Computing at the Center of Genomic Medicine

Isaac S. Kohane

The Challenge of a (Sometimes Forced) Interdisciplinary Pipeline

Finding the Needle in the Haystack: A Case History

- Cerebellum has pivotal roles in the coordination of posture and locomotion
- Laminar organization of the cerebellar cortex has facilitated understanding its basic circuitry, functions and ontogeny

Sonic Hedgehog (Shh), development and tumorigenesis

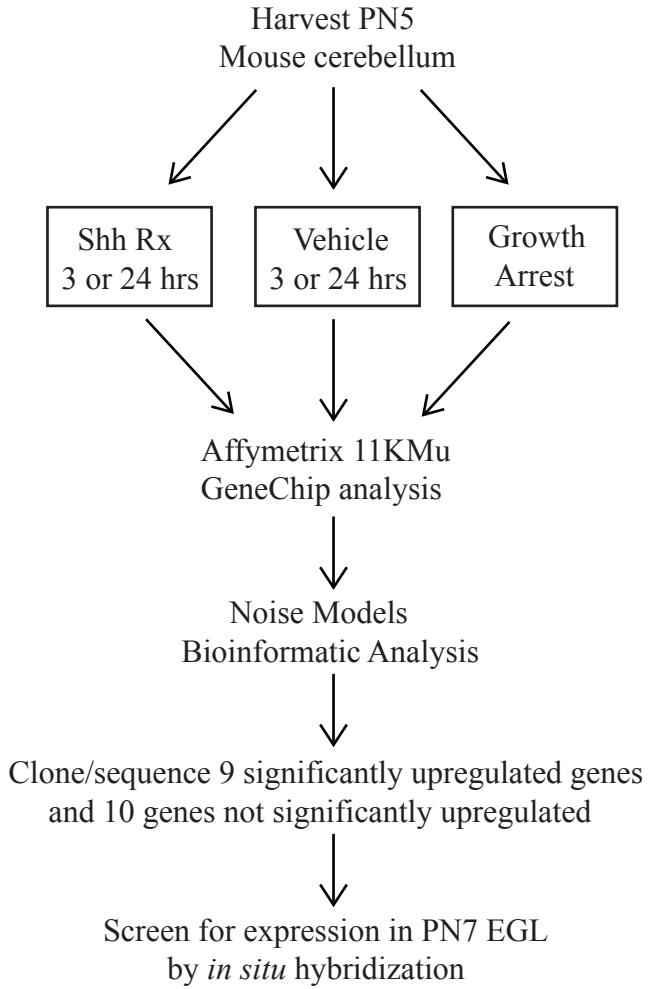
The challenge

- Find novel members (the needles) of the Shh pathway in a large haystack.
- The haystack
  - ✓ Large number of probes
  - ✓ Whole cerebellum
    - ☞ (**whole organ, multiple cell types)**
- But the signal of interest is confined to thin superficial layer

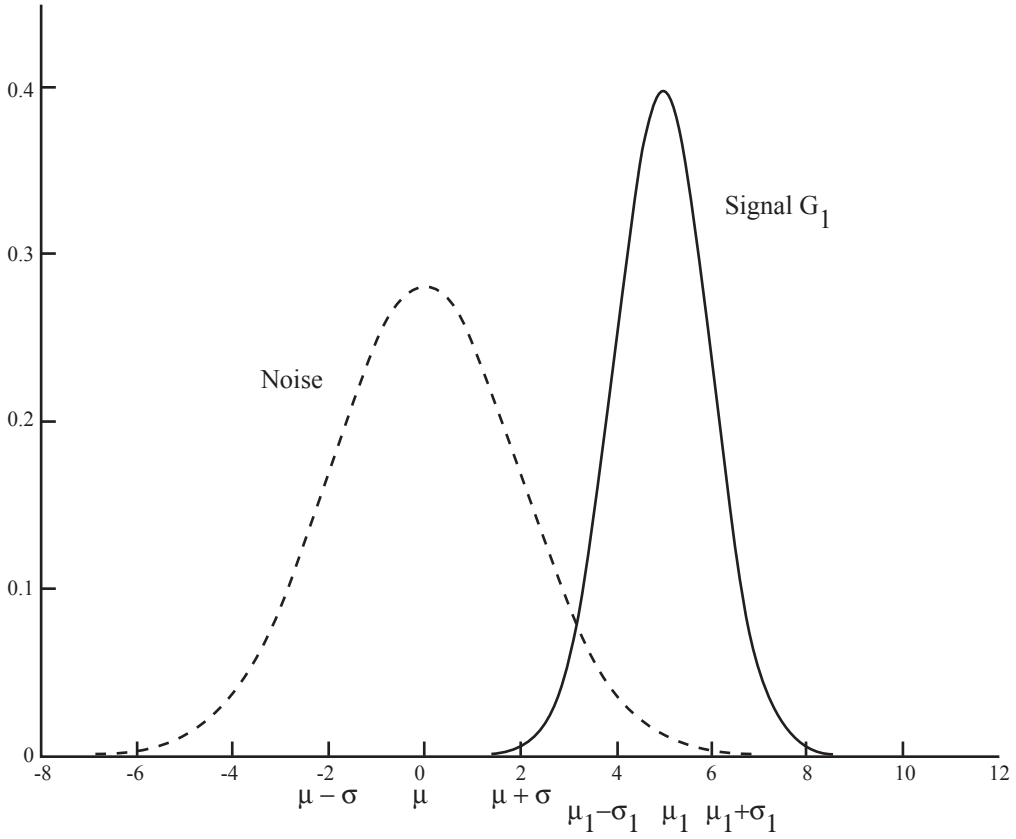
- Can we find the signal in time and space?

