Knowledge-Oriented Analysis of Mycroarray Data

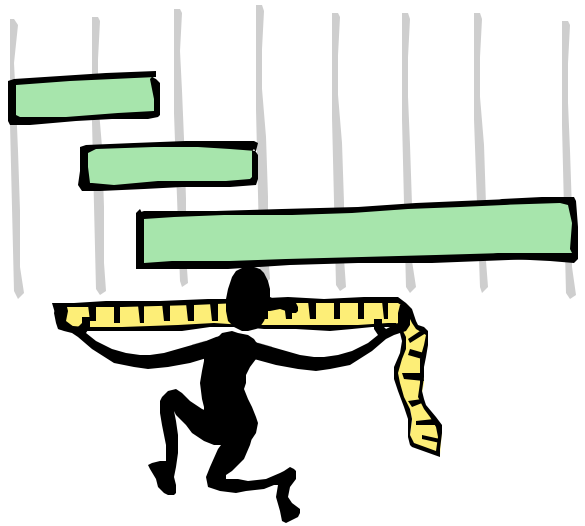
Avoiding Paralysis of Analysis:

Building an Intellectual Prosthesis

I. Jurisica

Goals

- Parallel analysis of gene expressions
  - Improved understanding of tumorigenesis
  - Tumor classification
- Individualized medicine
  - Improved diagnosis, prognostics, treatment planning & adjustment
  - Targetted therapy & drug design/use
  - Informed patient

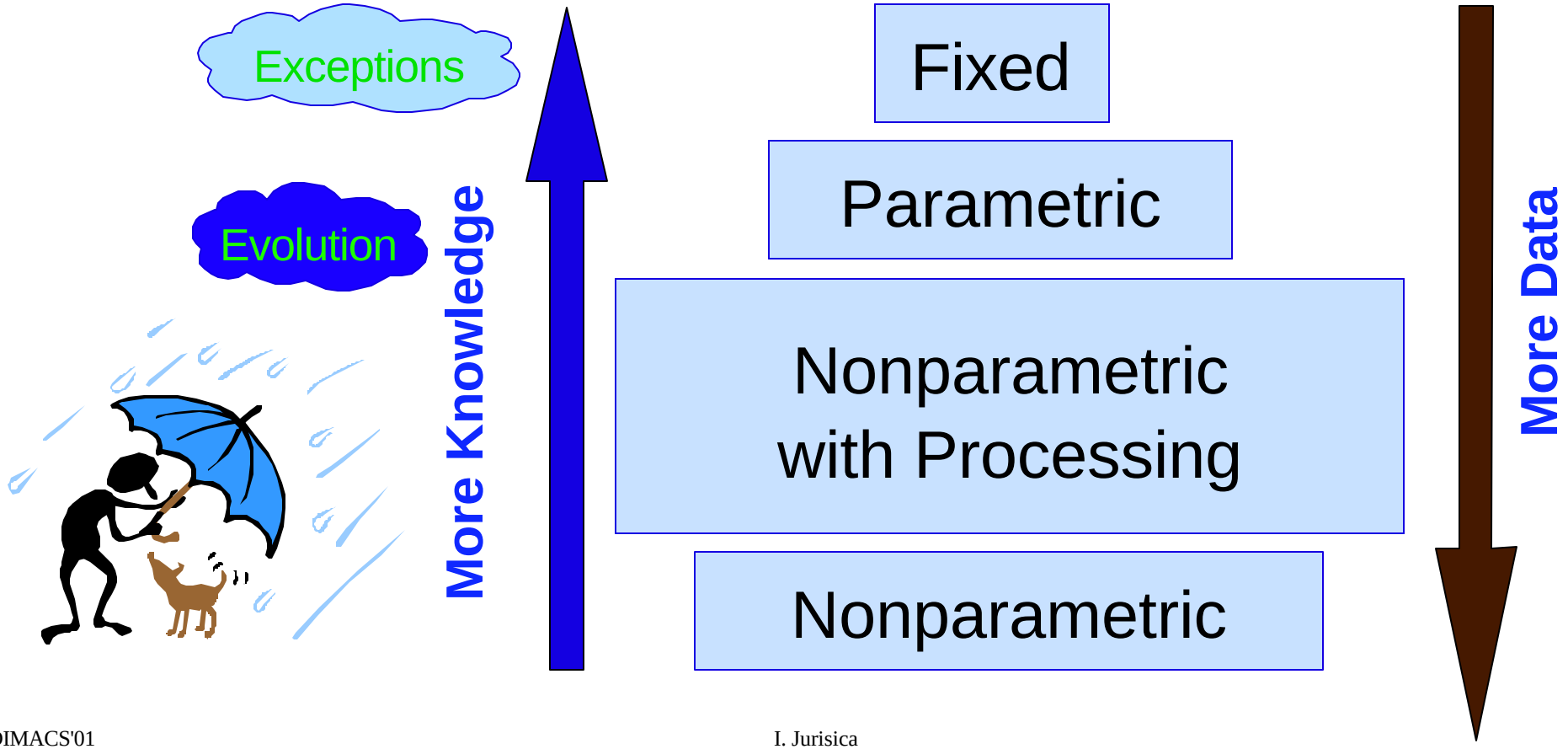


Problems

- Multi-dimensionality
  - many degrees of freedom, few datapoints
- Noise
  - Imprecision, variation
  - Low number of repeats
- Non-independebility
- Non-linearity
- DBs change
- Integration of results with other DBs & multiple experiments

Intellectual Prosthesis

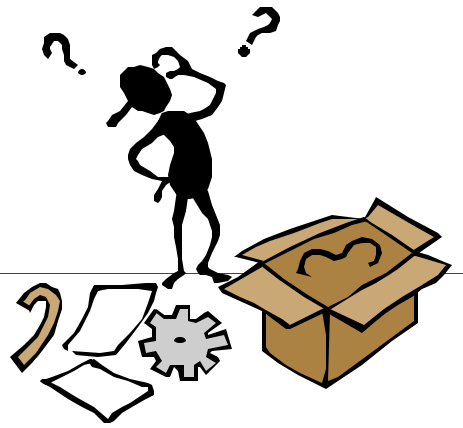
- Finding appropriate model to support reasoning



Analysis

- Clustering organizes observations into groups by max. iner-cluster and min. inter-cluster similarity
- Classification/prediction assigns an observation to a class (finite/infinite)
- Comparison describes the item by comparing it to other items
- Summarization describes common characteristics of a subset
- Discrimination describes minimum features needed to differentiate among classes
- Association finds common occurrence of observations

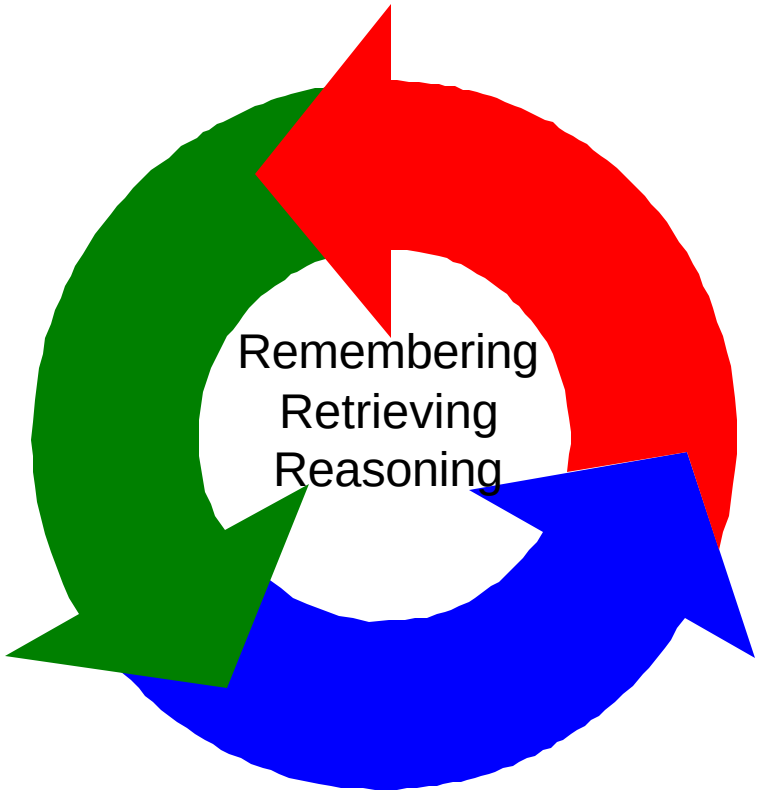
Paralysis



- Source
  - too slow to search the problem space
  - not enough data/processing time available for a system to generate a NP model
  - lack of domain knowledge
  - too much data (including noise) from HTP (high dimensionality)
- A solution
  - HTP & computation
  - Generate - analyze - reduce - test - validate

