Carolina Bioinformatics Research Center

Computational Infrastructure for Systems Toxicology

- Ivan Rusyn, M.D., Ph.D. (co-P.I.) – toxicology, genomics
- David Stotts, Ph.D. (co-P.I.) – computer science, software engineering
- Wei Wang, Ph.D. – computer science, data mining
- David Threadgill, Ph.D. – mammalian genetics, genomics
- Additional programmers and students

David Stotts
Department of Computer Science
University of North Carolina
Chapel Hill, NC 27599
919-962-1833
stotts@cs.unc.edu

Project Objectives

- **Develop and implement algorithms that streamline the analysis of multi-dimensional data streams in dose-response assessment and cross-species extrapolation.**

- **Facilitate the development of an industry-standard workflow for (i) analysis of the -omics data, (ii) linkages to classical indicators of adverse health effects, and (iii) integration with other types of biological information such as genome sequences and genetic differences between species.**

- **Build web-based, open-source and user-friendly graphical interfaces associated with interoperable computational tools for data analysis that facilitate incorporation of new data streams into basic research and decision-making pipelines (methods from Projects 1 and 2).**

- **Provide an interdisciplinary computer science resource to the environmental sciences and toxicology community**

- **Longer-term objectives include new software engineering methods for better execution and maintenance of above, and sharing and disseminating results**

Biostatistics Issues

- Data analysis procedures in concert with Project 1, including principal component analyses, distance-weighted discrimination, SAFE, etc.
- Specific data mining approaches also proposed, such as subspace clustering (SNPs vs. phenotypes, gene expression), that fall outside of typical statistical framework

Computational and Bioinformatics Issues

- Software technology – federated systems and architectures
- Execution platforms – workstations, grid computing, supercomputing
- Data access and management – data mining, formats and data interchange, common abstractions/metadata issues

Computing Infrastructure Activities

Rusyn, Stotts, Marron, MH Lee, et al.
Automation of visual inspection of genome analyses

Wei Wang, Xiang Zhang
Finding strongly correlated feature subsets where exploring the entire space is prohibitively expensive

D. Stotts, Keith Lee
AOP as middleware for toxicogenomics models
Modifications to ArrayTrack

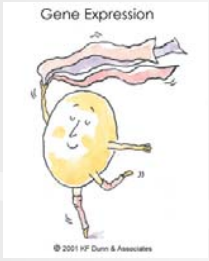
Alex Tropsha, Diane Pozefsky
Workflow automation for Toxicoinformatics

Mining High Dimensional Data

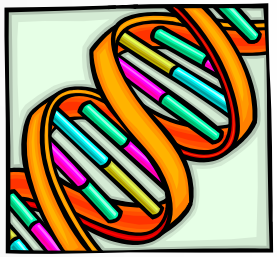
Wei Wang
Computer Science Dept.
http://www.cs.unc.edu/~weiwang

Biological Applications

Which genes underlie a disease?



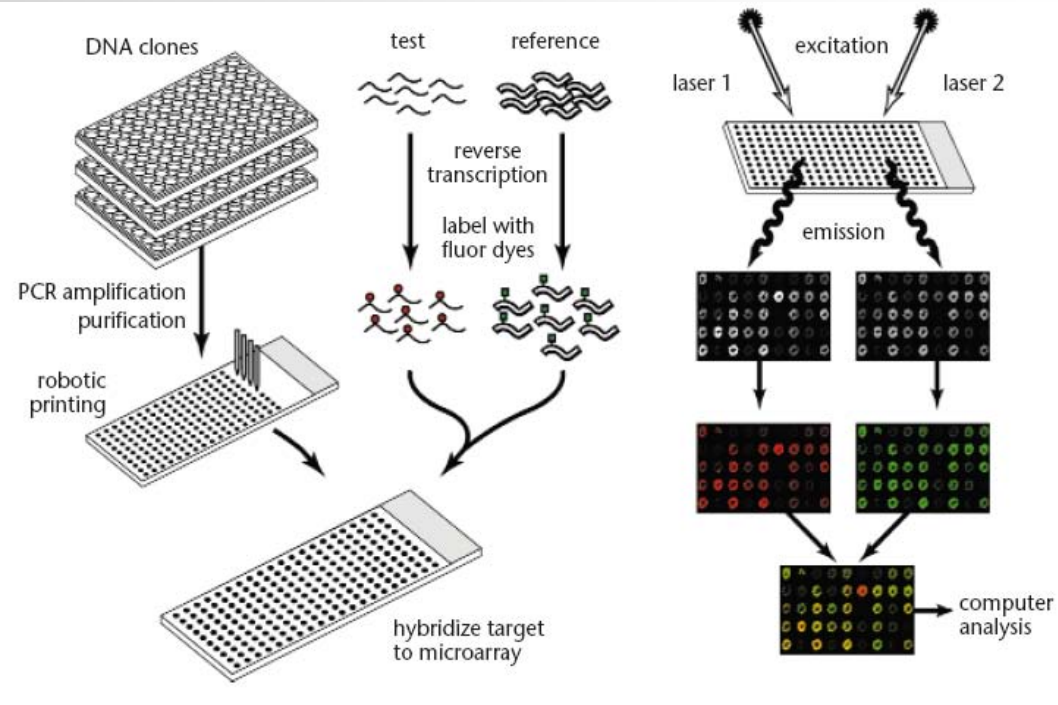
*The function of an organism is largely determined by **gene expression***



Which genes function differently in diseased people and healthy people?

Gene Expression Databases

- Goal of Microarray Techniques:
  - Determine which genes are activated and which genes are repressed.
  - Record the level activation or repression of all genes(10,000s).



Duggan et.al. 1999 Nature Genetics.

High Throughput Data Collection

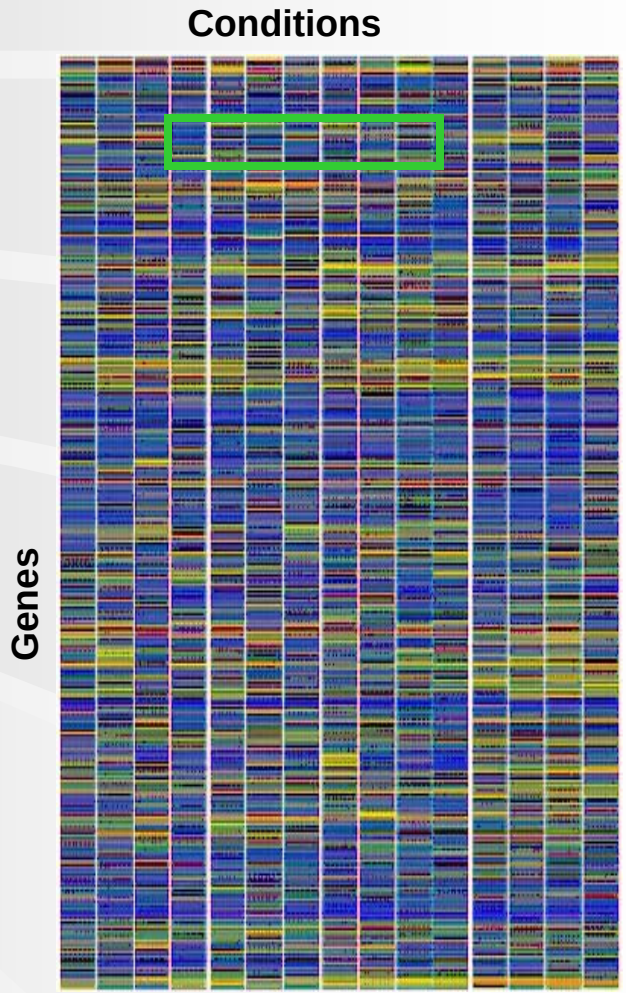
- Biological Applications
  - *Gene expression data*
    - 10, 000's genes by 100's conditions

$$\begin{matrix} \\ \text{Genes} \end{matrix} \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nm} \end{bmatrix}$$

