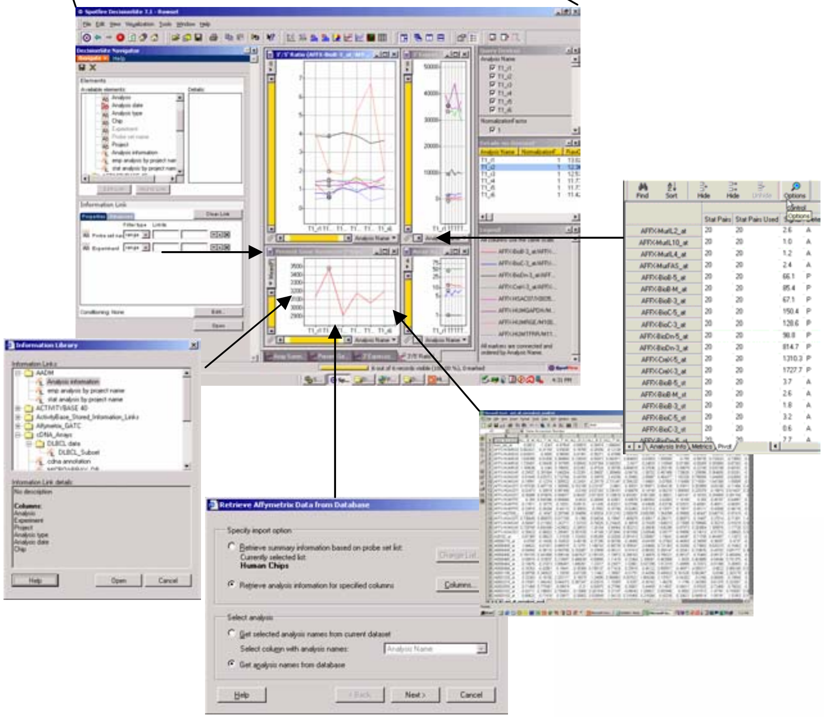
Data Import

Data Transformation

Analysis

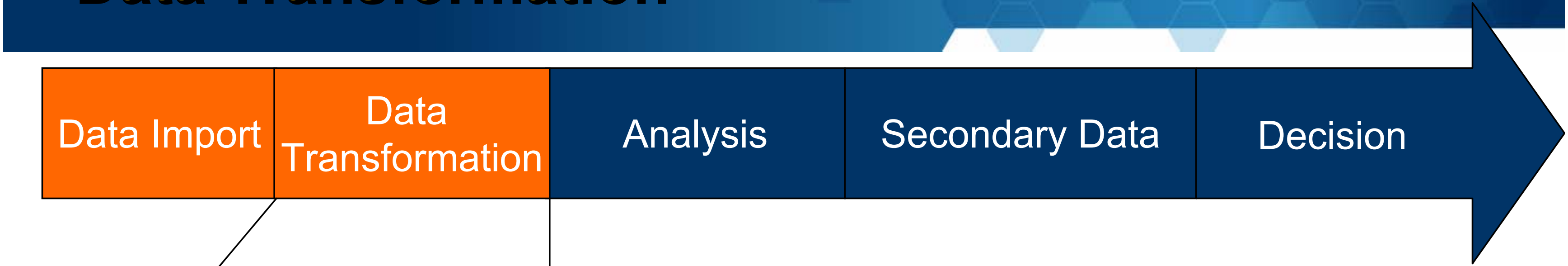
Secondary Data

Decision


Information Library
Retrieve Affymetrix Data from Database
Specify import option
Retrieve summary information based on probe set list
Retrieve analysis information for specified columns
Select analysis
Get selected analysis names from current dataset
Get analysis names from database
Help
Next >
Cancel

- Flexible, easy data import
  - Spreadsheet, text files, clipboard
  - Relational databases
  - Define columns on the fly
  - Merge additional columns of data
  - Support multiple types of gene expression data: cDNA arrays, GeneChip, A/P, experimental parameters, average and raw data together, p-values
  - data access from standard tools
    - Affymetrix MAS5, databases, QA/QC
    - Rosetta Resolver

Data Transformation



Data Import

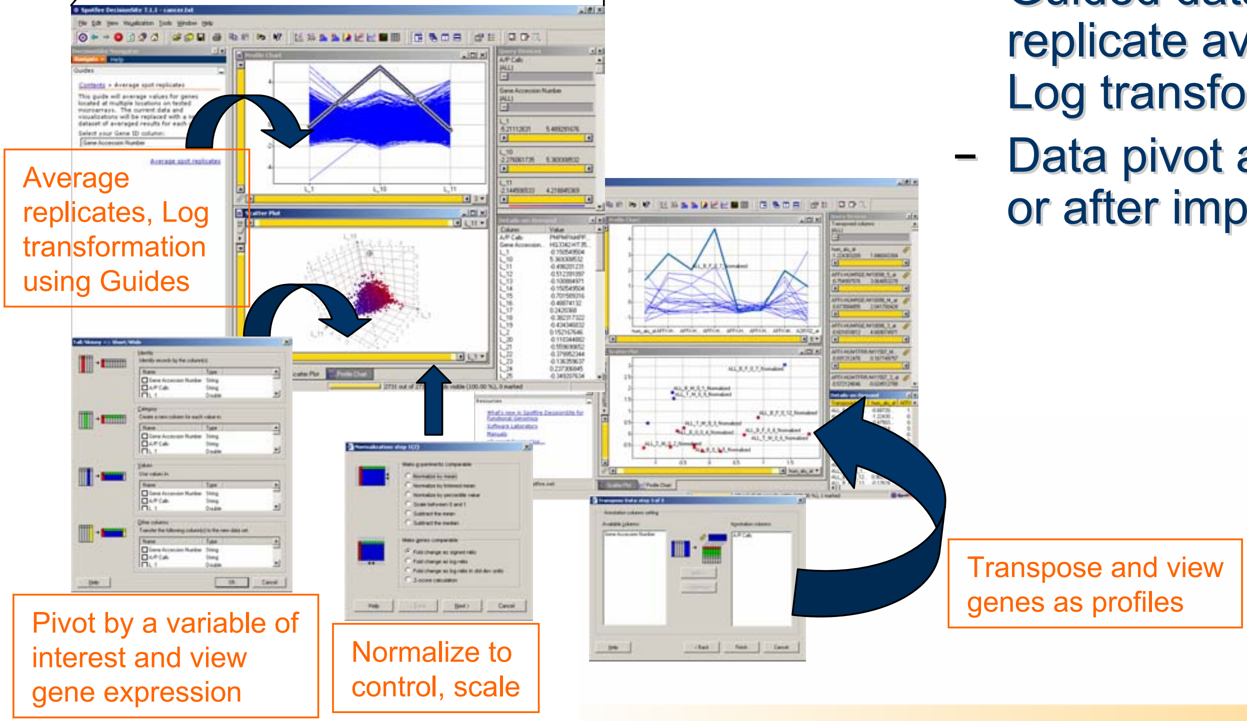
Data Transformation

Analysis

Secondary Data

Decision

- Flexible and easy data transformation
  - Guided data normalization, replicate average, Log transformations
  - Data pivot and transpose during or after import