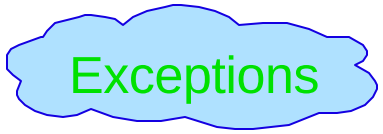
HTP

- Modified CBR approach
  - symbolic similarity
  - lazy learning combined with
    - clustering & classification
    - summarization
- Analysis-based research
  - DNA microarray analysis
  - annotation



Model-Building Solutions

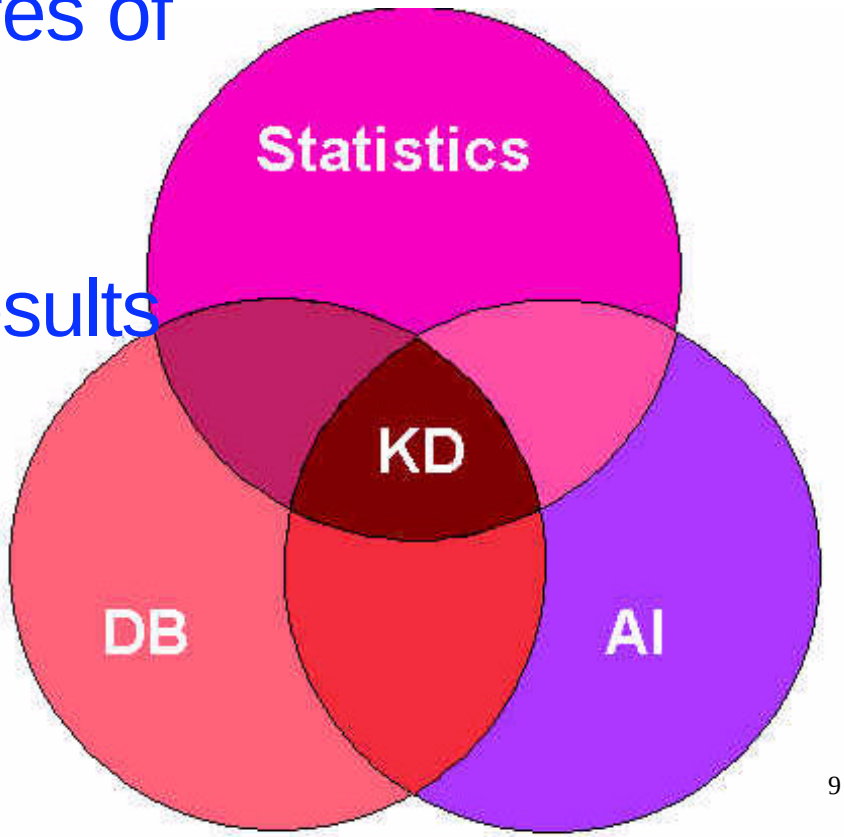
- Eager approach
  1. analyze data
  2. create a model
  3. use the model
- Lazy approach - data-driven model
  1. incrementally accumulate data
  2. incrementally analyze & evolve



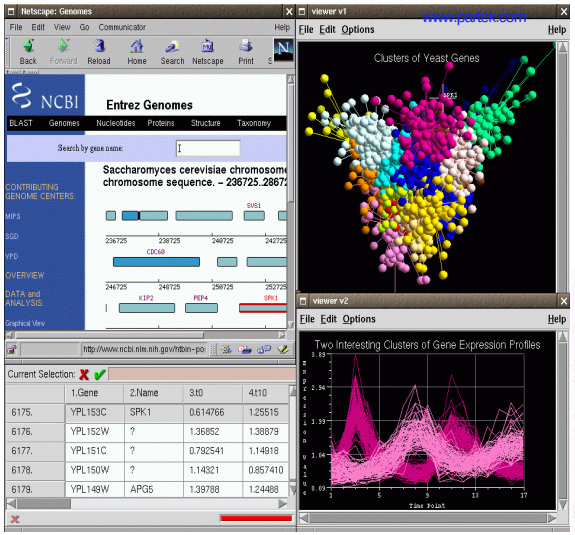
- Generate - analyze - reduce - test - validate

Analyzing and Using MA Data

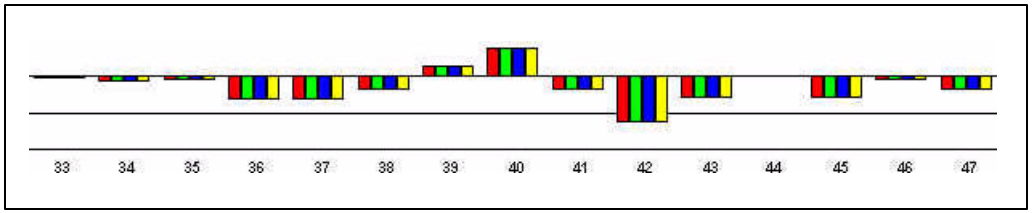
- Problems
  - Knowledge of classes
  - Providing parameters
  - Clinical attributes as measures of "meaningfulness"
  - Scalability
  - Annotating and explaining results
  - Quality assurance
  - Integratability