Conditions

Time points

Tissue samples

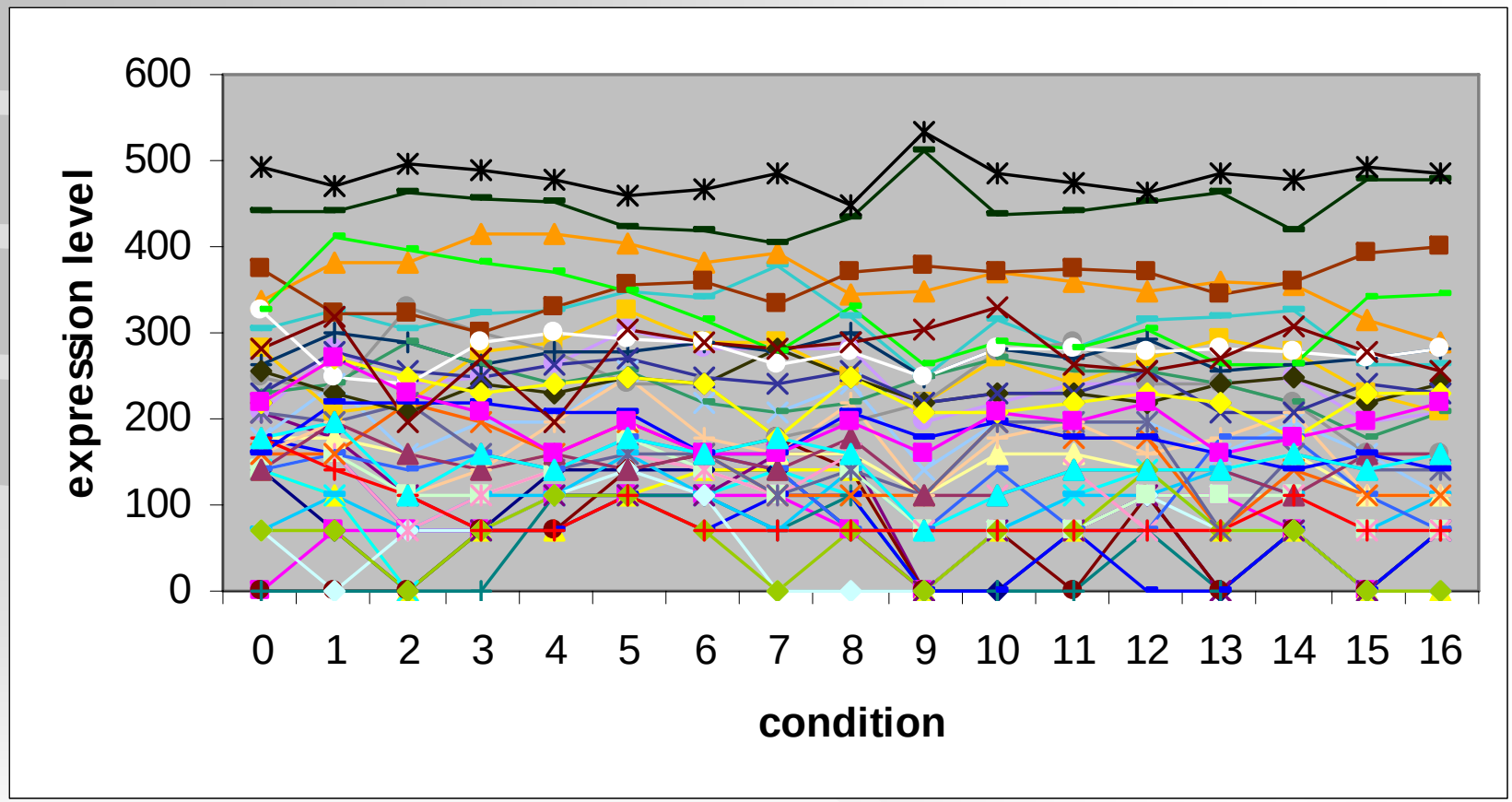


17 conditions

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
139	69	0	69	139	139	139	139	69	0	0	69	110	0	69	0	0
0	69	69	69	110	110	110	110	69	0	69	69	110	110	69	0	69
139	110	0	69	69	110	139	139	139	0	69	69	139	69	69	0	0
139	110	0	69	110	110	110	139	110	0	69	110	139	69	69	0	69
208	179	110	69	110	110	110	161	161	0	69	69	110	0	69	0	69
0	0	0	69	69	139	161	179	139	0	69	0	110	0	69	0	69
0	0	0	0	110	110	110	69	110	0	0	0	69	0	69	0	69
179	161	69	69	69	110	69	110	110	0	0	69	0	0	69	0	69
69	110	69	110	110	161	110	69	139	69	69	110	110	139	110	69	110
69	0	69	69	110	139	110	0	0	0	69	69	110	69	69	0	0
139	161	110	110	139	179	139	110	139	69	69	69	110	110	110	69	69
179	179	161	139	161	195	161	161	161	110	161	161	139	139	161	110	110
179	240	161	195	195	256	220	208	240	139	195	195	195	161	195	161	110
161	161	69	110	139	161	139	110	161	69	110	139	69	69	110	69	69
208	283	240	248	264	304	283	283	283	195	220	240	240	240	248	195	208
161	195	110	139	195	248	179	161	220	110	179	195	161	179	208	110	110
139	161	139	161	139	179	161	139	69	69	139	69	69	179	179	110	69
304	326	304	322	326	350	340	376	318	248	314	283	314	318	326	264	264
69	69	0	69	110	110	69	0	69	0	69	69	139	69	69	0	0
283	208	220	277	289	326	289	289	248	220	271	240	271	294	277	230	208
337	383	383	413	414	403	381	393	343	350	369	358	347	358	356	314	289
161	161	220	195	161	195	161	110	110	110	195	179	179	69	139	110	110
208	195	220	161	139	161	161	110	139	110	195	195	195	69	161	139	139
248	230	330	300	277	240	240	179	195	220	277	289	240	240	220	161	161
264	300	289	264	277	277	289	277	300	248	283	271	294	256	264	271	283
230	240	289	264	240	256	220	208	220	248	271	256	256	240	220	179	208
439	442	464	456	451	422	417	403	432	510	438	442	450	462	419	476	476
256	230	208	240	230	248	240	283	248	220	230	230	220	240	248	220	240
374	322	322	300	330	356	361	333	369	376	369	374	369	343	361	393	399
139	195	161	139	161	139	161	139	179	110	110	139	139	139	110	161	161
230	277	256	248	264	271	248	240	256	220	230	230	256	208	208	240	230
494	470	498	488	477	460	466	484	449	532	485	473	464	487	477	492	484
326	248	240	289	300	294	289	264	277	248	283	283	277	283	277	271	283
179	139	110	69	69	110	69	69	69	69	69	69	69	69	110	69	69
326	411	397	383	371	347	314	277	330	264	289	283	304	264	264	340	343
161	220	220	220	208	208	161	161	208	179	195	179	179	161	139	161	139
220	271	248	230	240	248	240	179	248	208	208	220	230	220	179	230	230
220	271	230	208	161	195	161	161	195	161	208	195	220	161	179	195	220
179	195	110	161	139	179	161	179	161	69	110	139	139	139	161	139	161
283	318	195	271	195	304	289	283	289	304	330	264	256	271	309	277	256