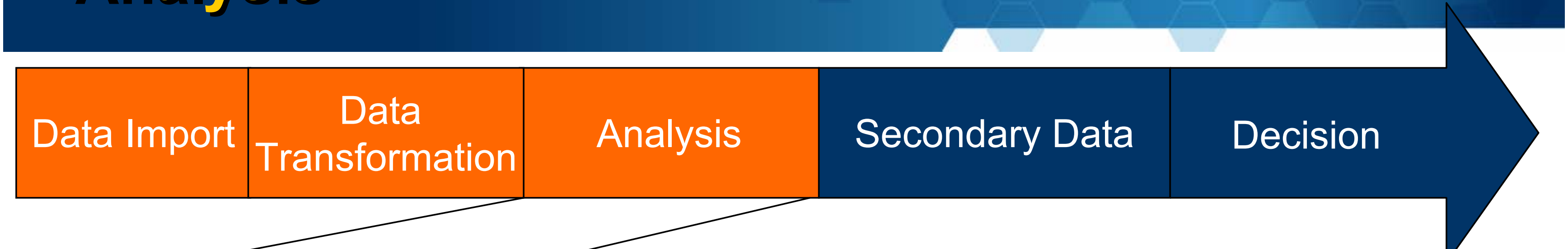
Average replicates, Log transformation using Guides

Pivot by a variable of interest and view gene expression

Normalize to control, scale

Transpose and view genes as profiles

Analysis



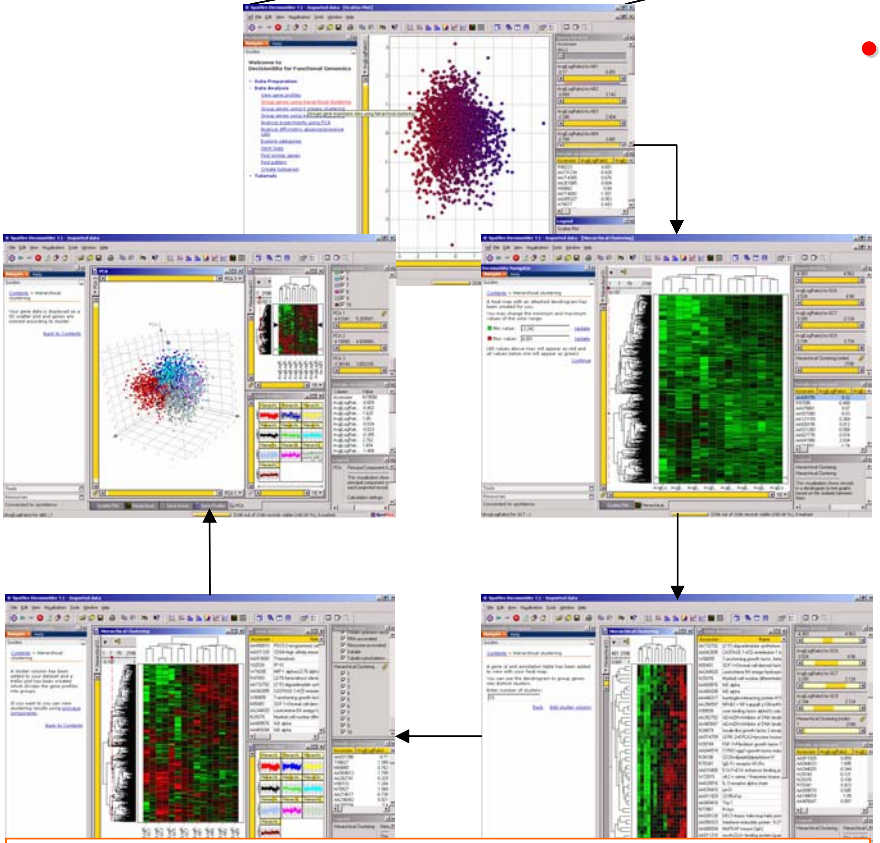
Data Import

Data Transformation

Analysis

Secondary Data

Decision

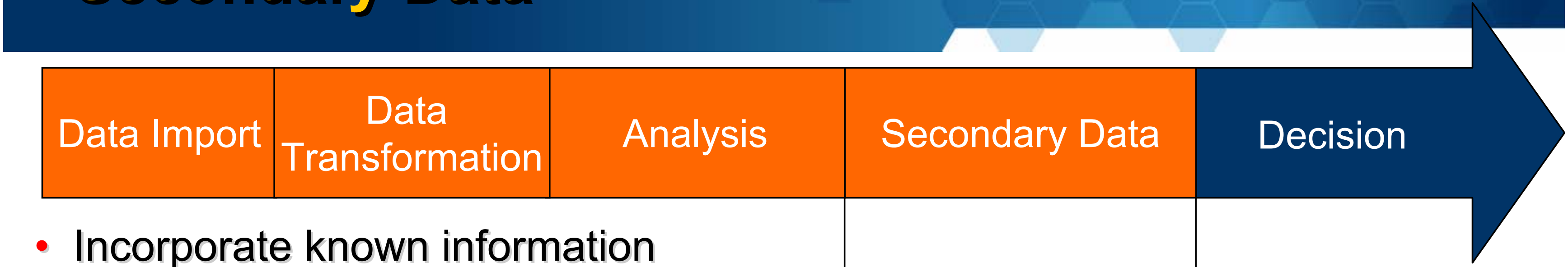


Guided hierarchical clustering prompts user for parameter input, creates appropriate views and extends the analysis through PCA