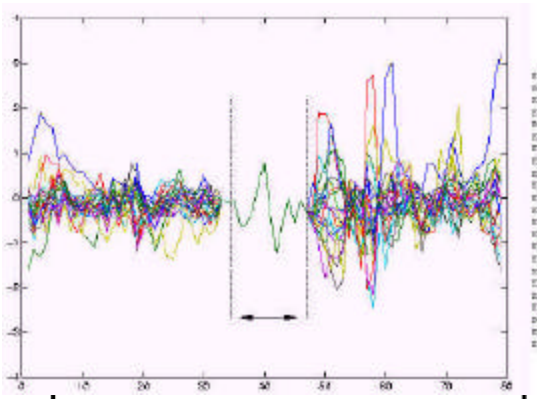


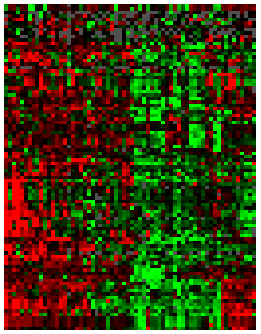
Discovery Algorithms

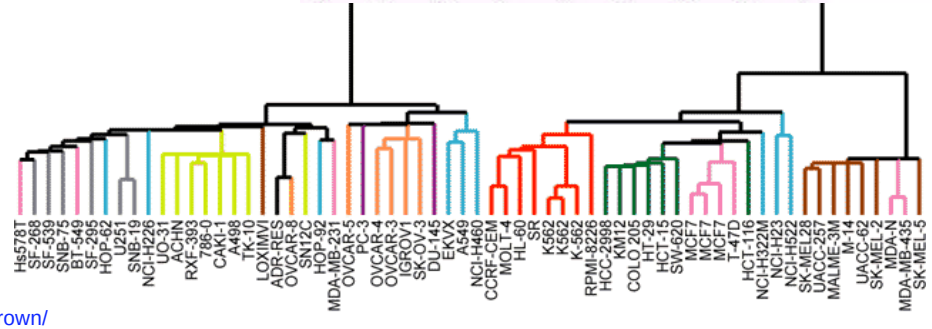


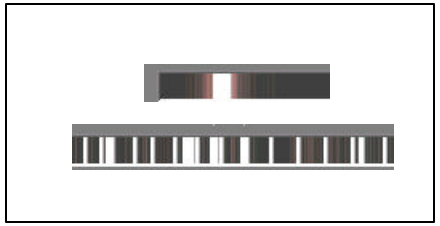
www.partek.com



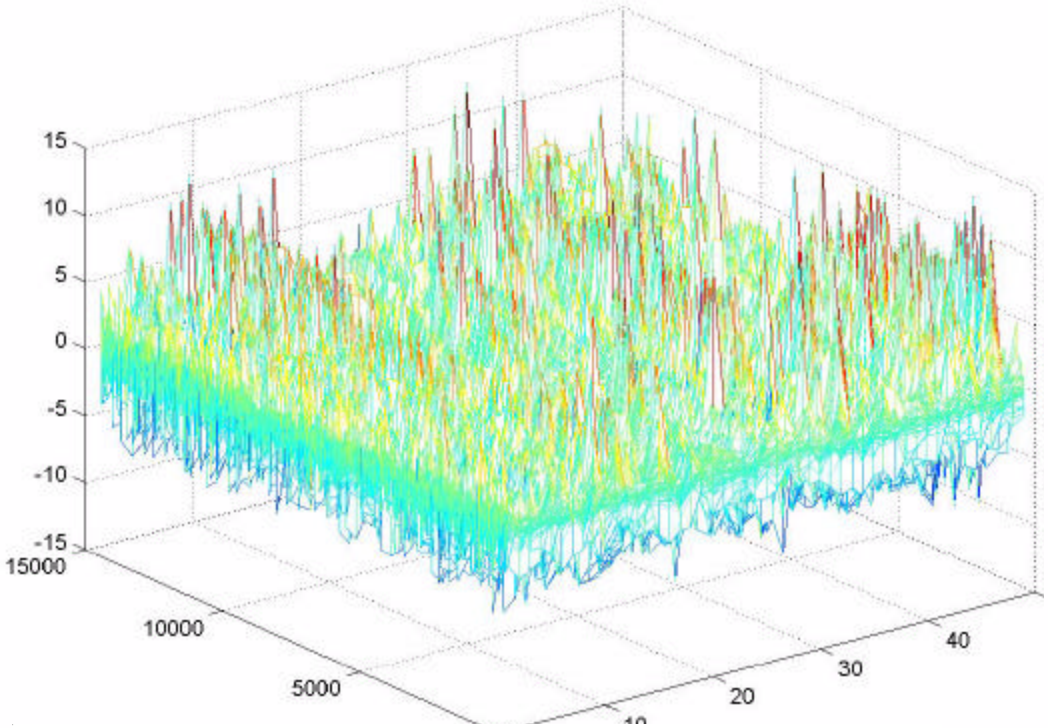


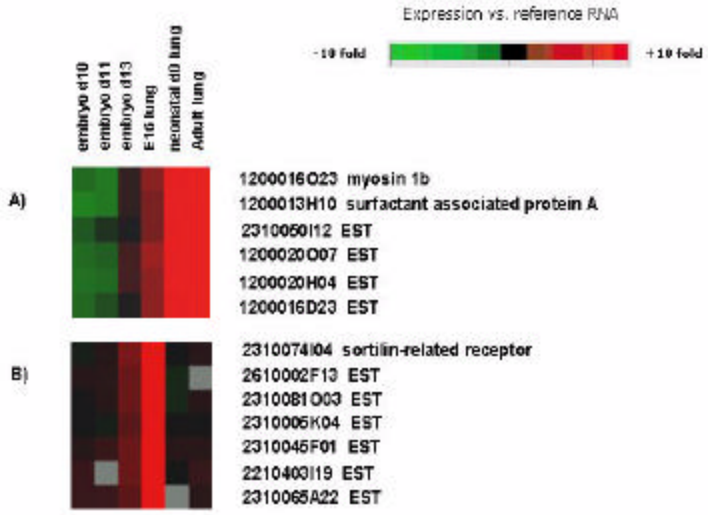


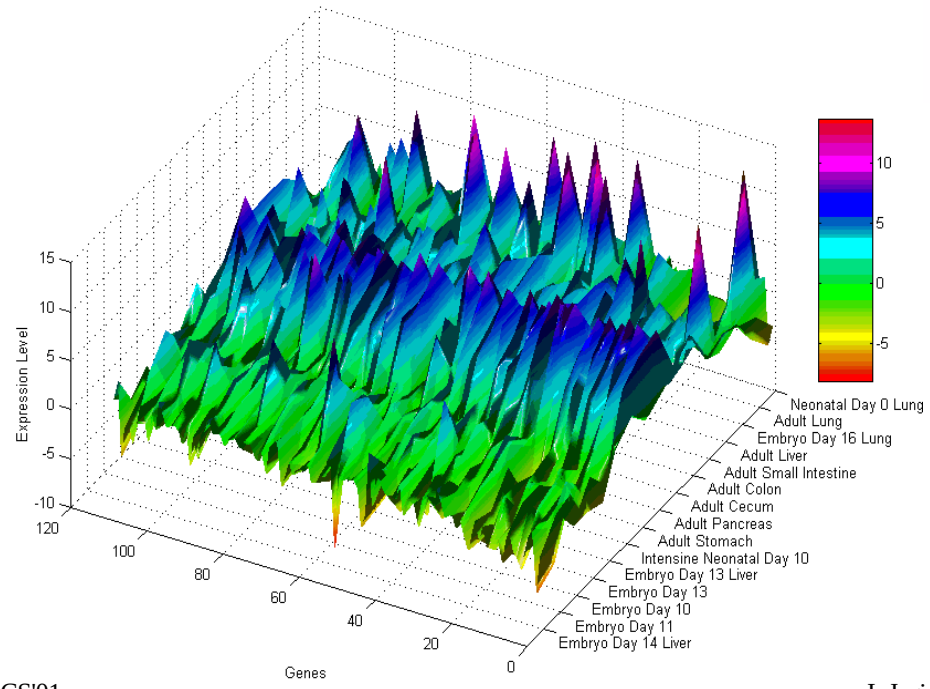


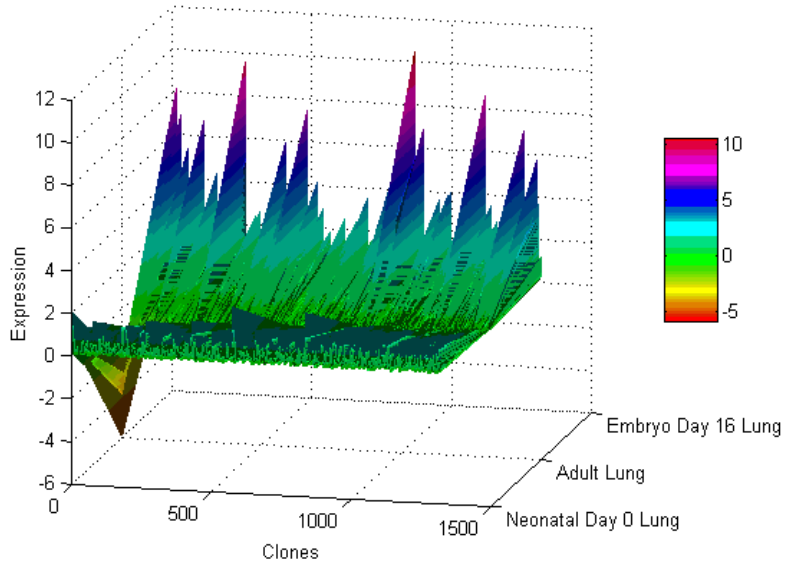


http://cmgm.stanford.edu/pbrown/

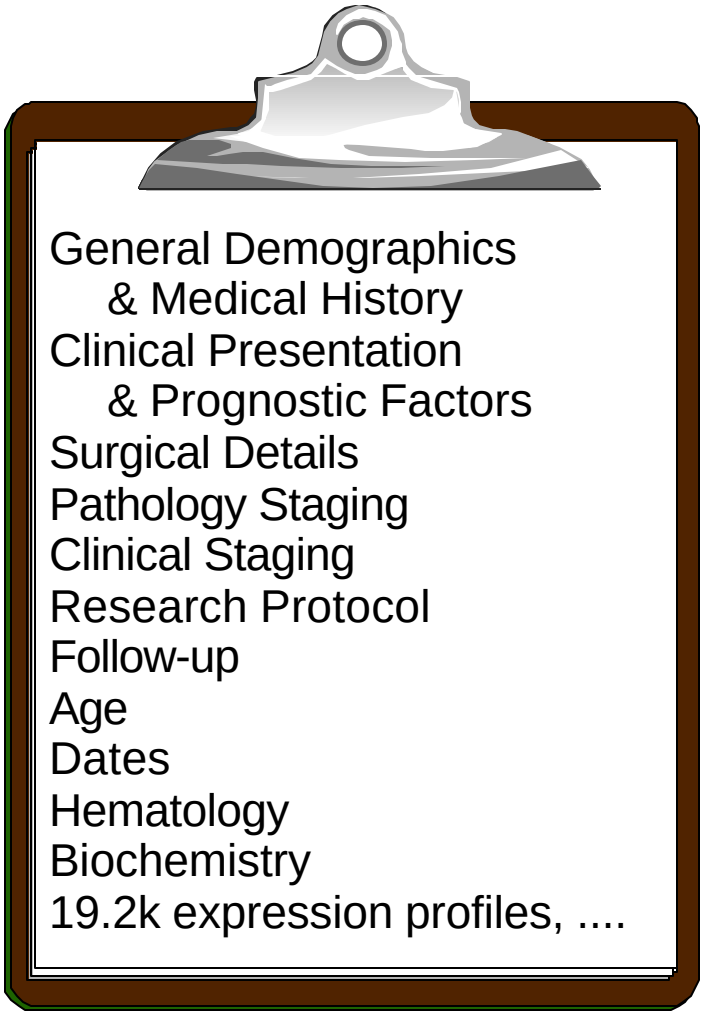

15
10
5
0
-5
-10
-15
15000
10000
5000
40
30
20

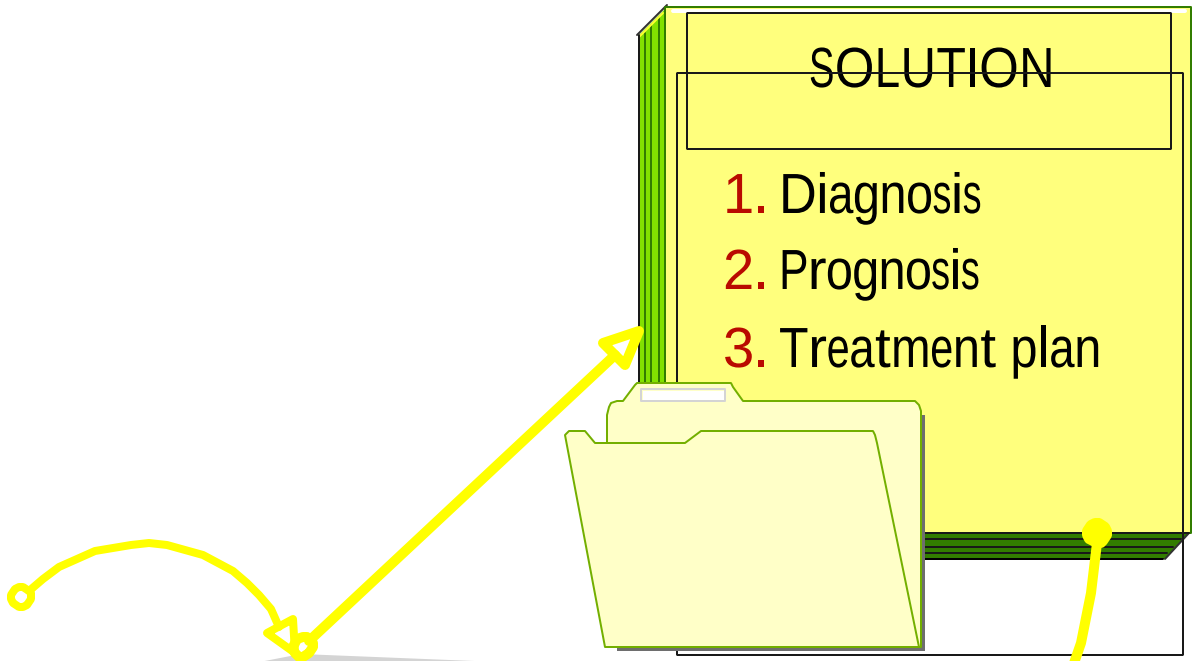

Expression vs. reference RNA
-10 fold
+10 fold
embryo d10
embryo d11
embryo d13
E16 lung
neonatal d0 lung
Adult lung
A)
1200016O23 myosin 1b
1200013H10 surfactant associated protein A
2310050I12 EST
1200020O07 EST
1200020H04 EST
1200016D23 EST
B)
2310074I04 sortilin-related receptor
2610002F13 EST
2310081O03 EST
2310005K04 EST
2310045F01 EST
2210403I19 EST
2310065A22 EST

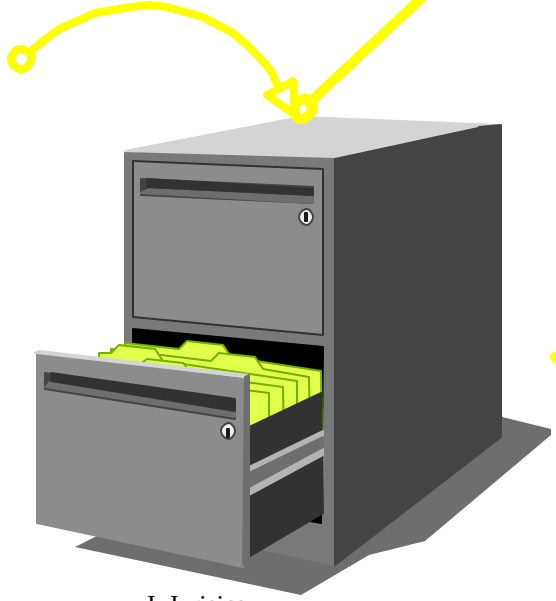

Expression Level
15
10
5
0
-5
-10
120
100
80
60
40
20
0
Genes
10
5
0
-5
Neonatal Day 0 Lung
Adult Lung
Embryo Day 16 Lung
Adult Liver
Adult Small Intestine
Adult Colon
Adult Cecum
Adult Pancreas
Adult Stomach
Intensine Neonatal Day 10
Embryo Day 13 Liver