40 genes

Coherent Cluster

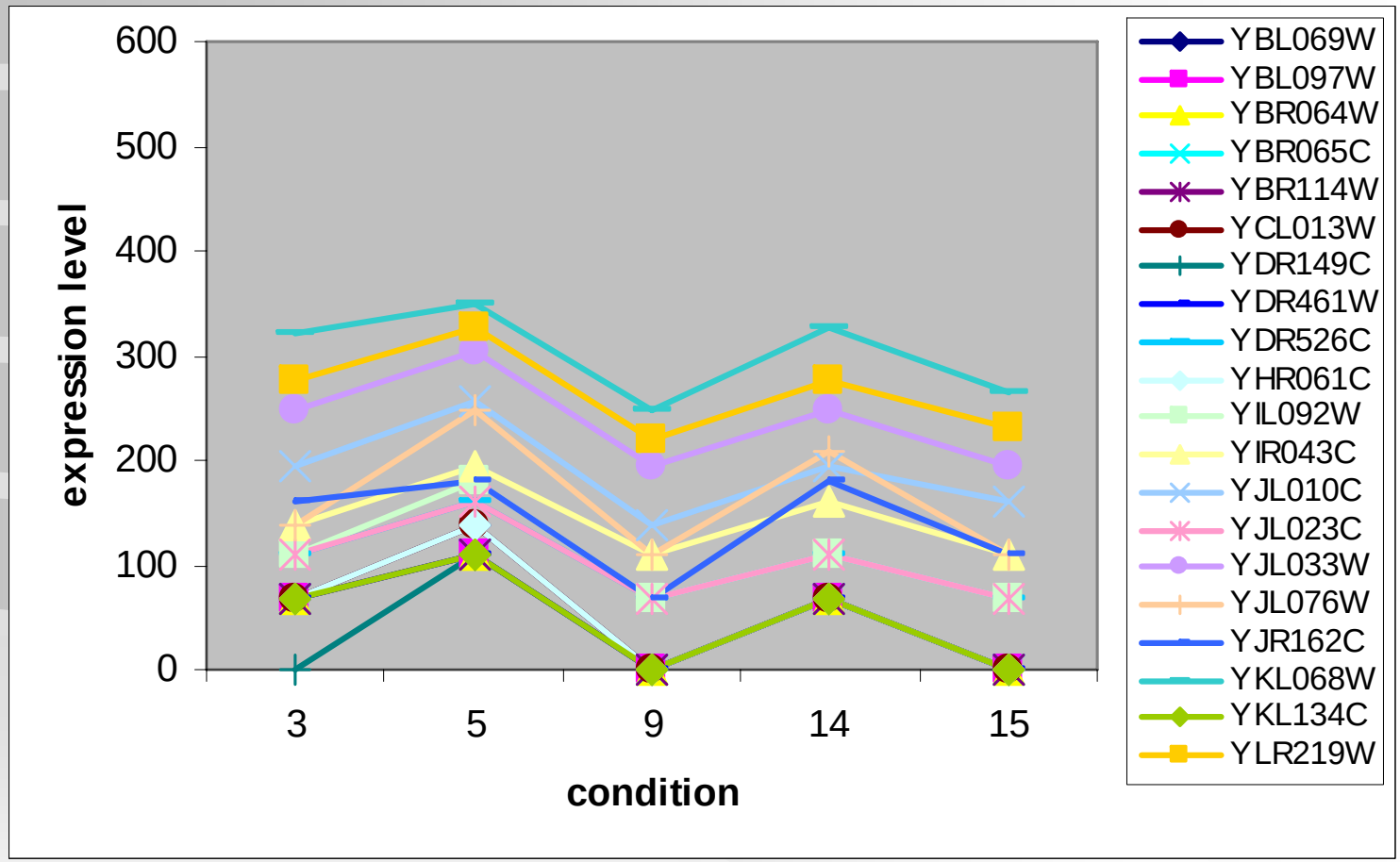


17 conditions

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
139	69	0	69	139	139	139	139	69	0	0	69	110	0	69	0	0
0	69	69	69	110	110	110	110	69	0	69	69	110	110	69	0	69
139	110	0	69	69	110	139	139	139	0	69	69	139	69	69	0	0
139	110	0	69	110	110	110	139	110	0	69	110	139	69	69	0	69
208	179	110	69	110	110	110	161	161	0	69	69	110	0	69	0	69
0	0	0	69	69	139	161	179	139	0	69	0	110	0	69	0	69
0	0	0	0	110	110	110	69	110	0	0	0	69	0	69	0	69
179	161	69	69	69	110	69	110	110	0	0	69	0	0	69	0	69
69	110	69	110	110	161	110	69	139	69	69	110	110	139	110	69	110
69	0	69	69	110	139	110	0	0	0	69	69	110	69	69	0	0
139	161	110	110	139	179	139	110	139	69	69	69	110	110	110	69	69
179	179	161	139	161	195	161	161	161	110	161	161	139	139	161	110	110
179	240	161	195	195	256	220	208	240	139	195	195	195	161	195	161	110
161	161	69	110	139	161	139	110	161	69	110	139	69	69	110	69	69
208	283	240	248	264	304	283	283	283	195	220	240	240	240	248	195	208
161	195	110	139	195	248	179	161	220	110	179	195	161	179	208	110	110
139	161	139	161	139	179	161	139	69	69	139	69	69	179	179	110	69
304	326	304	322	326	350	340	376	318	248	314	283	314	318	326	264	264
69	69	0	69	110	110	69	0	69	0	69	69	139	69	69	0	0
283	208	220	277	289	326	289	289	248	220	271	240	271	294	277	230	208
337	383	383	413	414	403	381	393	343	350	369	358	347	358	356	314	289
161	161	220	195	161	195	161	110	110	110	195	179	179	69	139	110	110
208	195	220	161	139	161	161	110	139	110	195	195	195	69	161	139	139
248	230	330	300	277	240	240	179	195	220	277	289	240	240	220	161	161
264	300	289	264	277	277	289	277	300	248	283	271	294	256	264	271	283
230	240	289	264	240	256	220	208	220	248	271	256	256	240	220	179	208
439	442	464	456	451	422	417	403	432	510	438	442	450	462	419	476	476
256	230	208	240	230	248	240	283	248	220	230	230	220	240	248	220	240
374	322	322	300	330	356	361	333	369	376	369	374	369	343	361	393	399
139	195	161	139	161	139	161	139	179	110	110	139	139	139	110	161	161
230	277	256	248	264	271	248	240	256	220	230	230	256	208	208	240	230
494	470	498	488	477	460	466	484	449	532	485	473	464	487	477	492	484
326	248	240	289	300	294	289	264	277	248	283	283	277	283	277	271	283
179	139	110	69	69	110	69	69	69	69	69	69	69	69	110	69	69
326	411	397	383	371	347	314	277	330	264	289	283	304	264	264	340	343
161	220	220	220	208	208	161	161	208	179	195	179	179	161	139	161	139
220	271	248	230	240	248	240	179	248	208	208	220	230	220	179	230	230
220	271	230	208	161	195	161	161	195	161	208	195	220	161	179	195	220
179	195	110	161	139	179	161	179	161	69	110	139	139	139	161	139	161
283	318	195	271	195	304	289	283	289	304	330	264	256	271	309	277	256

40 genes

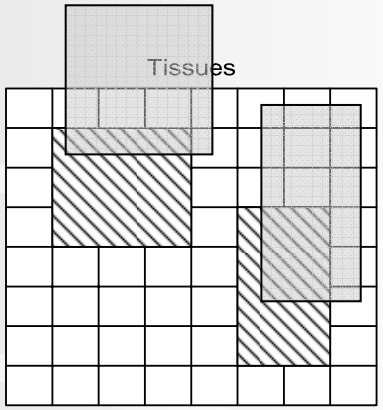
Coherent Cluster

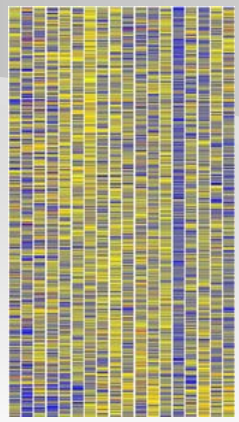


Co-regulated genes

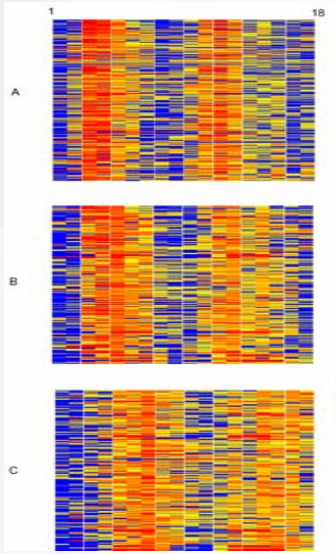
Subspace Clustering

- A subset of genes has local similarity under only a subset of experimental conditions or tissue samples.



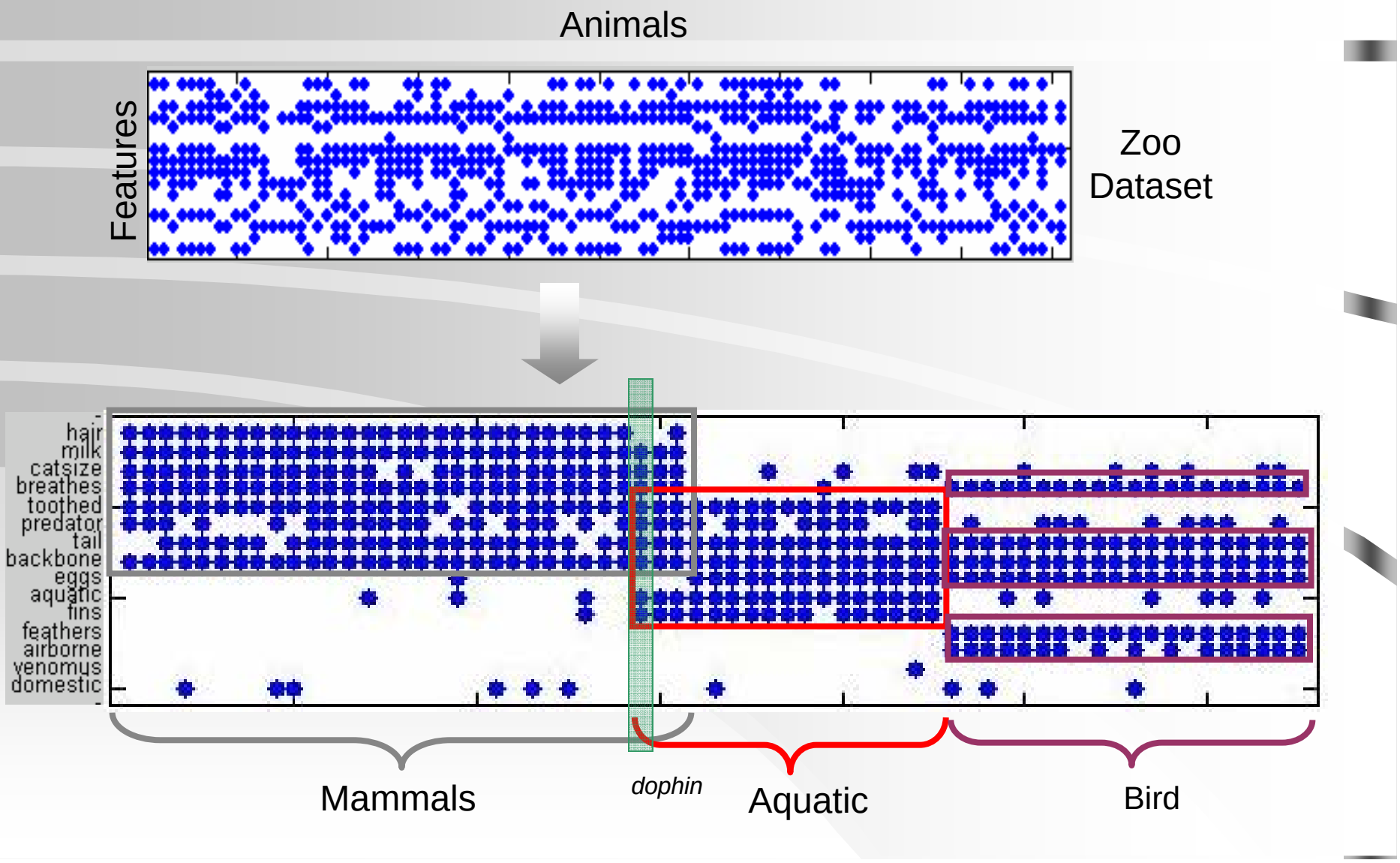


Subspace Clustering



Co-expressed genes

An example

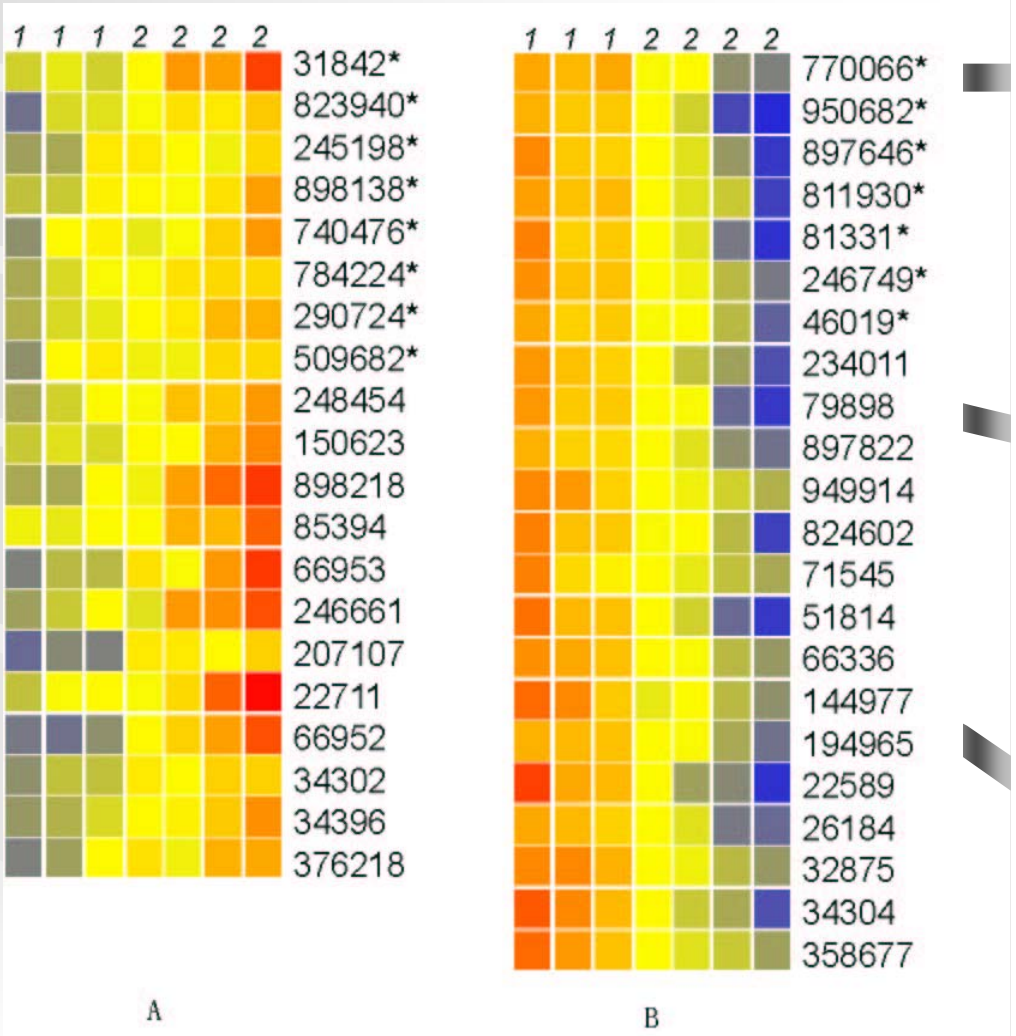

Animals
Features
Zoo
Dataset
hair
milk
catsize
breathes
toothed
predator
tail
backbone
eggs
aquatic
fins
feathers
airborne
venomus
domestic
dophin
Mammals
Aquatic
Bird