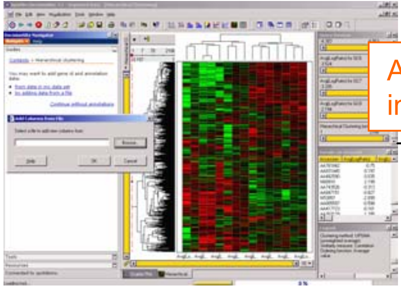
- Powerful data analysis
  - Guided workflow captures knowledge transfer and facilitates standardization
  - Analytical functions that everyone can use routinely and with ease
  - Informative views of the data generated at each step of the analysis
  - Linked interactive visualizations filtered simultaneously
  - Gene lists capture information on the most significant genes that can be flagged in future experiments/analysis

Spotfire®

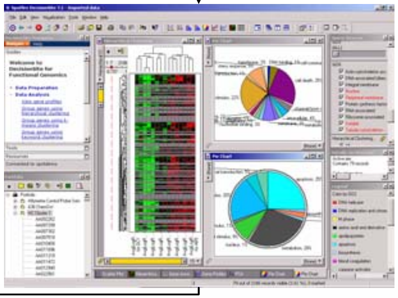
Secondary Data



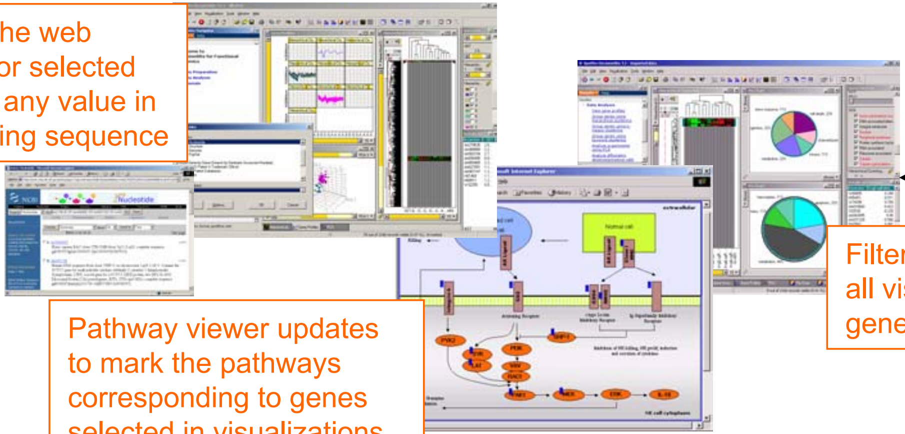
- Incorporate known information
  - Retrieve annotations from proprietary sources
  - Retrieve information from the web
  - Interact with pathway images to identify relationships
  - Use GO classifications to group genes
  - Dynamically interact with information



Add annotations from files or internal annotation source



Filtering to select a single cluster updates all visualizations and reveals majority of gene are involved in apoptosis



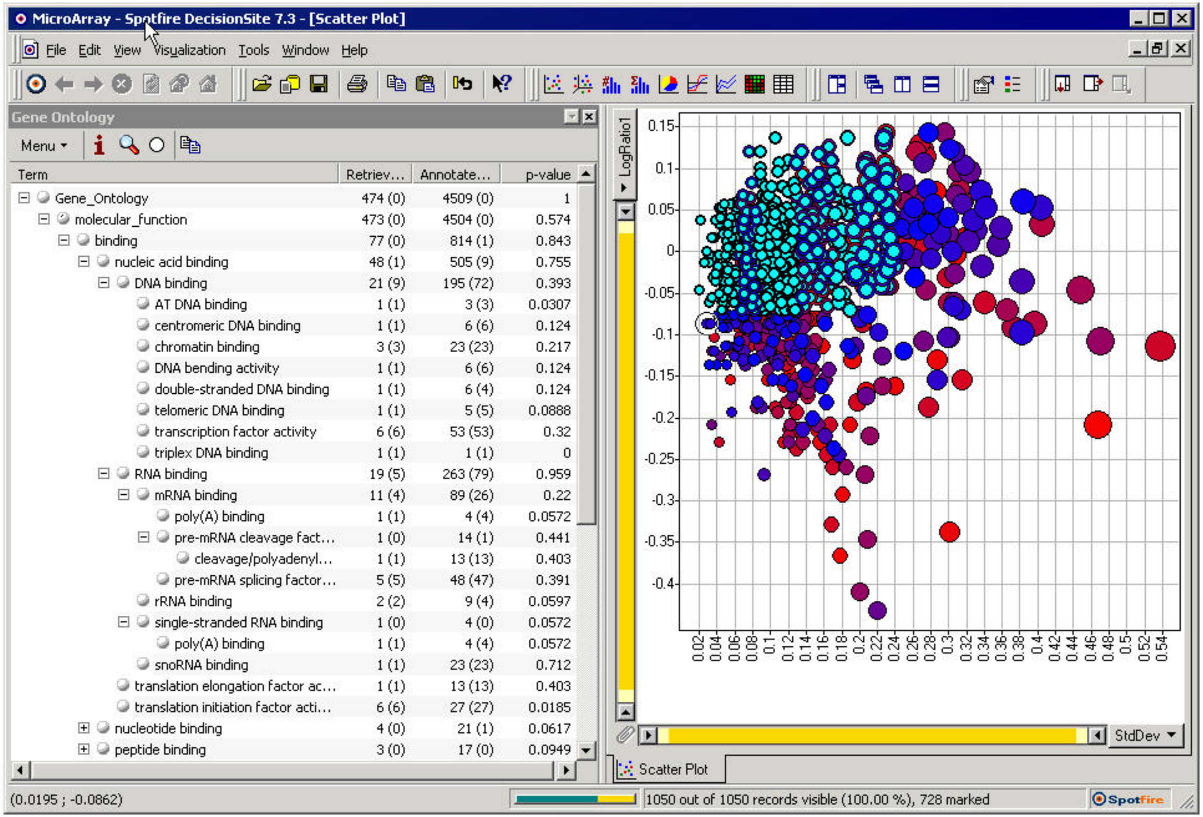
Information from the web can be retrieved for selected records based on any value in the dataset including sequence