Embryo Day 13
Embryo Day 10
Embryo Day 11
Embryo Day 14 Liver


Expression
12
10
8
6
4
2
0
-2
-4
-6
0
500
1000
1500
Clones
10
5
0
-5
Embryo Day 16 Lung
Adult Lung
Neonatal Day 0 Lung

Case-Based Reasoning



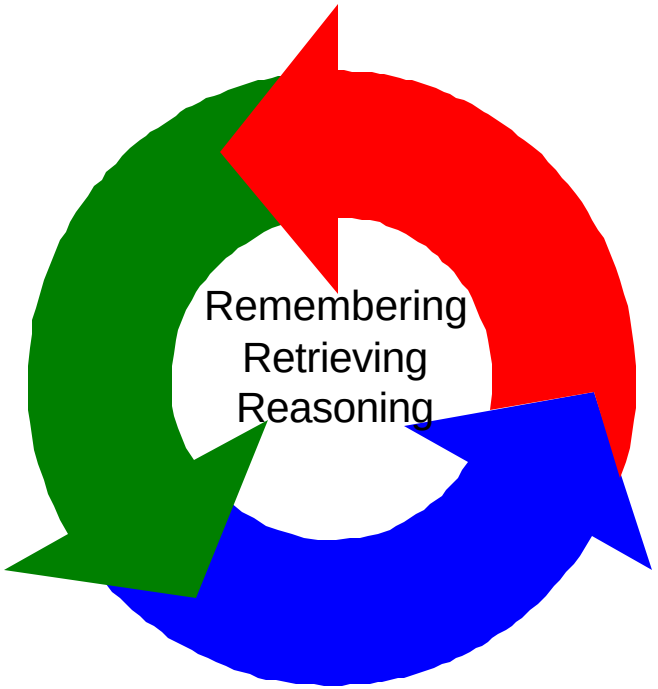




Store
Reason
Analyze

Case-Based Reasoning

- DSS
  - Cases represent experiential knowledge
  - Cases are patterns: context, problem, solution
  - Symbolic similarity - context-based
  - Retrieval - k-NN with context and structure
  - Anytime algorithm
- KM for evolving domains
  - Documenting, analyzing, transferring & sharing experience
  - Classification, prediction, guidance in hypothesis discovery
  - Clustering, summarization
  - Acquire now, process later

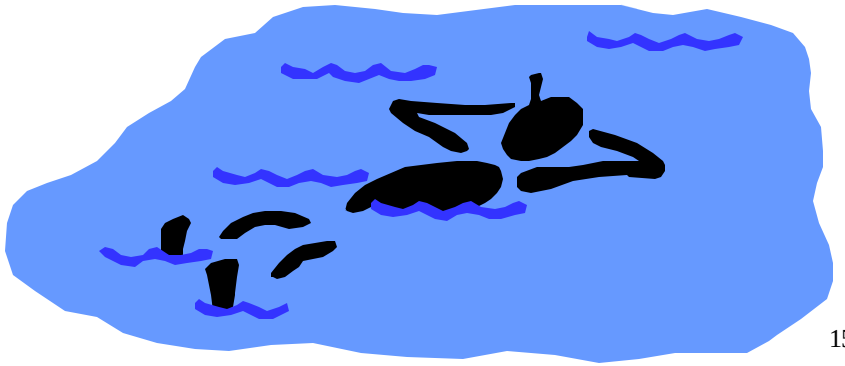


Patient Information Management

- we need detailed disease classification
- we need markers to improve diagnosis, prognosis and treatment planing
- we need new and systematic methods

CBR for DNA Micro Arrays

- Gene expression signature
- Find patients with similar signature
  - ***k***-NN approach - without prior domain knowledge
- Provide diagnosis, prognosis & treatment by analogy
- Apply ***Explain*** function for marker & cancer subtype summarization