Challenges

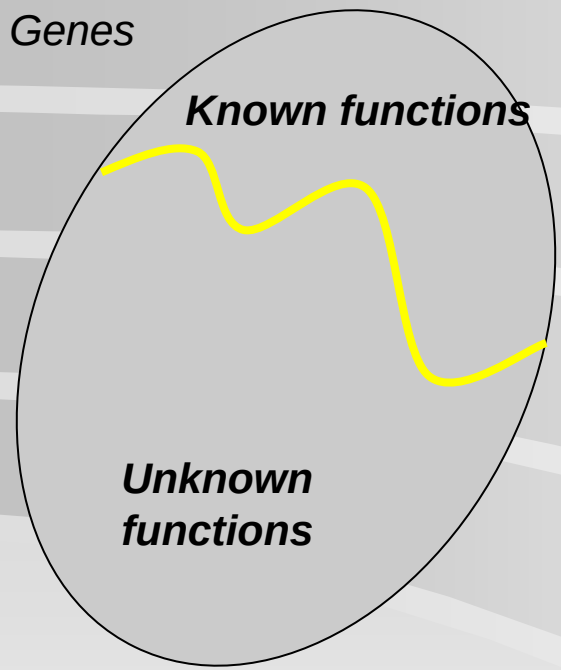
- Curse of high dimensionality
  - Perform clustering in subspaces
  - The number of subspaces is exponential
- Clusters are not independent
    - Clusters in different subspaces may overlap
    - Clusters should be relevant to application domains
- Noise may corrupt patterns
- The algorithms need to scale to large high dimensional datasets
- There are many open problems yet to be solved

Breast Cancer Classification

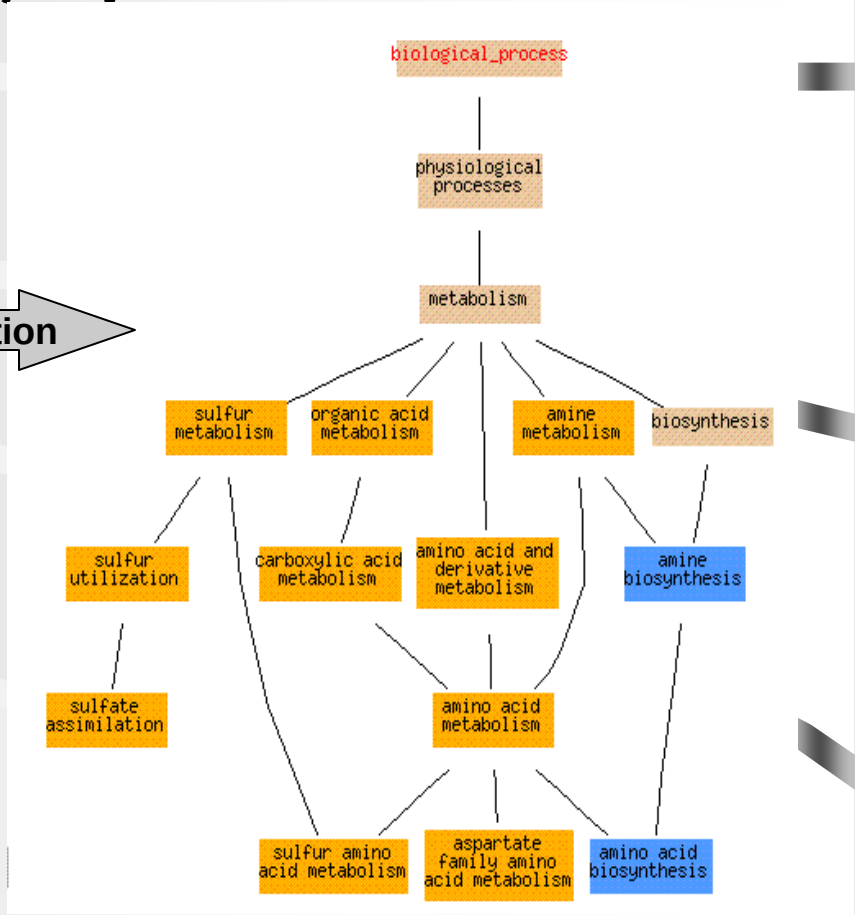
- Two clusters with opposite trends
  - * are previously discovered genes
  - 88% of 176 identified genes are found in at least one OP-Cluster
  - More genes with consistent trend are discovered.