Pathway viewer updates to mark the pathways corresponding to genes selected in visualizations

Spotfire®

Secondary Data – GO Browser

Locate records (Gene IDs) in the gene ontology hierarchy by function (molecular, biological, and cellular) and view associated annotations



Term	Retriev...	Annotate...	p-value
⊟ Gene_Ontology	474 (0)	4509 (0)	1
⊟ molecular_function	473 (0)	4504 (0)	0.574
⊟ binding	77 (0)	814 (1)	0.843
⊟ nucleic acid binding	48 (1)	505 (9)	0.755
⊟ DNA binding	21 (9)	195 (72)	0.393
AT DNA binding	1 (1)	3 (3)	0.0307
centromeric DNA binding	1 (1)	6 (6)	0.124
chromatin binding	3 (3)	23 (23)	0.217
DNA bending activity	1 (1)	6 (6)	0.124
double-stranded DNA binding	1 (1)	6 (4)	0.124
telomeric DNA binding	1 (1)	5 (5)	0.0888
transcription factor activity	6 (6)	53 (53)	0.32
triplex DNA binding	1 (1)	1 (1)	0
⊟ RNA binding	19 (5)	263 (79)	0.959
⊟ mRNA binding	11 (4)	89 (26)	0.22
poly(A) binding	1 (1)	4 (4)	0.0572
⊟ pre-mRNA cleavage fact...	1 (0)	14 (1)	0.441
cleavage/polyadenyl...	1 (1)	13 (13)	0.403
pre-mRNA splicing factor...	5 (5)	48 (47)	0.391
rRNA binding	2 (2)	9 (4)	0.0597
⊟ single-stranded RNA binding	1 (0)	4 (0)	0.0572
poly(A) binding	1 (1)	4 (4)	0.0572
snoRNA binding	1 (1)	23 (23)	0.712
translation elongation factor ac...	1 (1)	13 (13)	0.403
translation initiation factor acti...	6 (6)	27 (27)	0.0185
⊞ nucleotide binding	4 (0)	21 (1)	0.0617
⊞ peptide binding	3 (0)	17 (0)	0.0949

(0.0195 ; -0.0862)

1050 out of 1050 records visible (100.00 %), 728 marked

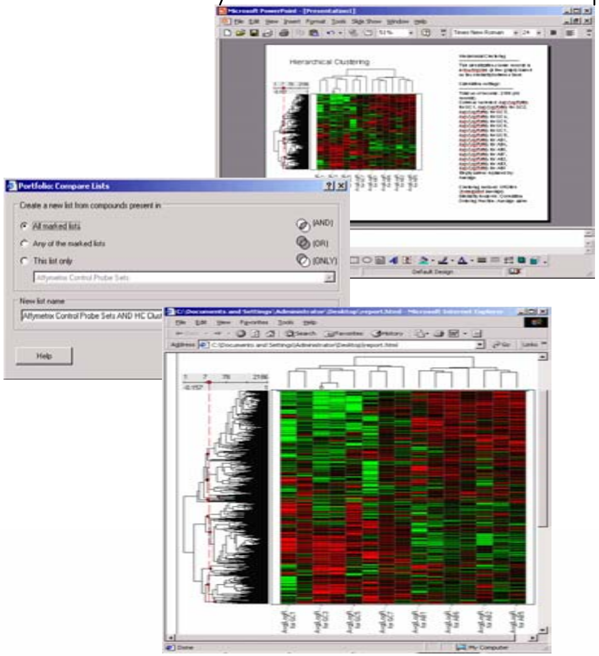
- Node information including a customizable live web links to additional sources of information about the GOID, term or a gene
- Search capabilities allowing to search on substring, ID or exact match
- Search for p-values below a certain threshold

Spotfire®

Decision Making

Data Import | Data Transformation | Analysis | Secondary Data | Decision

- Report results, save summary information and share thought processes
  - Use visualization and dynamic filtering in project team meetings
  - Identify gene clusters, compare lists and confirm expression patterns in new experiments
  - Save lists of significant, co-regulated genes
  - Share and archive the decision making process


Microsoft PowerPoint - [Presentation1]
Hierarchical Clustering
Portfolio: Compare Lists
Help
C:\Documents and Settings\Administrator\Desktop\report.html - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Search
Favorites
History
Address
Done
My Computer

Create a new list from compounds present in

● All marked lists (AND)
○ Any of the marked lists (OR)
○ This list only (ONLY)

Affymetrix Control Probe Sets

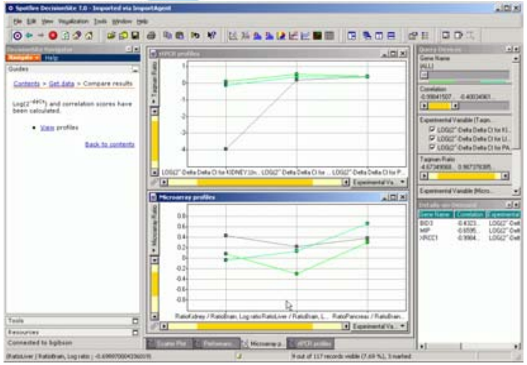
New list name

Affymetrix Control Probe Sets AND HC Clus

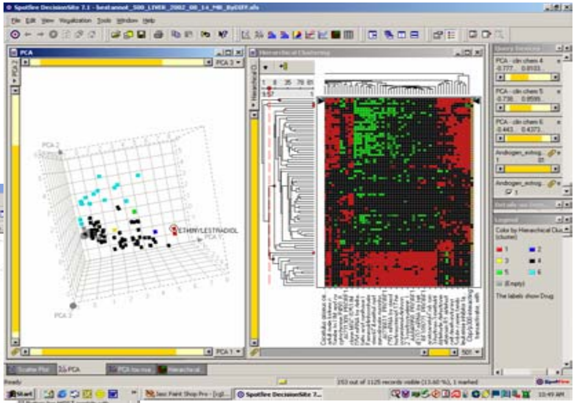
Spotfire®

Extending the analysis

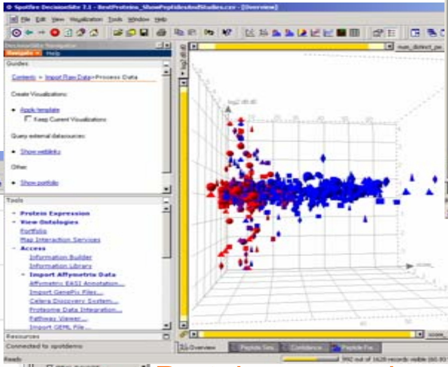
- What other information is available?
  - Validate microarray results with rtPCR data
  - Toxicogenomics: integrate Toxicology data
  - Compare gene expression with protein expression
  - Chemogenomics: integrate chemistry support