Advantage of CBR

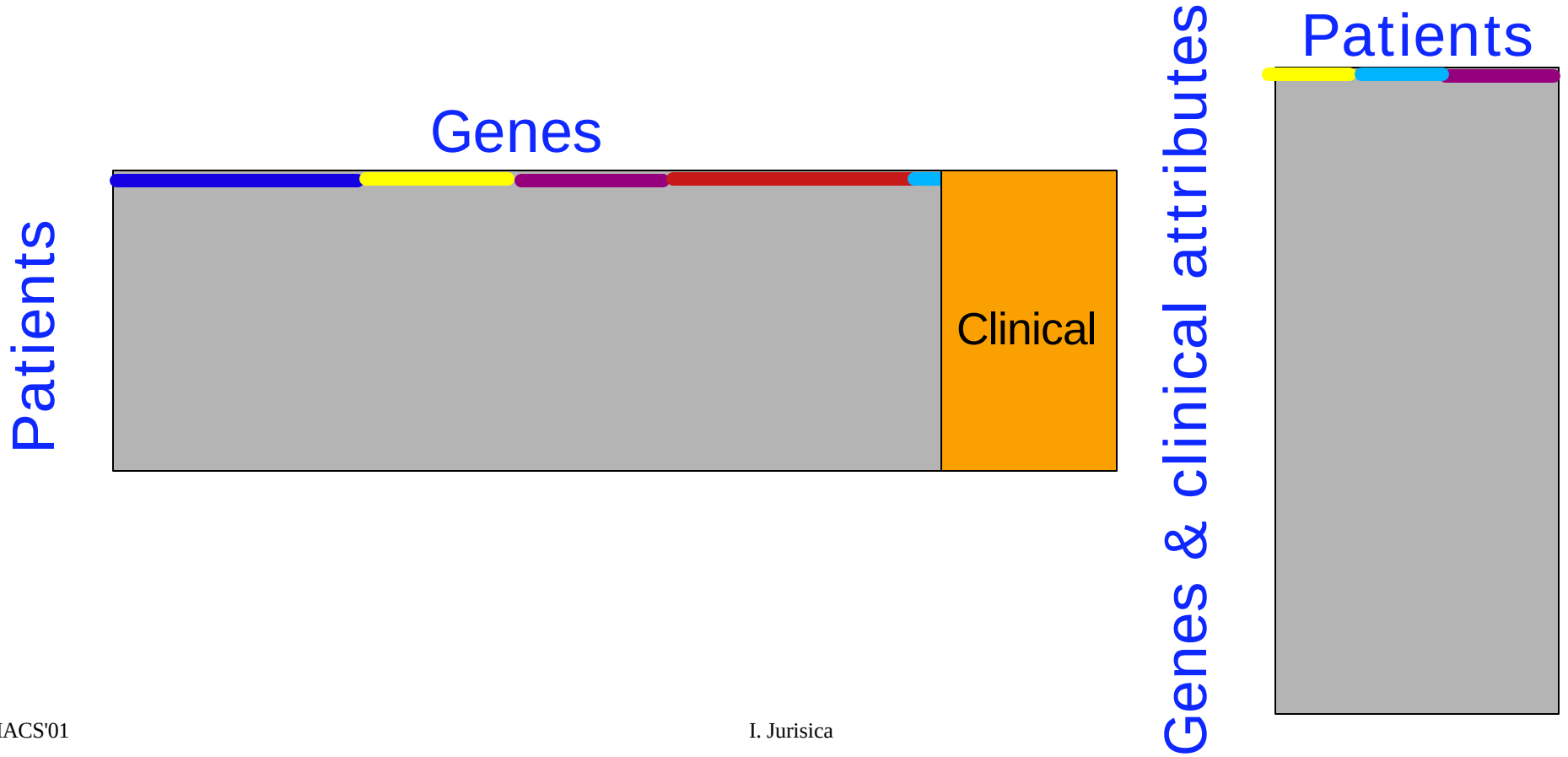
- Supports reasoning, not just analysis
- Measure of similarity is based on gene expression profile
- Does not require prior knowledge
- Supports evolution & is more flexible
- Handles inconsistencies
  - Inconsistencies get resolved at run-time with contextual information
  - CBR can be used to find inconsistencies
- Supports discovery & validation

Outliers

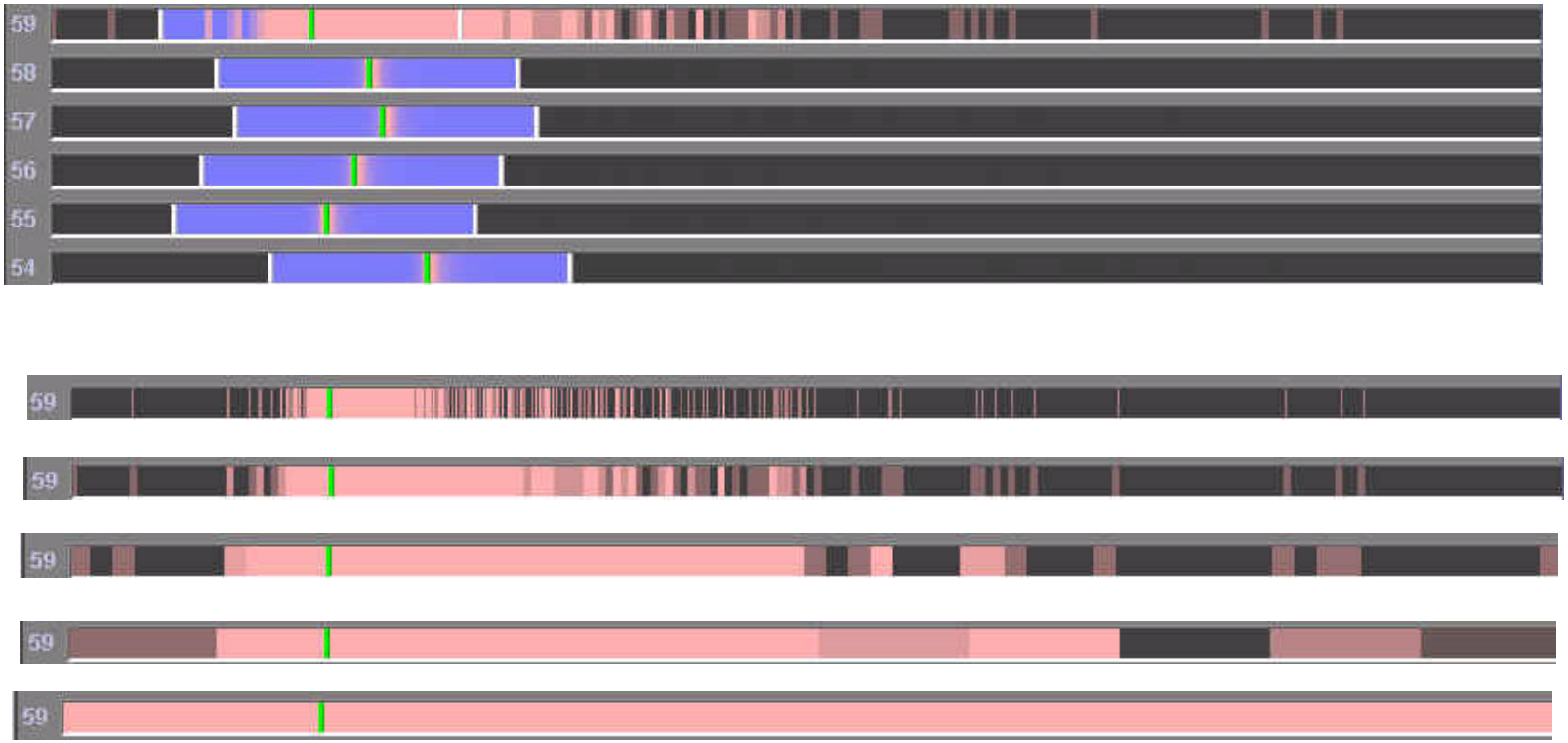
- Represent change and deviation
  - data outside of normal region of input
    - unusual but correct
    - unusual & incorrect
  - for numeric attributes
    - detect with histogram
      - remove with threshold filter
    - identify by calculating the mean & stdev
      - remove by specifying "window", e.g., 2 standard deviations from the mean

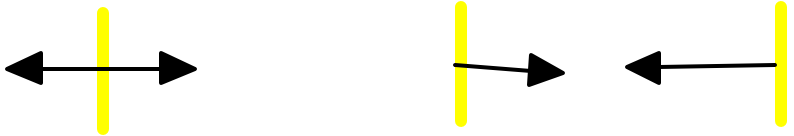
KD and CBR

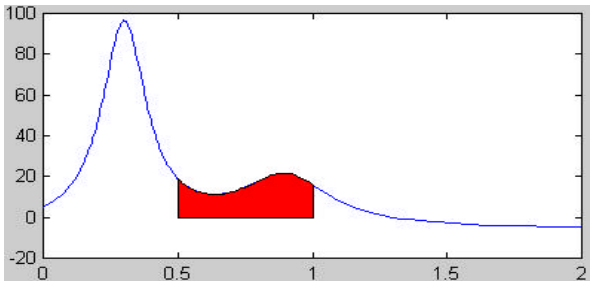
- Organize genes into groups
- Organize attribute values into taxonomies