Gene Function Annotation



Function Annotation

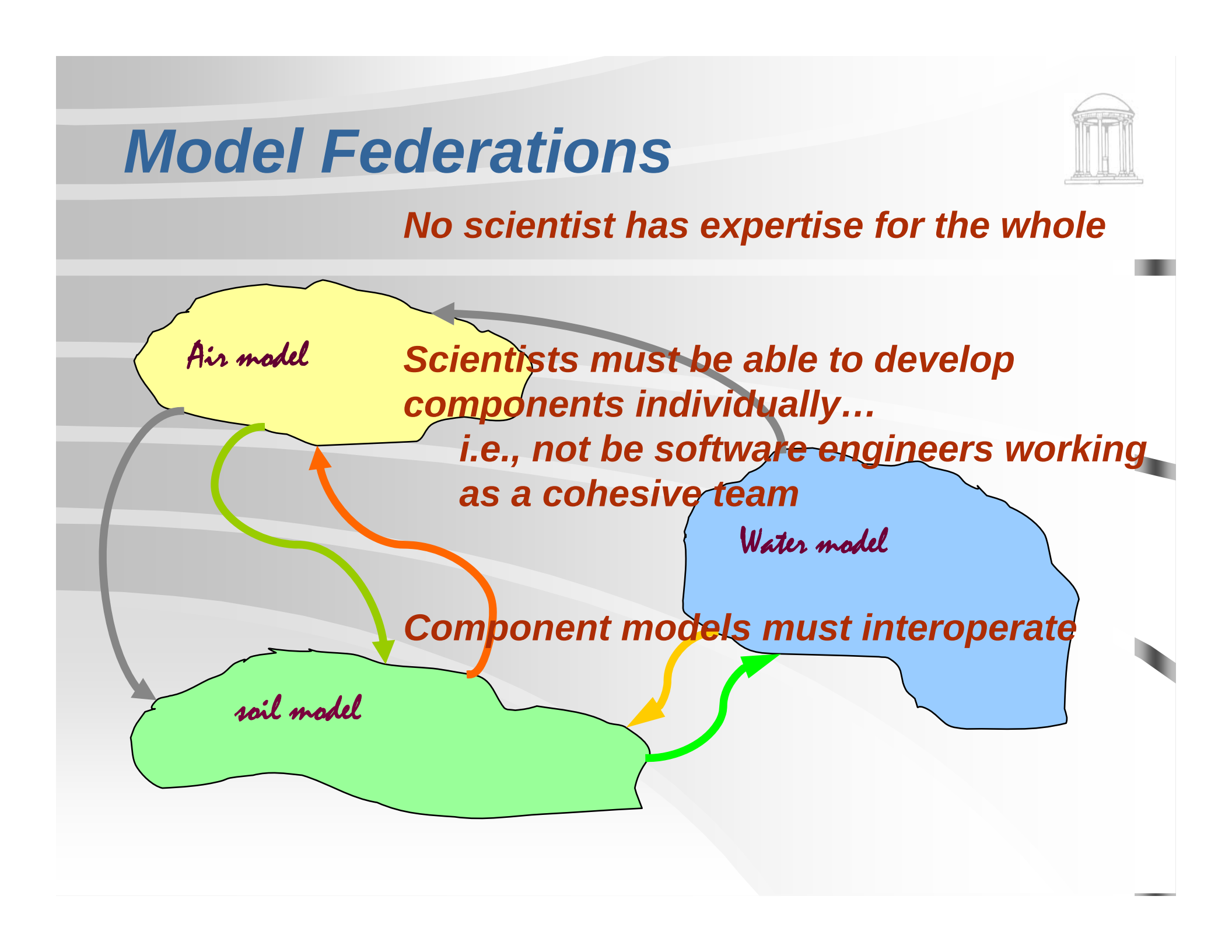


A fragment of Gene Ontology

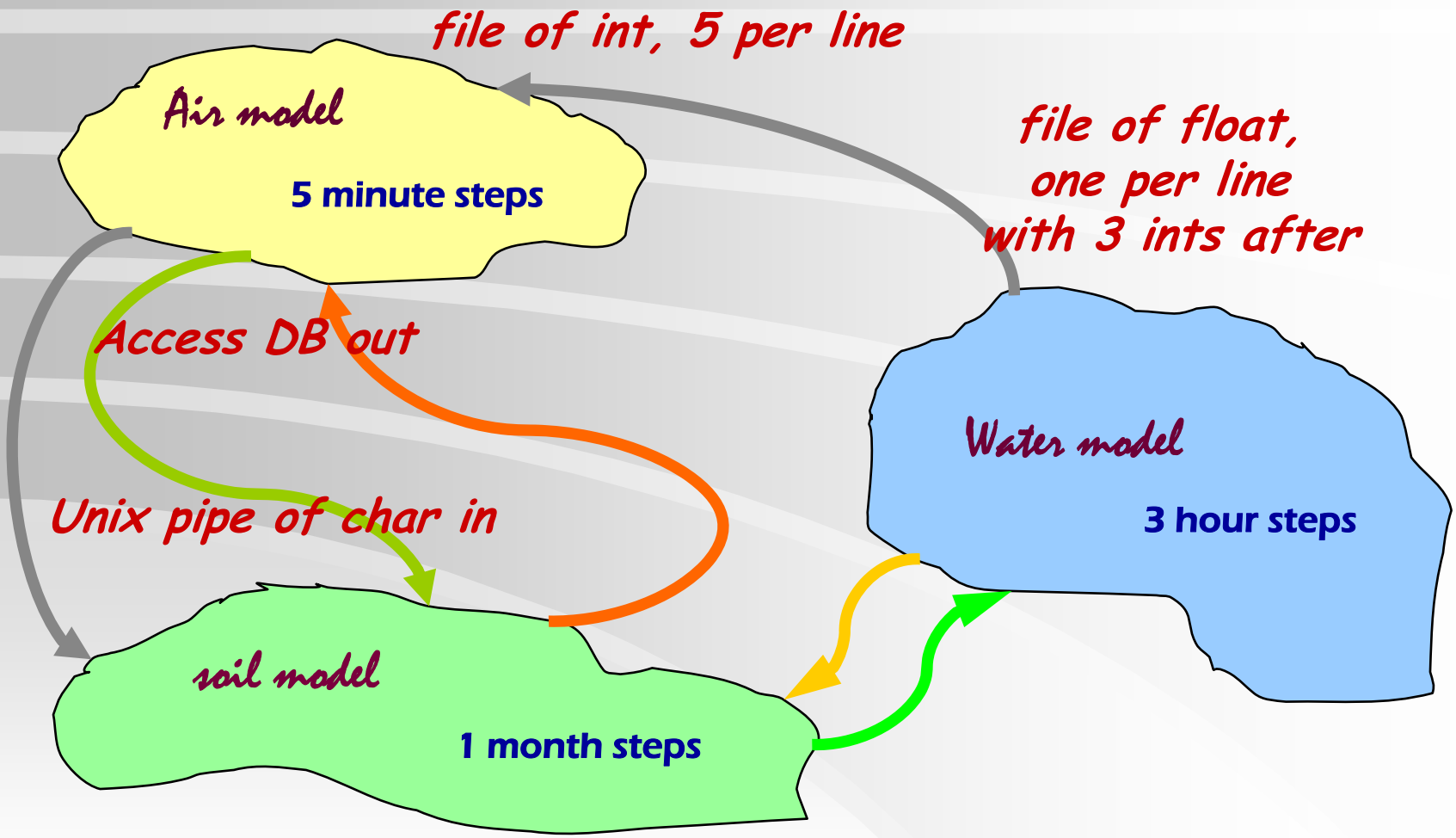
Scientific Model Federations

David Stotts
Computer Science Dept.
http://www.cs.unc.edu/~stotts

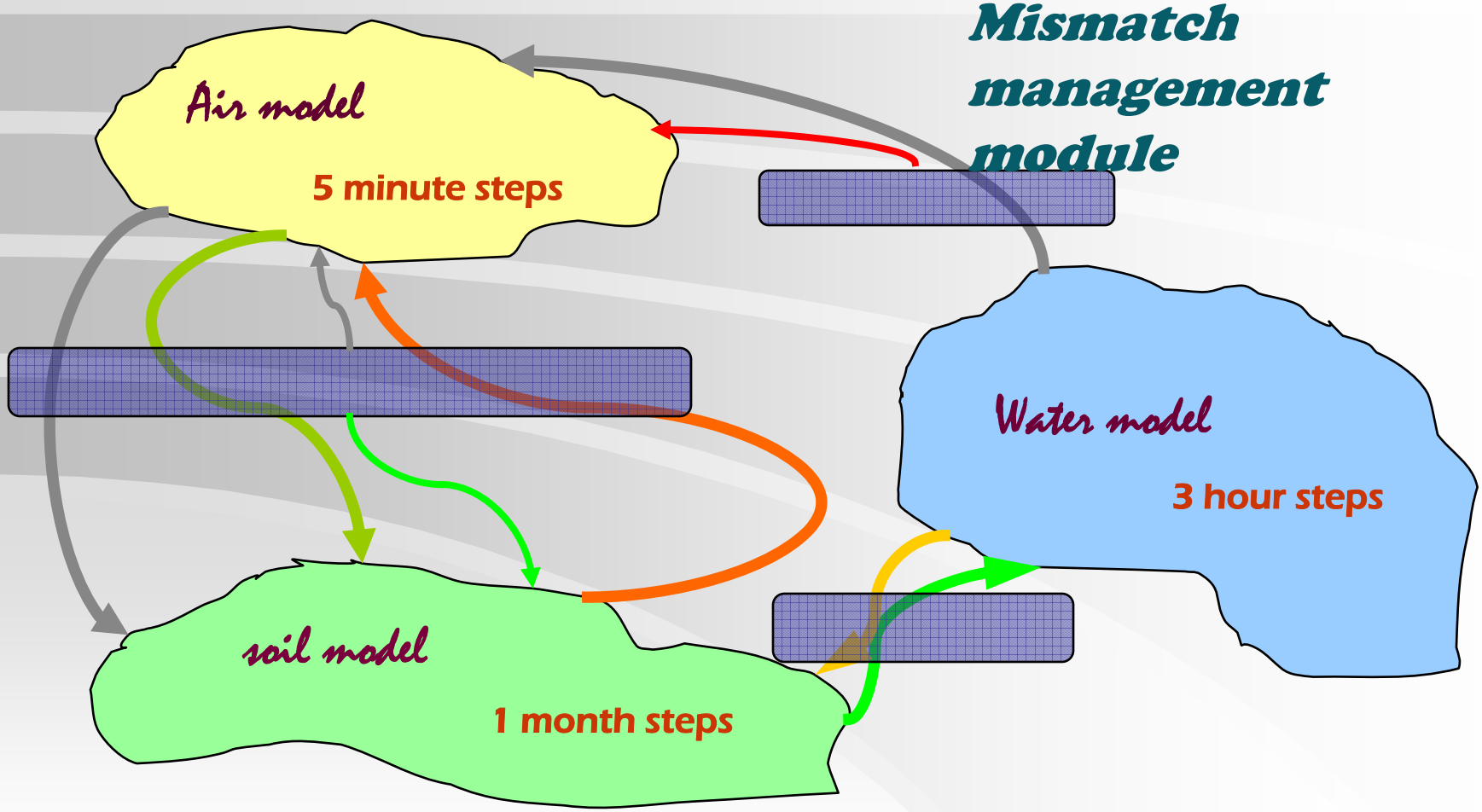
Model Federations



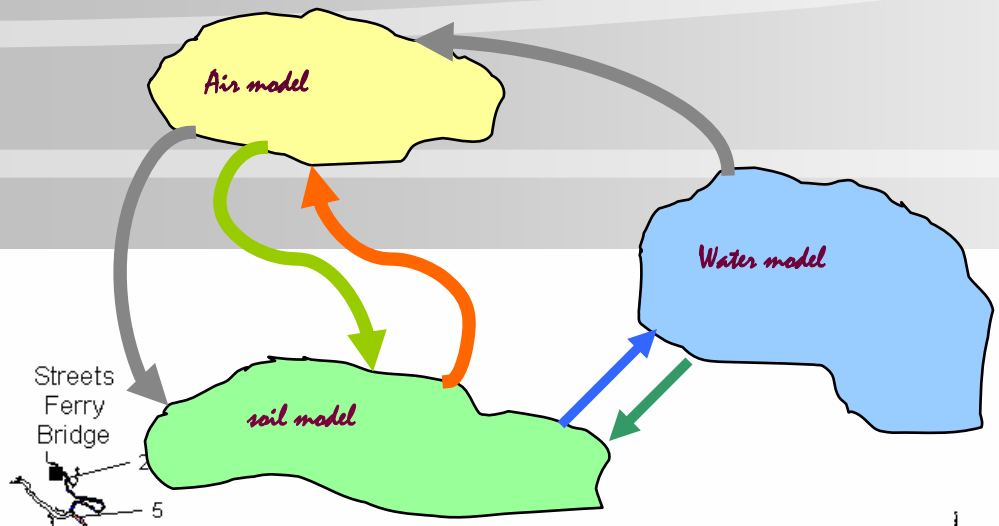
"Multi-media" Modeling