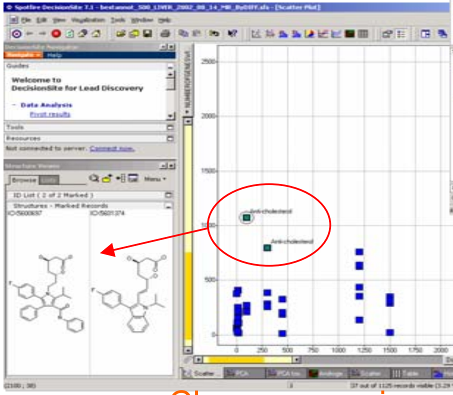
Compare microarray and rtPCR data



Toxicogenomics



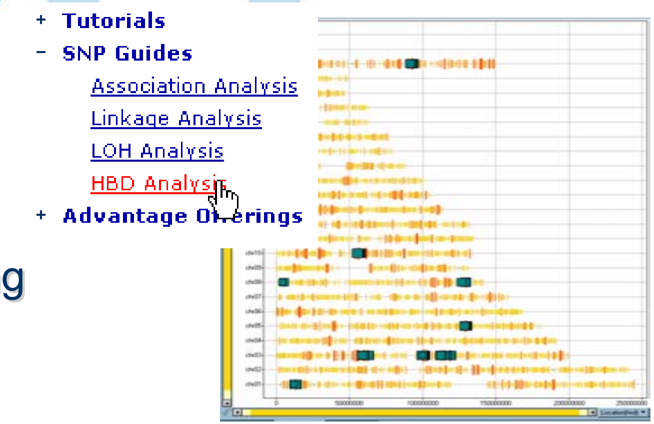
Protein expression


Spotfire DecisionSite 7.1 - [Scatter Plot]
Welcome to DecisionSite for Lead Discovery
Anti-cholesterol

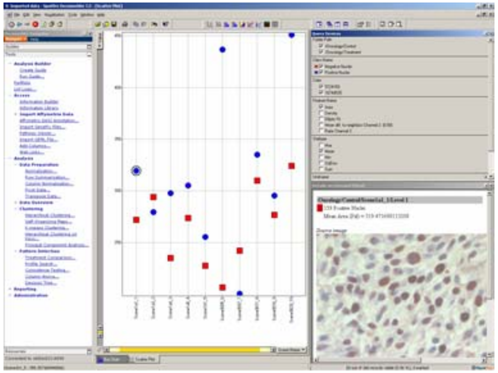
Chemogenomics

Extending the analysis

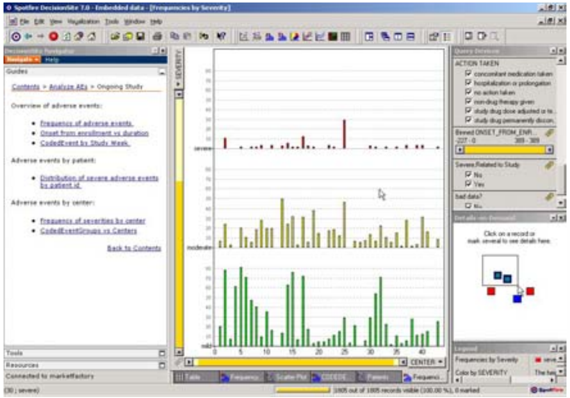
- What other information is available?
  - Investigate statistically significant regions of homozygosity
  - Validate data extracted from histological imaging
  - Adverse event analysis



SNP Analysis



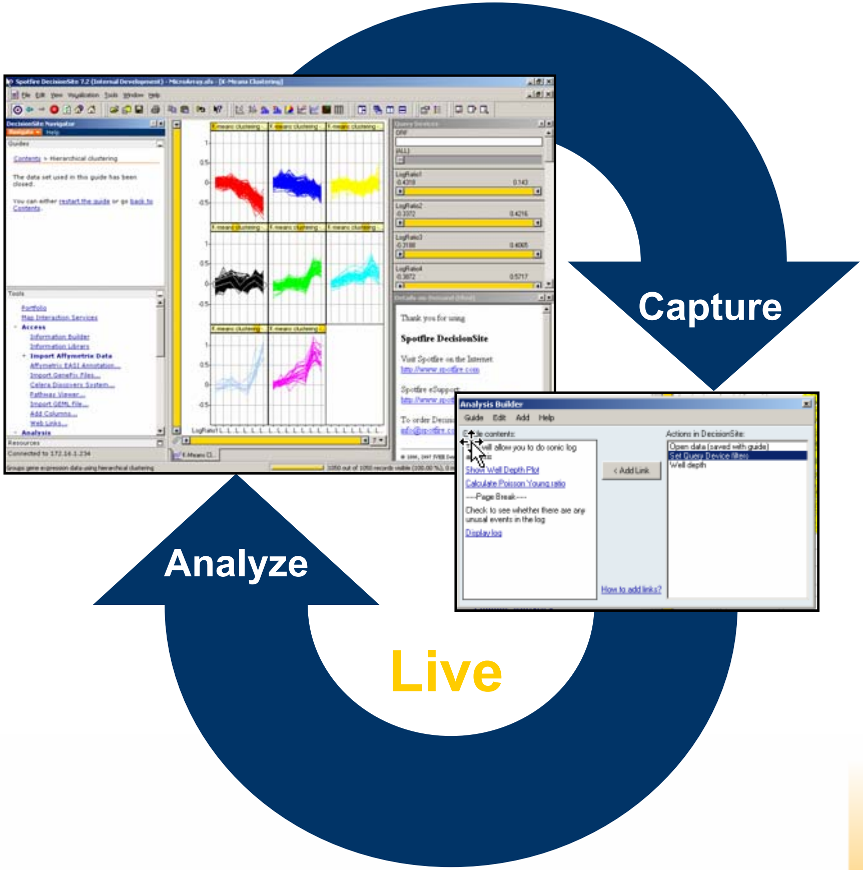
Histological Imaging


Spotfire DecisionSite 7.0 - Embedded data - [Frequencies by Severity]
Contents > Analyze AEs > Ongoing Study
Overview of adverse events:
Frequency of adverse events
Onset from enrollment vs duration
CodedEvent by Study Week
Adverse events by patient:
Distribution of severe adverse events by patient id
Adverse events by center:
Frequency of severities by center
CodedEventGroups vs Centers
Back to Contents
ACTION TAKEN
severe
moderate

Drug safety/Pharmacovigilance

Creating Custom Guides

- Expert analysis processes captured live
- End-user generated guided analytic applications
  - Ready for deployment with no additional development



The data set used in this guide has been closed.