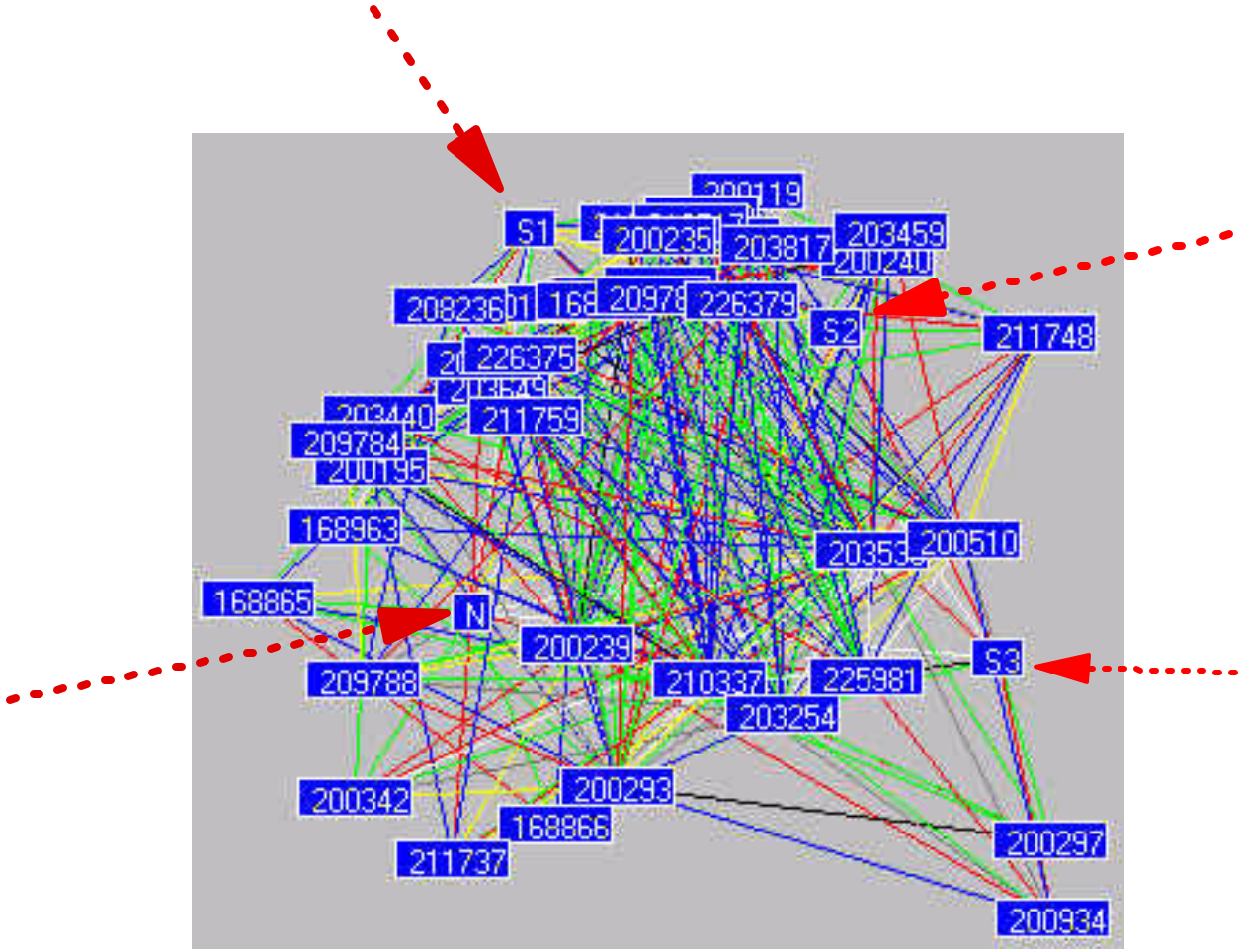
Context Relaxation

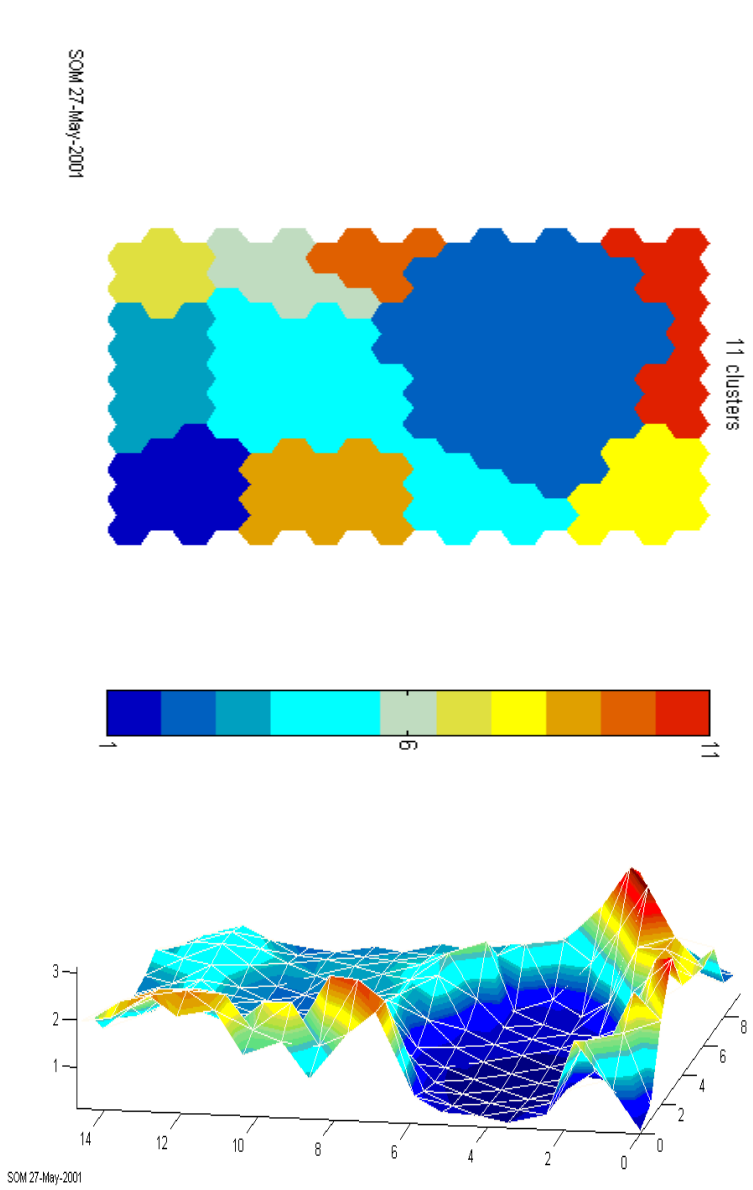


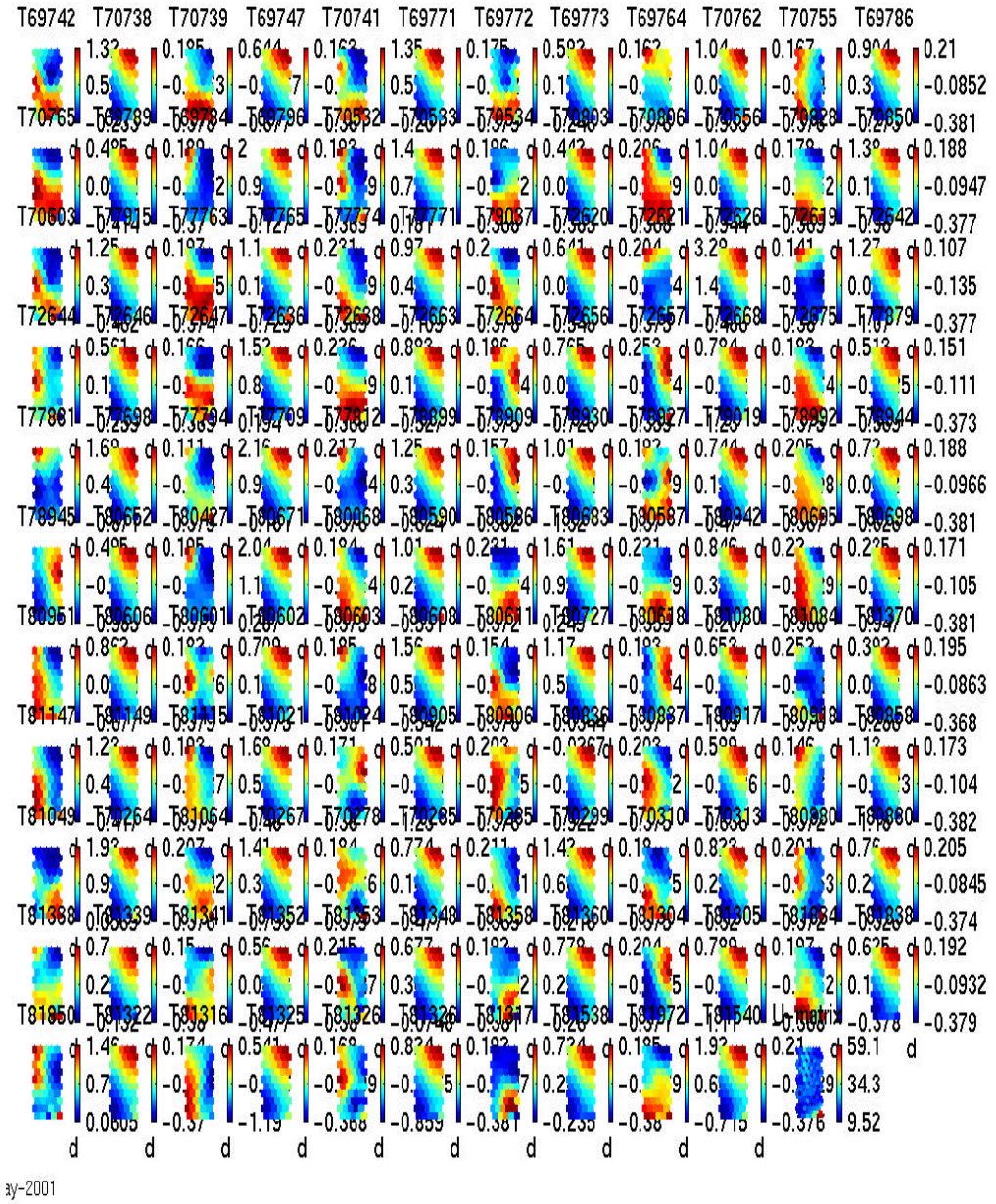


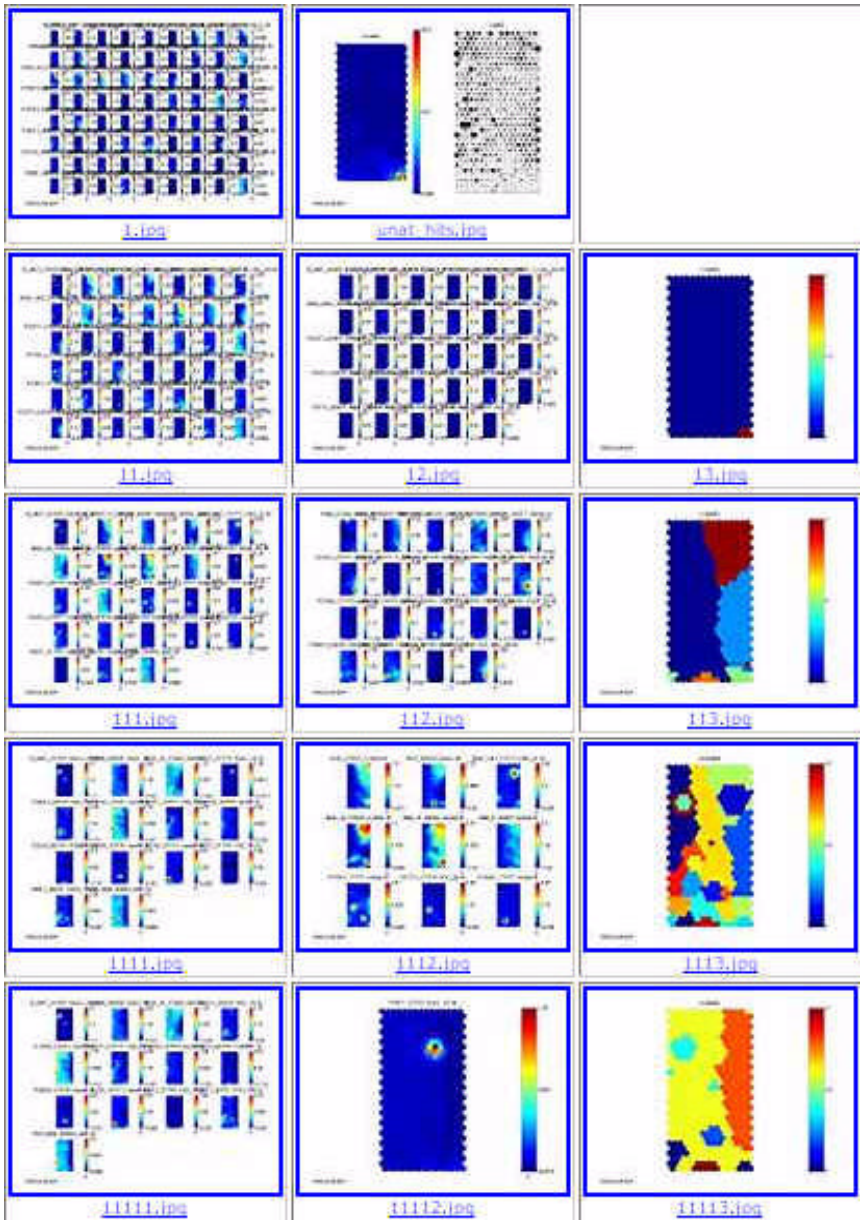


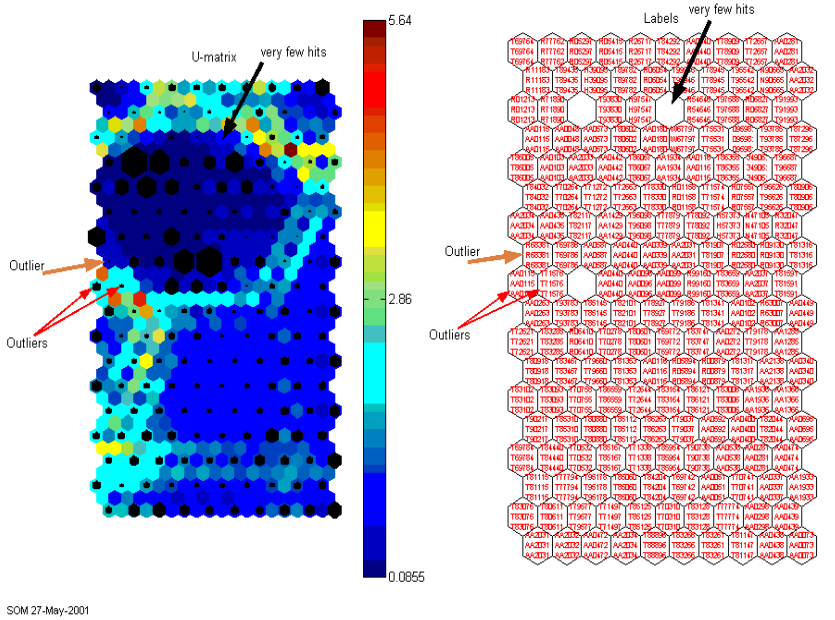
Patient-Patient Similarity

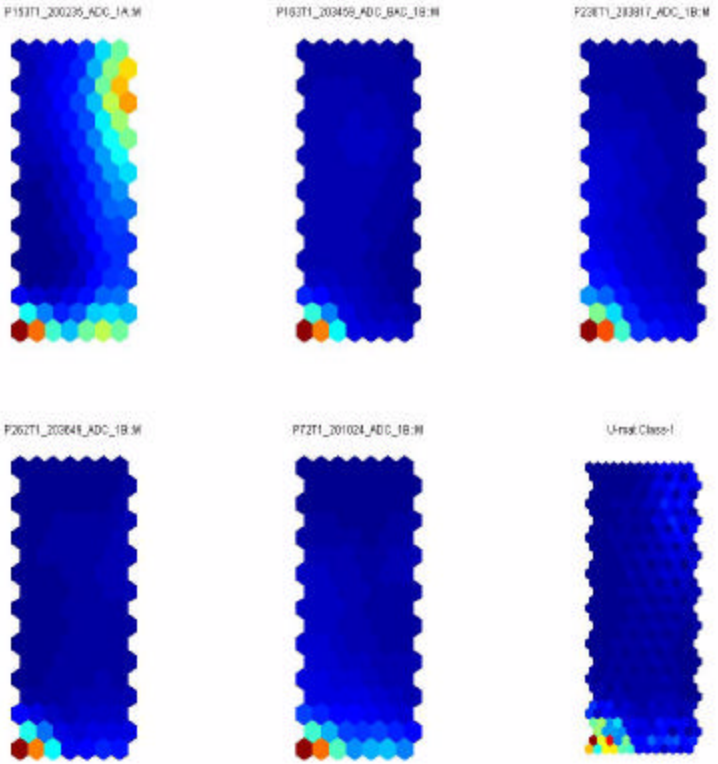



SOM 27-May-2001
11 clusters
1
6
11
3
2
1
14
12
10
8
6
4
2
0
0
2
4
6
8
SOM 27-May-2001


T69742
T70738
T70739
T69747
T70741
T69771
T69772
T69773
T69764
T70762
T70755
T69786
U-matrix

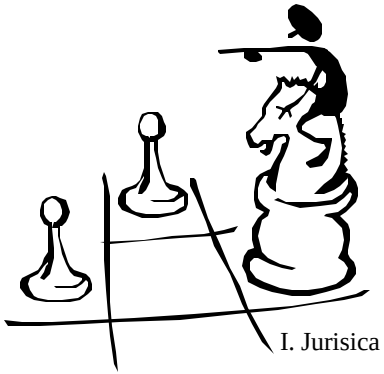

1.jpg
unat_hits.jpg
11.jpg
12.jpg
13.jpg
111.jpg
112.jpg
113.jpg
1111.jpg
1112.jpg
1113.jpg
11111.jpg
11112.jpg
11113.jpg