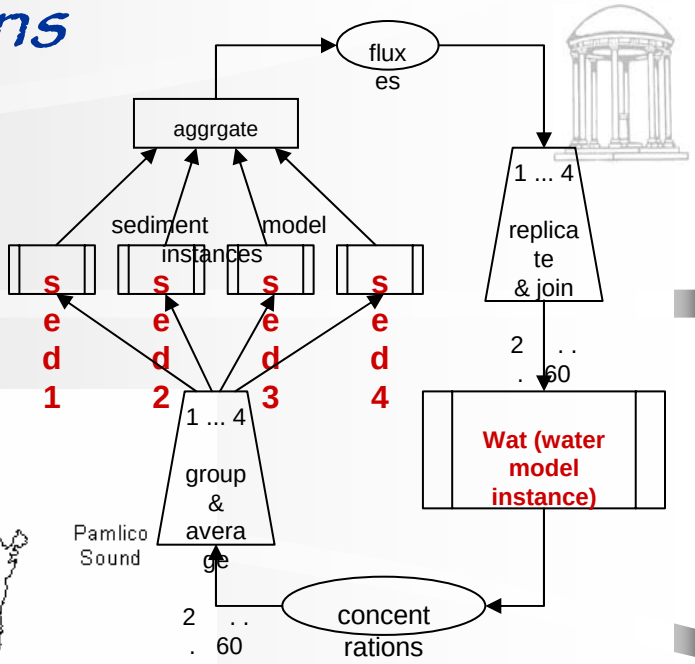


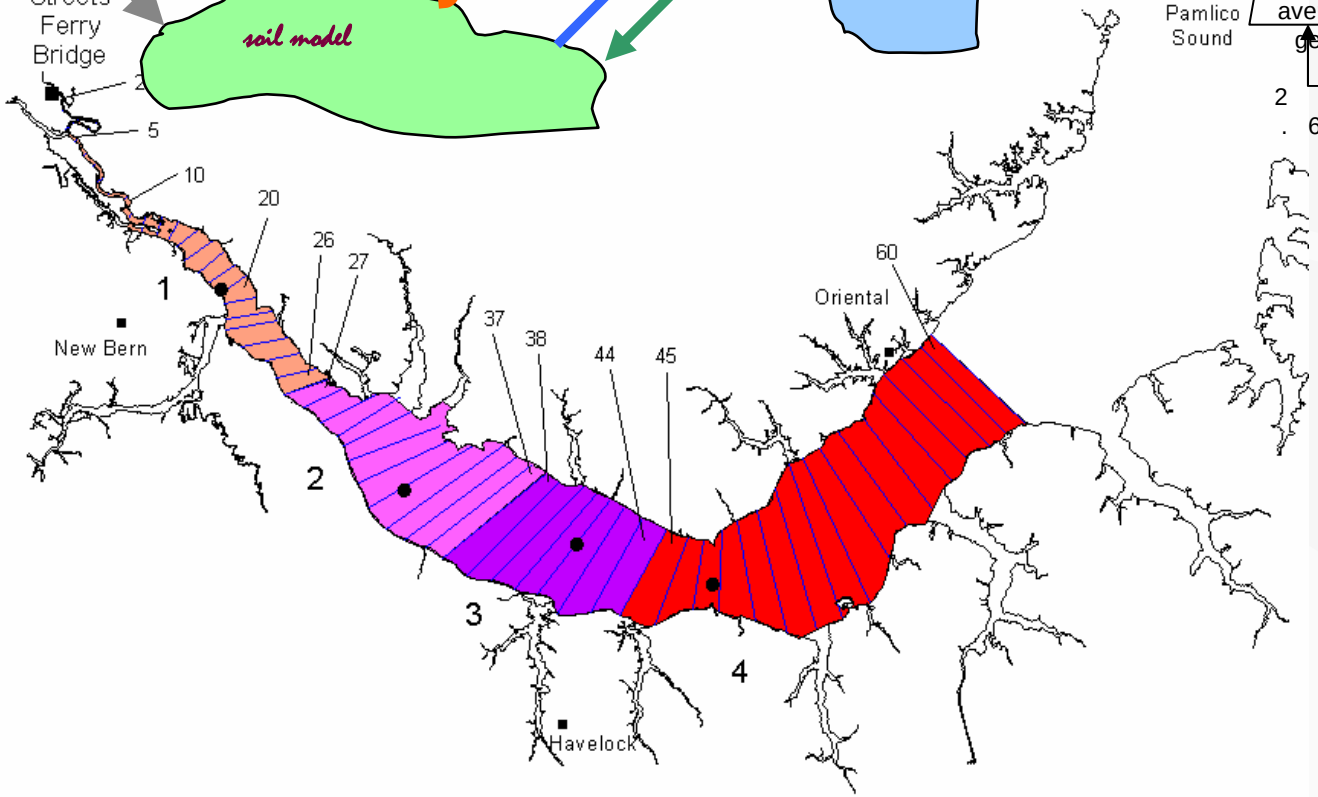
Model Federations



JDeco Functional Federations







New Areas

Virtual Human Lung

biochemistry, fluid dynamics, protein motors and cilia mechanics, cell physiology...
for cystic fibrosis research

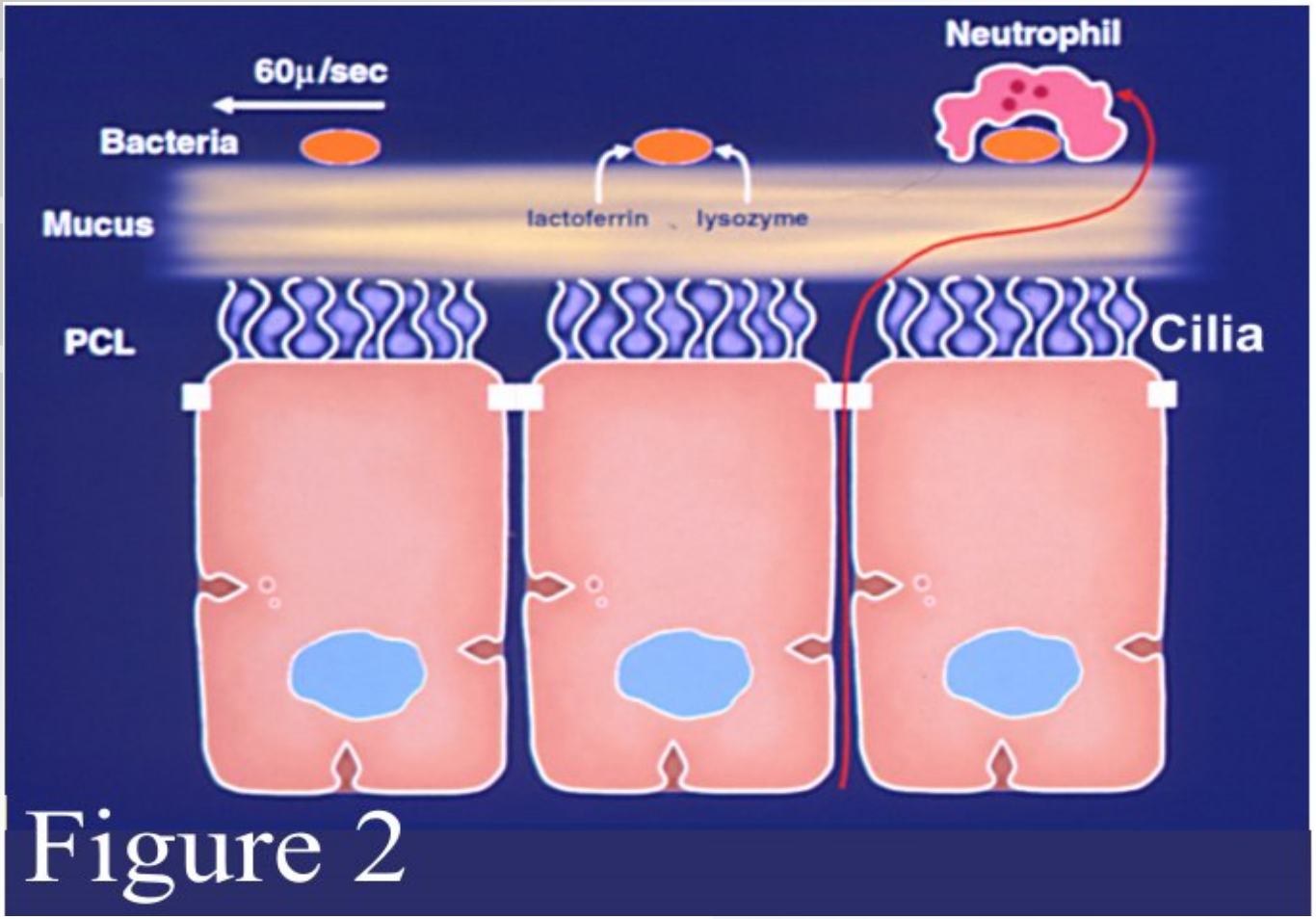
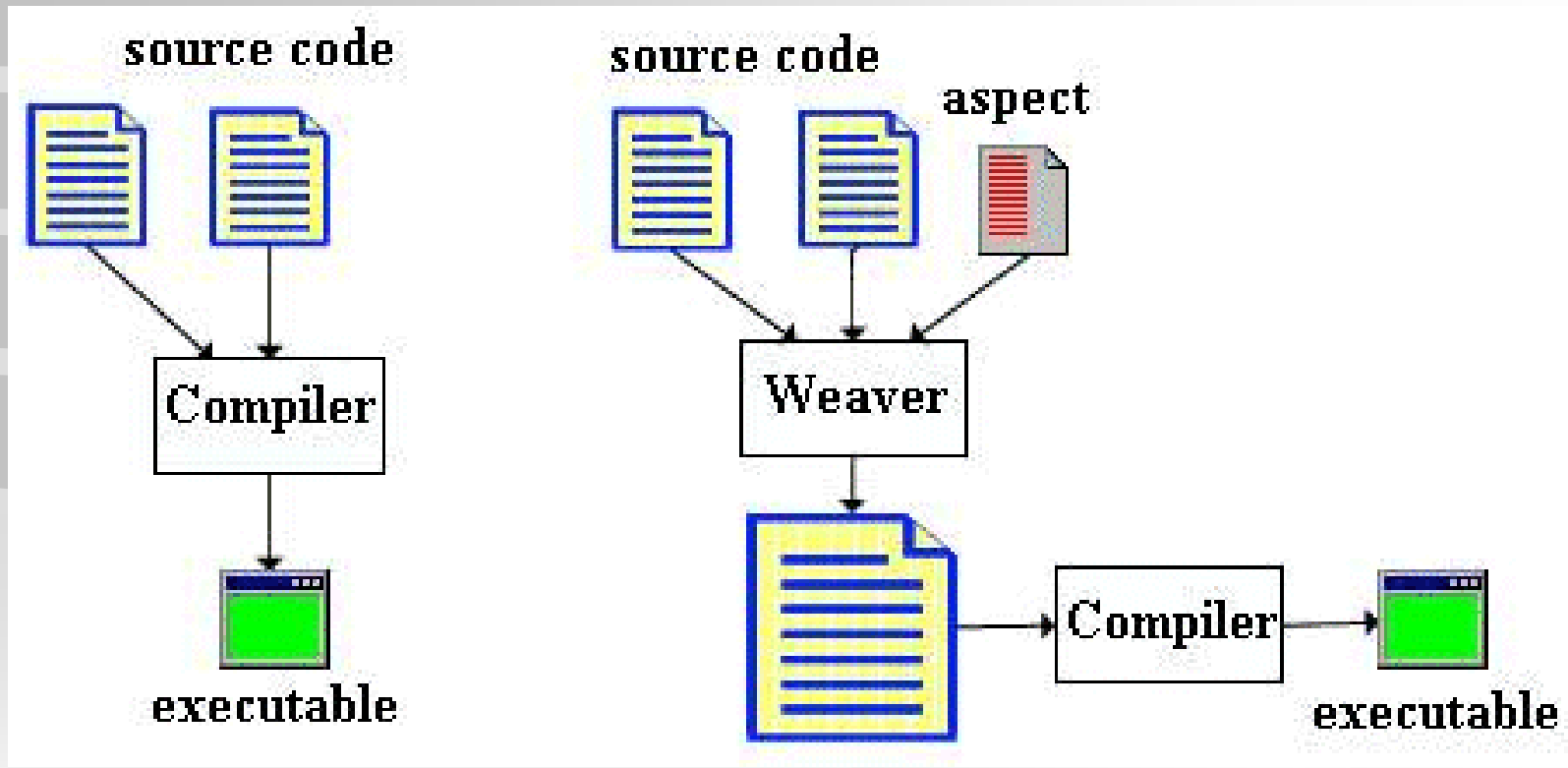
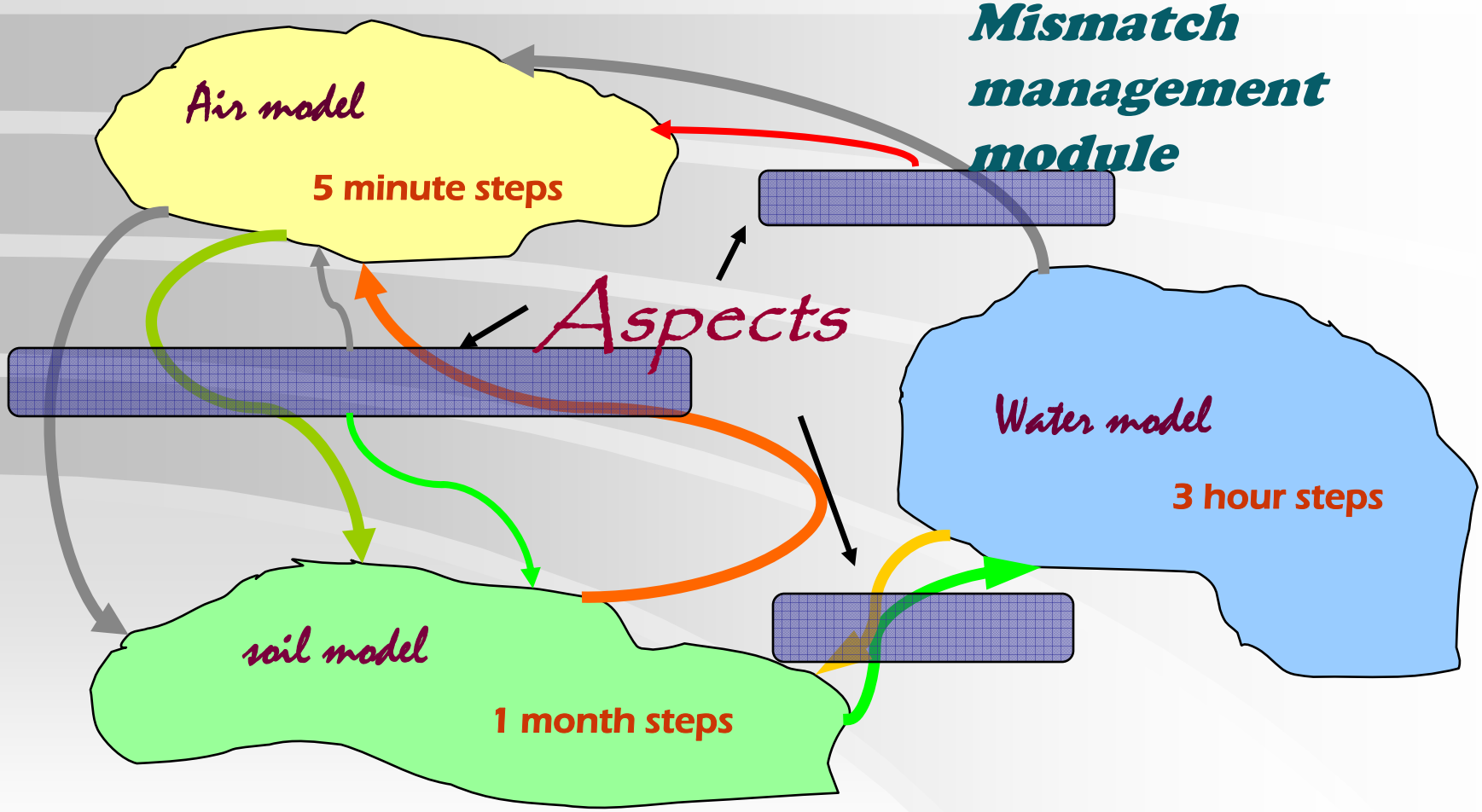