You can either restart the guide or go back to Contents.

Connected to 172.16.1.234

Groups gene expression data using hierarchical clustering

1050 out of 1050 records visible (100.00 %)

Thank you for using

Spotfire DecisionSite

Visit Spotfire on the Internet
http://www.spotfire.com

Spotfire eSupport
http://www.spot

To order Decisi
info@spotfire.co

This will allow you to do sonic log analysis

Show Well Depth Plot

Calculate Poisson Young ratio

---Page Break---

Check to see whether there are any unusual events in the log

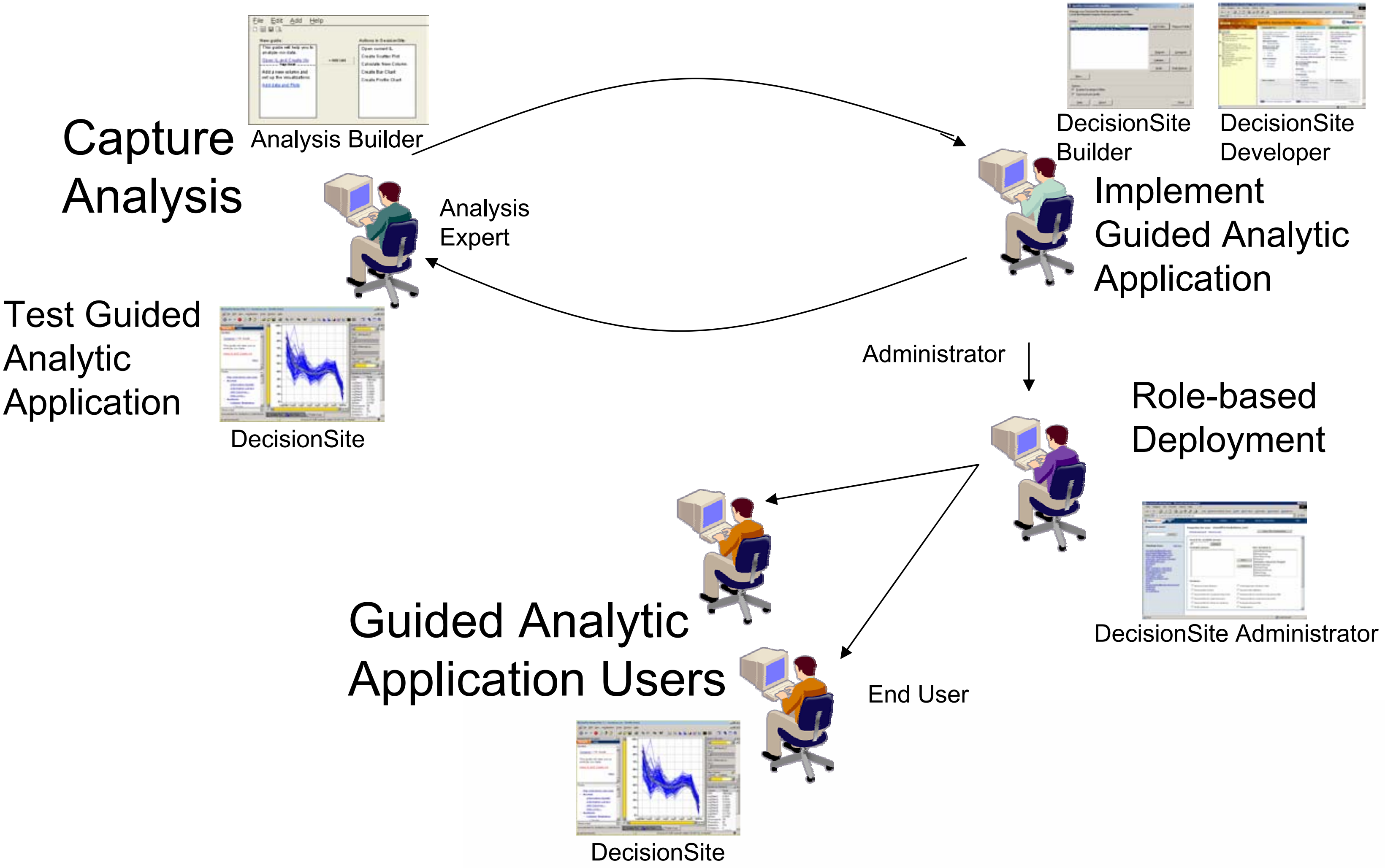
Display log

Open data (saved with guide)

Set Query Device filters

Well depth

Capturing and Reusing Analysis Processes



Capture Analysis

Analysis Builder

Analysis Expert

DecisionSite Builder

DecisionSite Developer

Implement Guided Analytic Application

Test Guided Analytic Application

DecisionSite

Administrator

Role-based Deployment

DecisionSite Administrator

Guided Analytic Application Users

End User

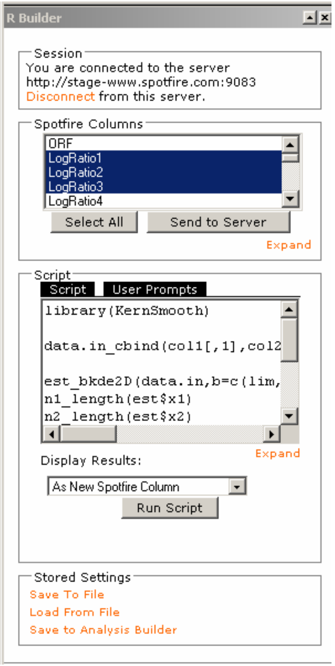
DecisionSite

Statistics in R

- Interactive multiplatform statistics environment
- Open-source implementation of the S language
- Easily extendable to include additional algorithms
- Active community developing algorithms and extensions
- For more info on R statistics and the Bioconductor Project see: **www.bioconductor.org**

R Integration

Develop R Code



Session
You are connected to the server http://stage-www.spotfire.com:9083
Disconnect from this server.