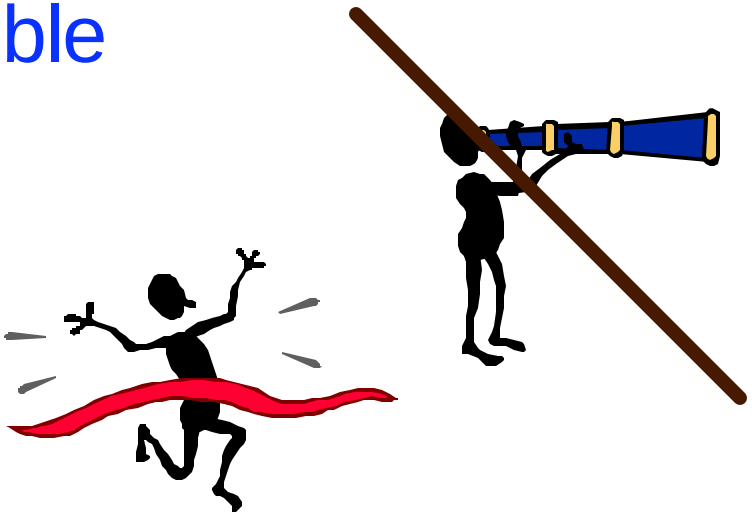

U-matrix
very few hits
Outlier
Outliers
5.64
2.86
0.0855
Labels
very few hits
Outlier
Outliers
SOM 27-May-2001


P153T1_200235_ADC_1A:M
P163T1_203458_ADC_BAC_1B:M
P228T1_203817_ADC_1B:M
P262T1_203649_ADC_1B:M
P72T1_201024_ADC_1B:M
U-mat Class-1

Open Source BIOdb

- Automated annotation
- Schema integration, info validation
- Querying and analysis
- Reasons for local source:
  - certain tasks are more efficient and effective
  - certain tasks become possible





WebOQL

http://www.cs.toronto.edu/~weboql

- A system for supporting data restructuring operations
  - to integrate data from different sources (documents, relational tables, hypertexts)
  - to restructure an instance of a given source into an instance of another one
- We used WebOQL to write wrappers for UniGene
  - more generic, dynamic, incremental

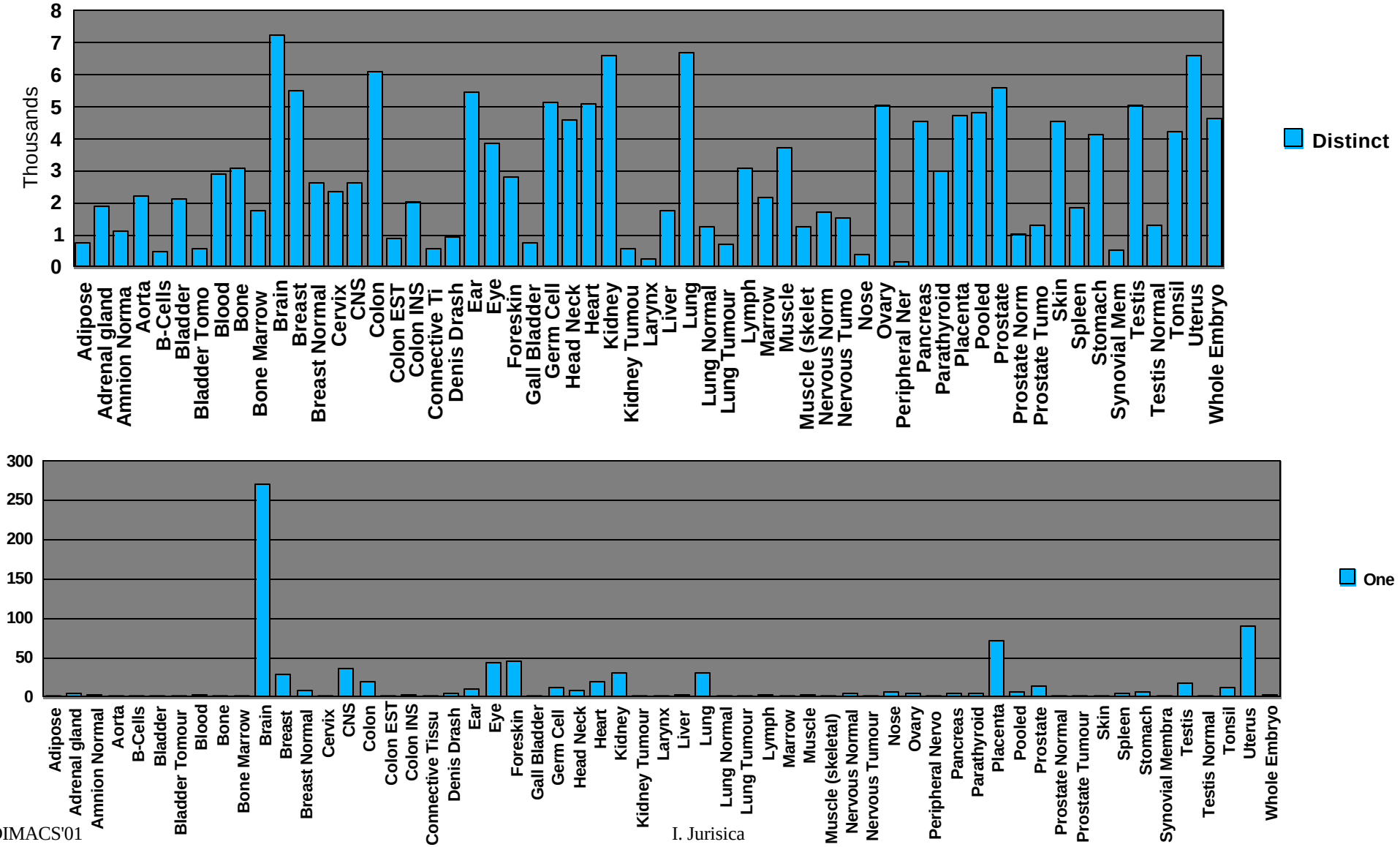
Autoannotations

- Information may not be downloadable
- Information may not be complete

ID=1
TITLE=Hippocampus,_Stratagene_(cat.__936205)
TISSUE=brain, hippocampus
VECTOR=lambdaZAP-II

Lib.1
Infant, 2 yrs, female
brain, hippocampus
lambdaZAP-II
453 ESTs have been classified, 411 gene sets

Expression Distribution



Lung

Lung	15,410
Lung-tumor	67
Lung-tumor & suppressor	26
Lung-tumor & necrosis	20
Lung-tumor & antigen	5
Lung-tumor & susceptibility	3

Hs.241493	M. musculus	PIR:B47328	B47328 natural killer cell tumor-recognition protein - mouse"	1511	79 %
Hs.241493	H. sapiens	SP:P30414	NKCR_HUMAN NK-TUMOR RECOGNITION PROTEIN"	1461	100 %
Hs.19074	H. sapiens	PID:g7212790	large tumor suppressor 2"	1045	100 %
Hs.48499	H. sapiens	PID:g7144644	AF102177 1 tumor antigen SLP-8p"	965	100 %
Hs.116875	M. musculus	PID:g7637845	AF172722 1 tumor-rejection antigen SART3"	962	87 %
Hs.211600	M. musculus	SP:Q60769	TNP3 MOUSE TUMOR NECROSIS FACTOR, ALPHA-INDUCED PROTEIN 3"	789	88 %
Hs.211600	H. sapiens	SP:P21580	TNP3_HUMAN TUMOR NECROSIS FACTOR, ALPHA-INDUCED PROTEIN 3"	789	100 %

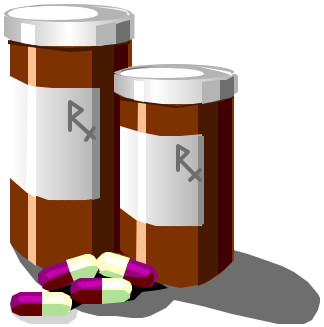
Conclusions

- Management - representation - reasoning - discovery
  - moving from hypothesis-driven to exploration-driven research (analysis)
  - systematically analyzing the problem space
- HTP
  - automation, systematicity, reproducibility
  - hypothesis search - generation & evaluation

The Future

- "Most disease processes and treatments are manifested at the protein level"
- "Gene-based expression analysis alone will (in certain cases) be totally inadequate for drug discovery"
- "Only 2% of diseases are believed to be monogenic - we need to understand protein-protein interactions"

DDT 4(3):129-133, 1999



Thanks

- P. Rogers, M. Sultan
- A. Rehaag, G. Quon
- D. Wigle, O. Huner
- P. Macgregor, M. Albert

- J. Glasgow

- A. Barta
- M. Maziarz
- W. Andreopoulos

NSERC, CITO,
NIH, IBM, OCI

http://www.cs.utoronto.ca/~juris