Figure 2

AOP: Aspect-Oriented Programming



Model Federations



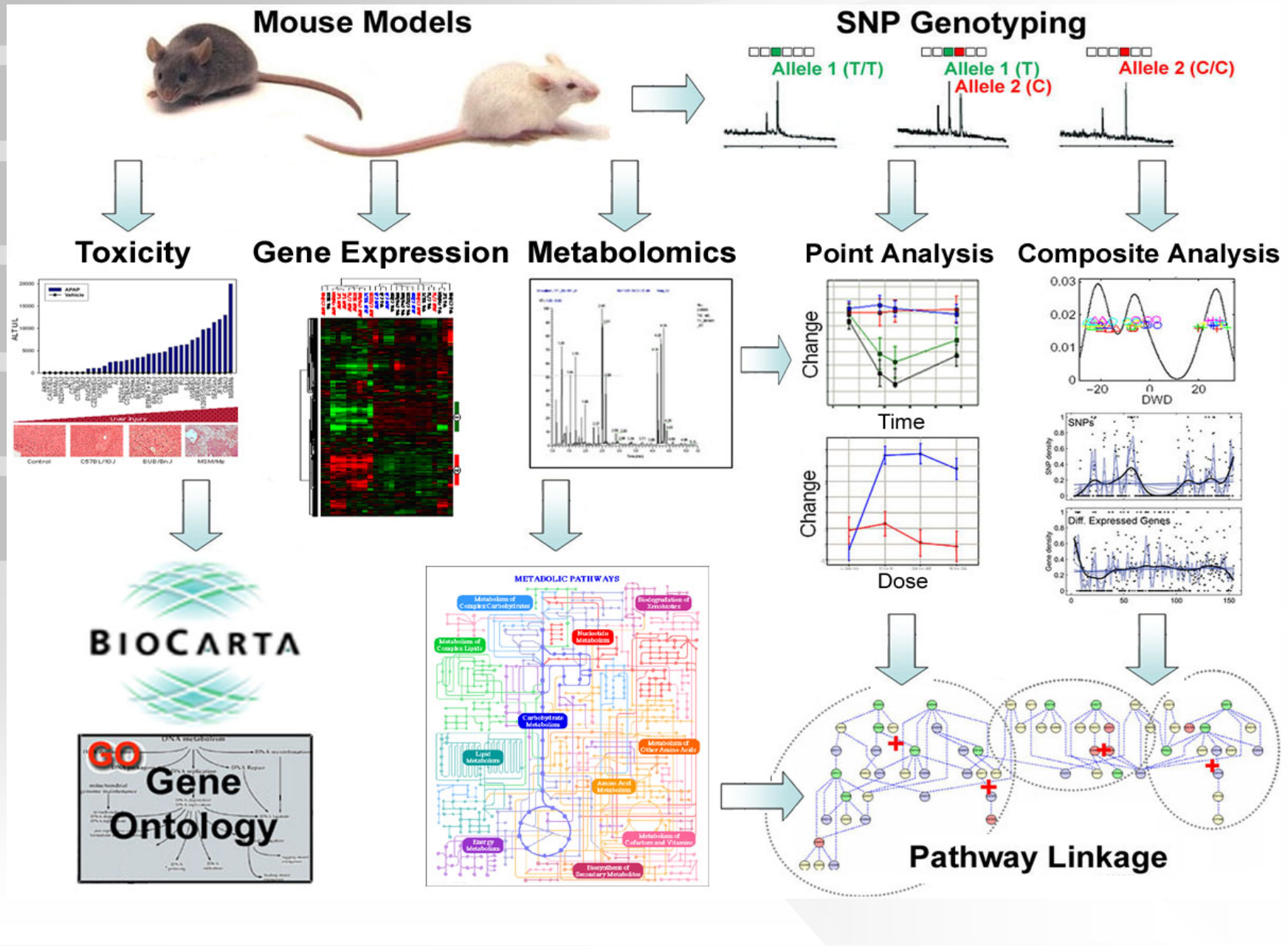
Extension of Existing Tools

Federations are for situations where tools are disjoint, have different developers

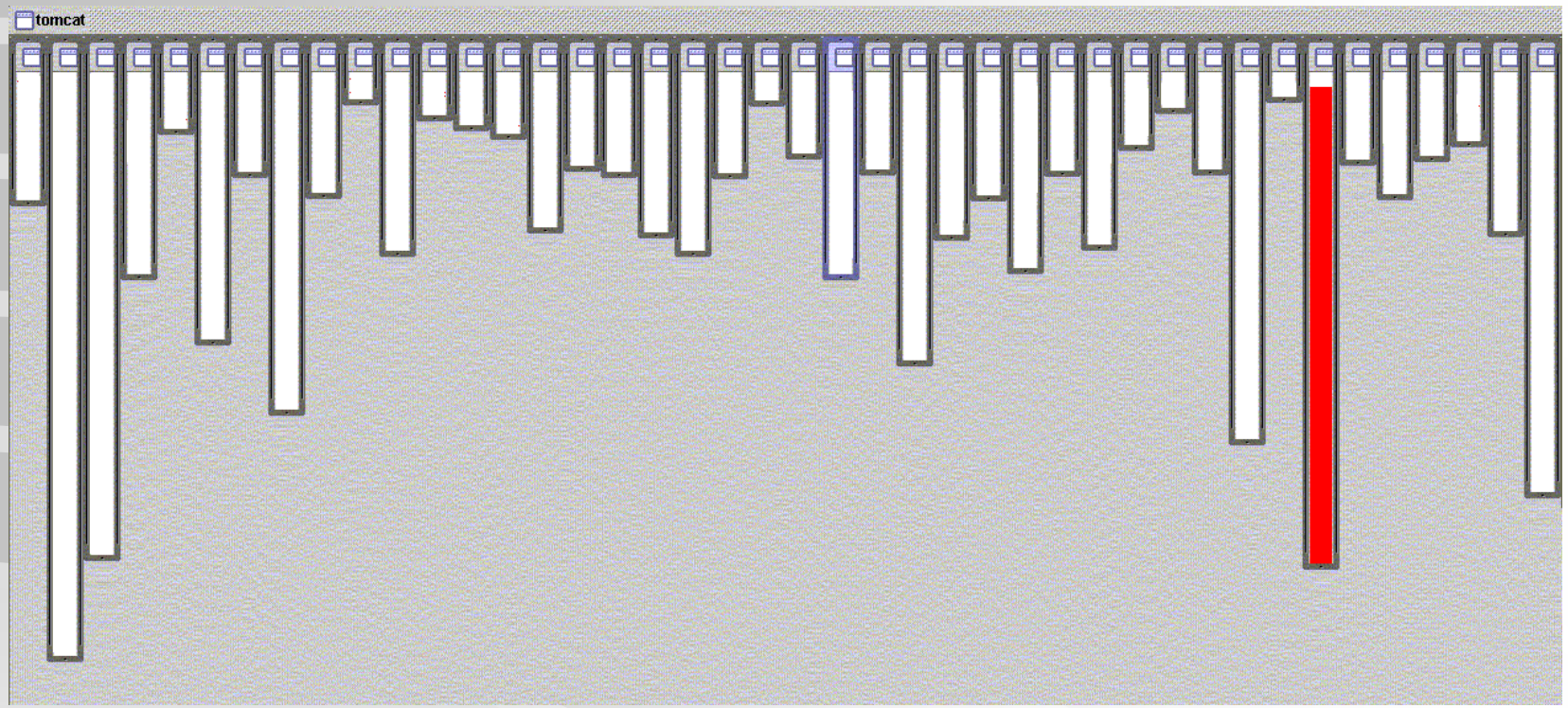
Example: ArrayTrack

We have no source code
but users like the data management
We wish to add new analysis functions