Spotfire Columns
ORF
LogRatio1
LogRatio2
LogRatio3
LogRatio4

Script

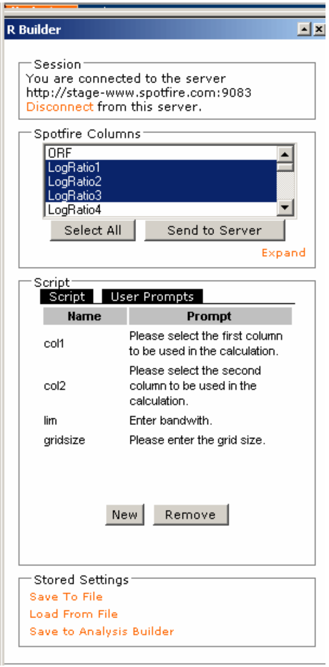
```
library(KernSmooth)

data.in_cbind(col1[,1],col2

est_bkde2D(data.in,b=c(lim,
n1_length(est$x1)
n2_length(est$x2)
```

Display Results: As New Spotfire Column

Stored Settings
Save To File
Load From File
Save to Analysis Builder



Session
You are connected to the server http://stage-www.spotfire.com:9083
Disconnect from this server.

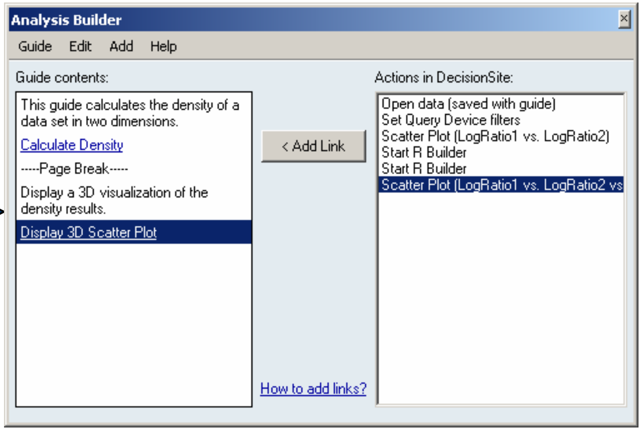
Spotfire Columns
ORF
LogRatio1
LogRatio2
LogRatio3
LogRatio4

Script

Name	Prompt
col1	Please select the first column to be used in the calculation.
col2	Please select the second column to be used in the calculation.
lim	Enter bandwith.
gridsize	Please enter the grid size.

Stored Settings
Save To File
Load From File
Save to Analysis Builder

Assemble DecisionSite Guide



Guide contents:

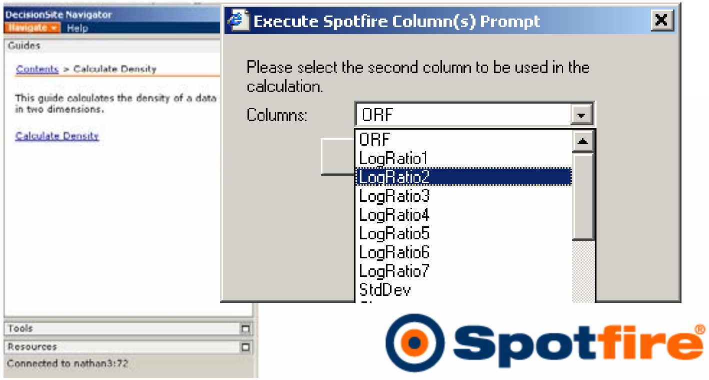
This guide calculates the density of a data set in two dimensions.
Calculate Density
-----Page Break-----
Display a 3D visualization of the density results.
Display 3D Scatter Plot

How to add links?

Actions in DecisionSite:

Open data (saved with guide)
Set Query Device filters
Scatter Plot (LogRatio1 vs. LogRatio2)
Start R Builder
Start R Builder
Scatter Plot (LogRatio1 vs. LogRatio2 vs

Run DecisionSite Application



Contents > Calculate Density

This guide calculates the density of a data in two dimensions.

Calculate Density

Connected to nathan3:72

Please select the second column to be used in the calculation.

Columns: ORF
ORF
LogRatio1
LogRatio2
LogRatio3
LogRatio4
LogRatio5
LogRatio6
LogRatio7
StdDev

Development Resources

SUCCESSFULLY DEPLOY NEW METHODS

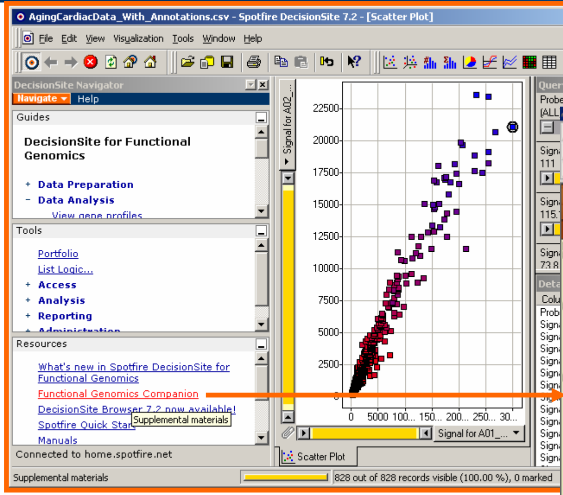
R/Bioconductor

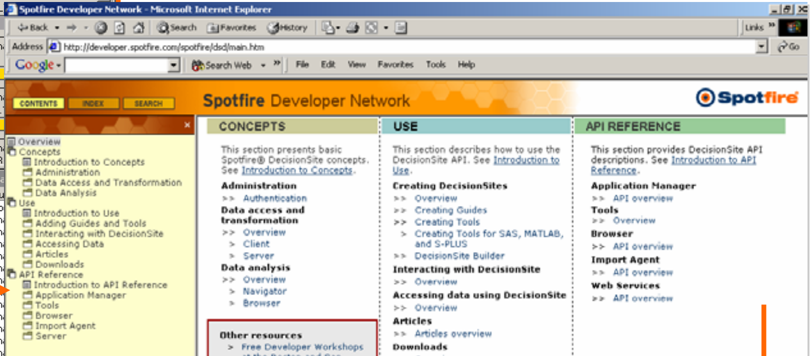
MATLAB

VisualBasic

C++

JavaScript





CONCEPTS

This section presents basic Spotfire® DecisionSite concepts. See Introduction to Concepts.