“Systems Toxicology” Approach

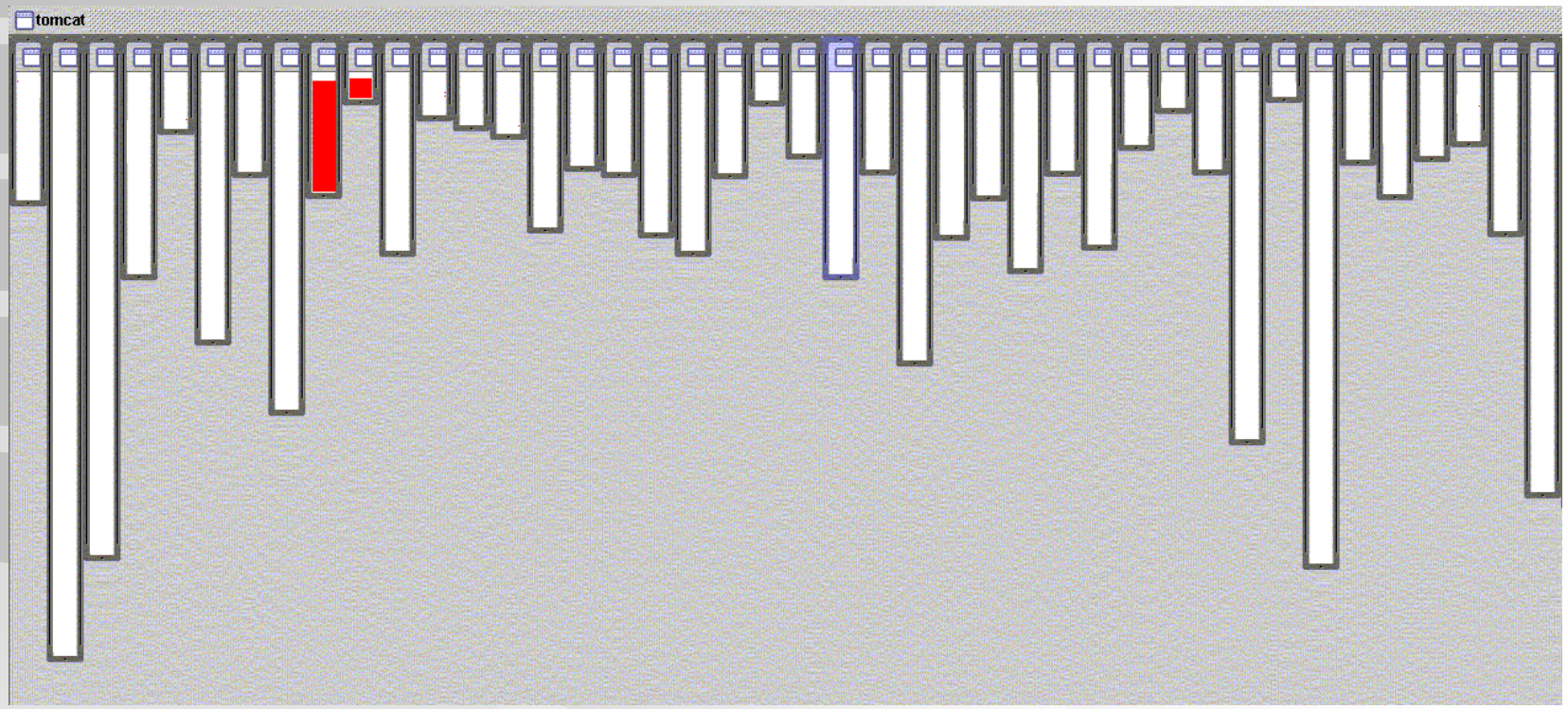


Good modularity



- XML parsing in org.apache.tomcat
  - red shows relevant lines of code
  - nicely fits in one box

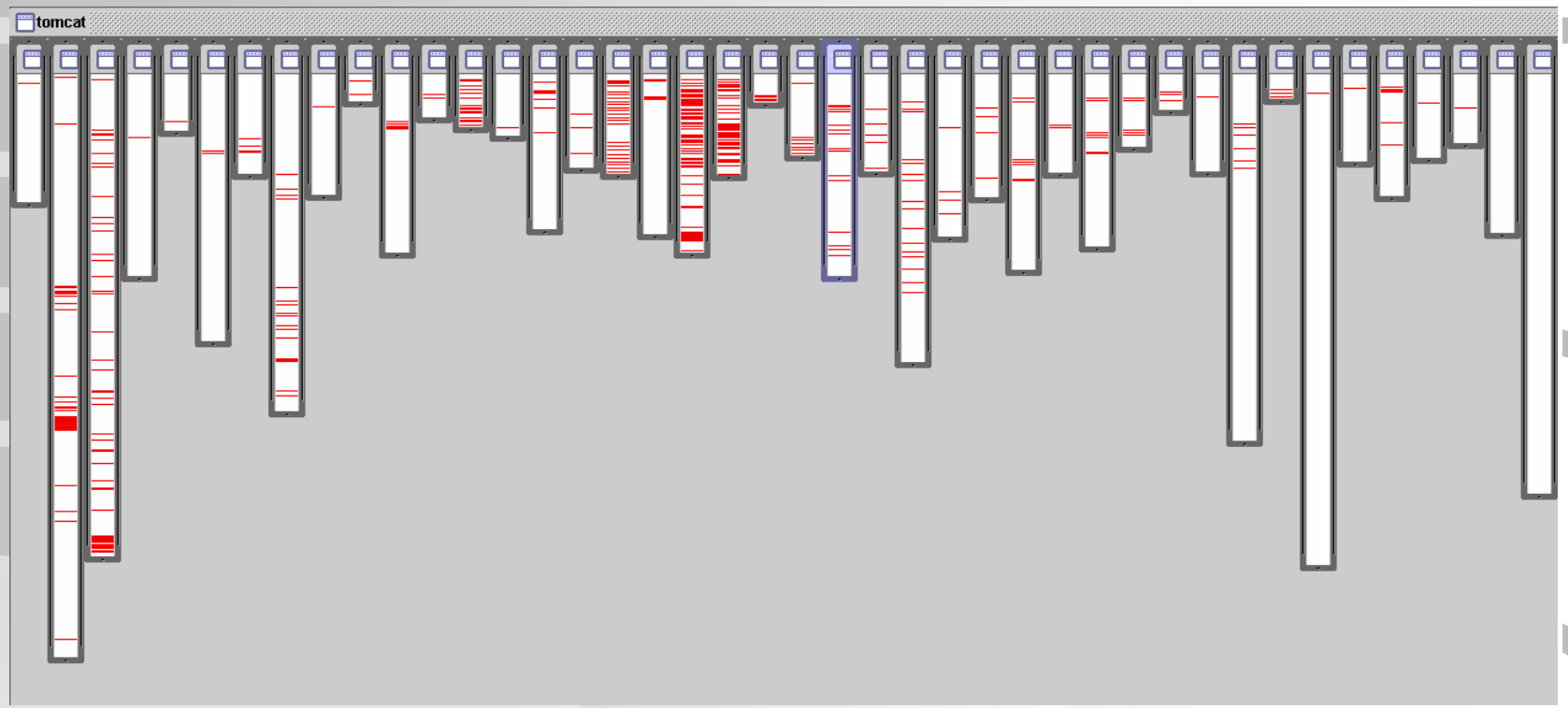
Good modularity



- URL pattern matching in org.apache.tomcat
  - red shows relevant lines of code
  - nicely fits in two boxes (using inheritance)

logging is not modularized

But problems exist...



- logging in org.apache.tomcat
  - red shows lines of code that handle logging
  - not in just one place
  - not even in a small number of places

Logging, zoomed in

//From ContextManager

```
public void service( Request rrequest, Response rresponse ) {
 // log( "New  request " + rrequest );
 try {
     //     System.out.print("A");
     rrequest.setContextManager( this );
     rrequest.setResponse(rresponse);
     rresponse.setRequest(rrequest);

     // wront request - parsing error
     int status=rresponse.getStatus();

     if( status < 400 )
  status= processRequest( rrequest );

     if(status==0)
  status=authenticate( rrequest, rresponse );
     if(status == 0)
  status=authorize( rrequest, rresponse );
     if( status == 0 ) {
  rrequest.getWrapper().handleRequest(rrequest,
          rresponse);
     } else {
  // something went wrong
  handleError( rrequest, rresponse, null, status );
     }
 } catch (Throwable t) {
     handleError( rrequest, rresponse, t, 0 );
 }
 // System.out.print("B");
 try {
     rresponse.finish();
     rrequest.recycle();
     rresponse.recycle();
 } catch( Throwable ex ) {
     if(debug>0) log( "Error closing request " + ex);
 }
 // log( "Done with request " + rrequest );
 // System.out.print("C");
 return;
}
```

```
// log( "New  request " + rrequest );

// System.out.print(“A”);

// System.out.print("B");

if(debug>0)
     log("Error closing request " + ex);

// log("Done with request " + rrequest);

// System.out.print("C");
```

Contacts

- Imran Shah, RTP
  - Virtual liver project
- Weida Tong, FDA NCTR
  - ArrayTrack