Administration
>> Authentication

Data access and transformation
>> Overview
> Client
> Server

Data analysis
>> Overview
> Navigator
> Browser

Other resources
> Free Developer Workshops at the Boston and San Fransisco User C... Space limited!
> Deploying R app... within Spotfire D... Take the R chall...

Premium Developer Su...

USE

This section describes how to use the DecisionSite API. See Introduction to Use.

Creating DecisionSites
>> Overview
>> Creating Guides
>> Creating Tools
> Creating Tools for SAS, MATLAB, and S-PLUS
>> DecisionSite Builder

Interacting with DecisionSite
>> Overview

Accessing data using DecisionSite
>> Overview

Articles
>> Articles overview

Downloads

API REFERENCE

This section provides DecisionSite API descriptions. See Introduction to API Reference.

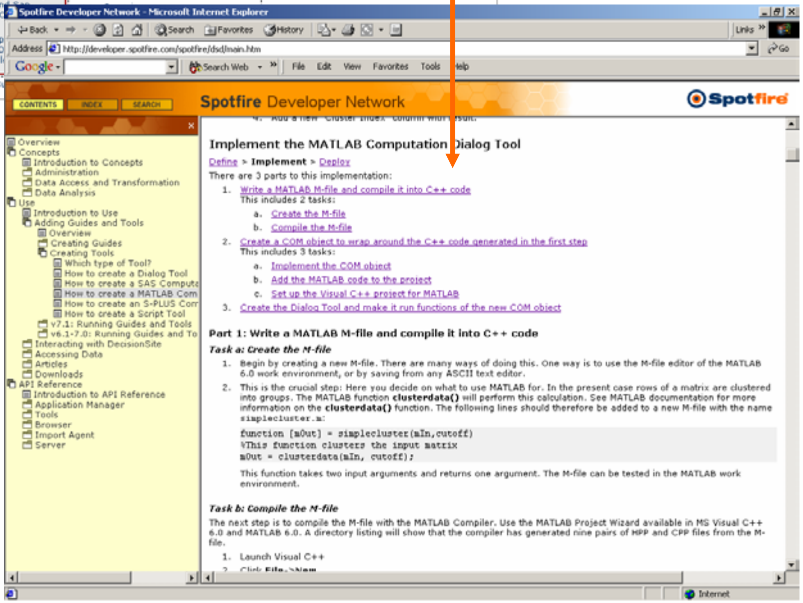
Application Manager
>> API overview

Tools
>> Overview

Browser
>> API overview

Import Agent
>> API overview

Web Services
>> API overview


Spotfire Developer Network - Microsoft Internet Explorer
Back
Search
Favorites
History
Address http://developer.spotfire.com/spotfire/dsd/main.htm
Go
Links
Google
Search Web
File Edit View Favorites Tools Help
CONTENTS
INDEX
SEARCH
Spotfire Developer Network
Spotfire
Overview
Concepts
Introduction to Concepts
Administration
Data Access and Transformation
Data Analysis
Use
Introduction to Use
Adding Guides and Tools
Overview
Creating Guides
Creating Tools
Which type of Tool?
How to create a Dialog Tool
How to create a SAS Computa
How to create a MATLAB Com
How to create an S-PLUS Com
How to create a Script Tool
v7.1: Running Guides and Tools
v6.1-7.0: Running Guides and To
Interacting with DecisionSite
Accessing Data
Articles
Downloads
API Reference
Introduction to API Reference
Application Manager
Tools
Browser
Import Agent
Server
Internet

4. Add a new "Cluster Index" column with result.

Implement the MATLAB Computation Dialog Tool

Define > **Implement** > Deploy

There are 3 parts to this implementation:

1. Write a MATLAB M-file and compile it into C++ code
   This includes 2 tasks:
   a. Create the M-file
   b. Compile the M-file
2. Create a COM object to wrap around the C++ code generated in the first step
   This includes 3 tasks:
   a. Implement the COM object
   b. Add the MATLAB code to the project
   c. Set up the Visual C++ project for MATLAB
3. Create the Dialog Tool and make it run functions of the new COM object

Part 1: Write a MATLAB M-file and compile it into C++ code

Task a: Create the M-file

1. Begin by creating a new M-file. There are many ways of doing this. One way is to use the M-file editor of the MATLAB 6.0 work environment, or by saving from any ASCII text editor.
2. This is the crucial step: Here you decide on what to use MATLAB for. In the present case rows of a matrix are clustered into groups. The MATLAB function **clusterdata()** will perform this calculation. See MATLAB documentation for more information on the **clusterdata()** function. The following lines should therefore be added to a new M-file with the name `simplecluster.m`:

```
function [mOut] = simplecluster(mIn,cutoff)
%This function clusters the input matrix
mOut = clusterdata(mIn, cutoff);
```

   This function takes two input arguments and returns one argument. The M-file can be tested in the MATLAB work environment.

Task b: Compile the M-file

The next step is to compile the M-file with the MATLAB Compiler. Use the MATLAB Project Wizard available in MS Visual C++ 6.0 and MATLAB 6.0. A directory listing will show that the compiler has generated nine pairs of HPP and CPP files from the M-file.

1. Launch Visual C++
2. Click **File->New**

- **Create your Analytic Application**:
  - Extensive API documentation that is searchable, current & online
  - Instruction on how to add Matlab, SAS, S-Plus computations
  - Instruction on how to add your algorithm or specialized tools

Thank you for your attention!

Spotfire®

/freedom to wonder_power to decide

For more information and archived webcast recordings:

www.spotfire.com

Contact: *mark@spotfire.com*

Spotfire®