Overall schema

Please see Figure 1 of Proc Natl Acad Sci U S A. 2002 Apr 16; 99(8): 5704-9. Identification of genes expressed with temporal-spatial restriction to developing cerebellar neuron precursors by a functional genomic approach. Zhao Q, Kho A, Kenney AM, Yuk Di DI, Kohane I, Rowitch DH.



A noise model with triplicate measurements



Fold noise and signal distributions a gene, G_1

What average are we calculating?

$s_1\ s_2\ s_3$ $v_1\ v_2\ v_3$

$$\frac{\frac{1}{3}\sum_{i=1}^{3} s_i}{\frac{1}{3}\sum_{i=1}^{3} v_i}$$

How do you calculate average interest rate?

- If 4 different interest rates over four years
- $(1+r1)(1+r2)(1+r3)(1+r4)\mathbf{P} \circ Q^4\mathbf{P}$
- **Q=**

$$\sqrt[4]{(1+r1)(1+r2)(1+r3)(1+r4)}$$

Biological validation: In situ hybridization

Please see Figure 3 of Proc Natl Acad Sci U S A. 2002 Apr 16; 99(8): 5704-9.

Identification of genes expressed with temporal-spatial restriction to developing cerebellar neuron precursors by a functional genomic approach.

Zhao Q, Kho A, Kenney AM, Yuk Di DI, Kohane I, Rowitch DH.

1/2 FP of ratio of means

Same sensitivity

Relevance to Human Disease

- How can we leverage this developmental view of the mouse?
- Human medulloblastoma microarray data
  - ✓ Pomeroy et al, Nature 2002
- Find the mouse homologues of the genes up and down regulated in the tumors.
- Principal Component Analysis to find the main sources of variance in the developmental time series

Mouse Cerebellar time course by first two principal compents

Up and Down Regulated Human Genes Mapped onto The Mouse Genes

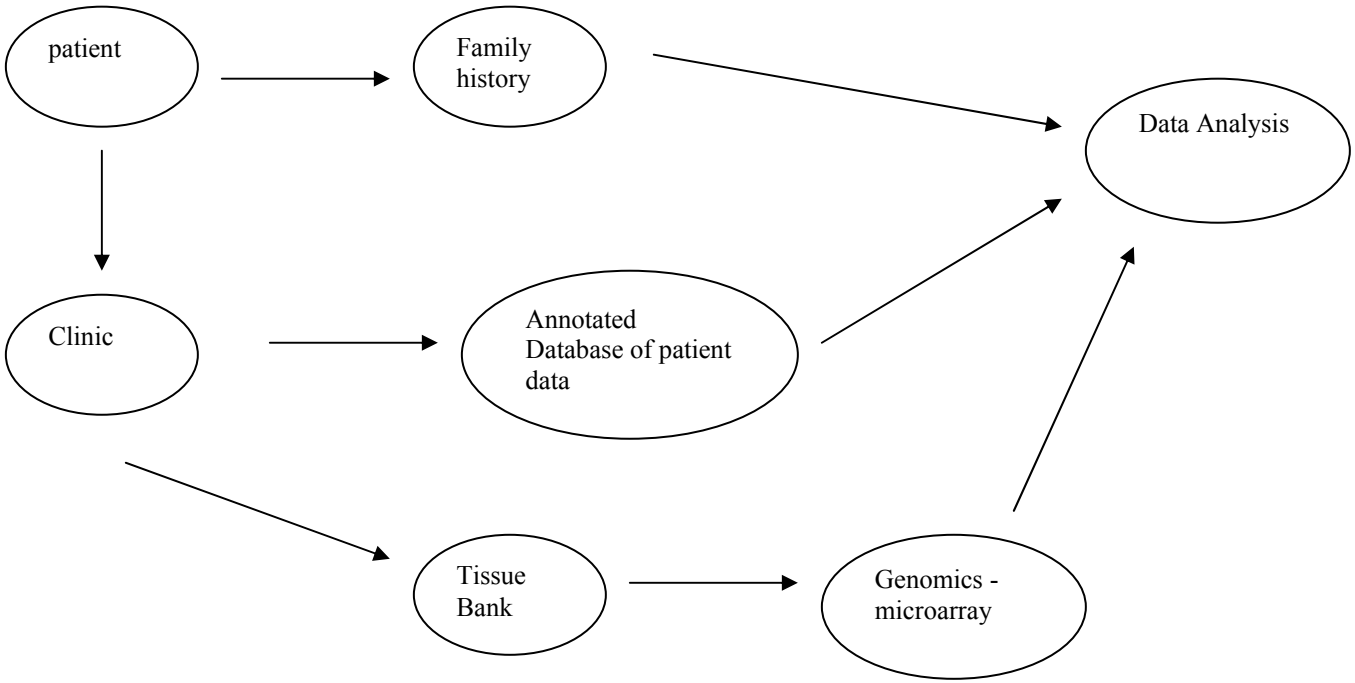
The New Histopatholgy and History Repeated

- Lobstein (1829) and Cohnheim (1887) were amongst the first to theorize similarities between human embryogenesis and the biology of cancer cells
- The brain tumor classification system devised by Bailey and Cushing in 1926, from which modern taxonomies derive, emphasizes histological resemblance to cells of the developing central nervous system

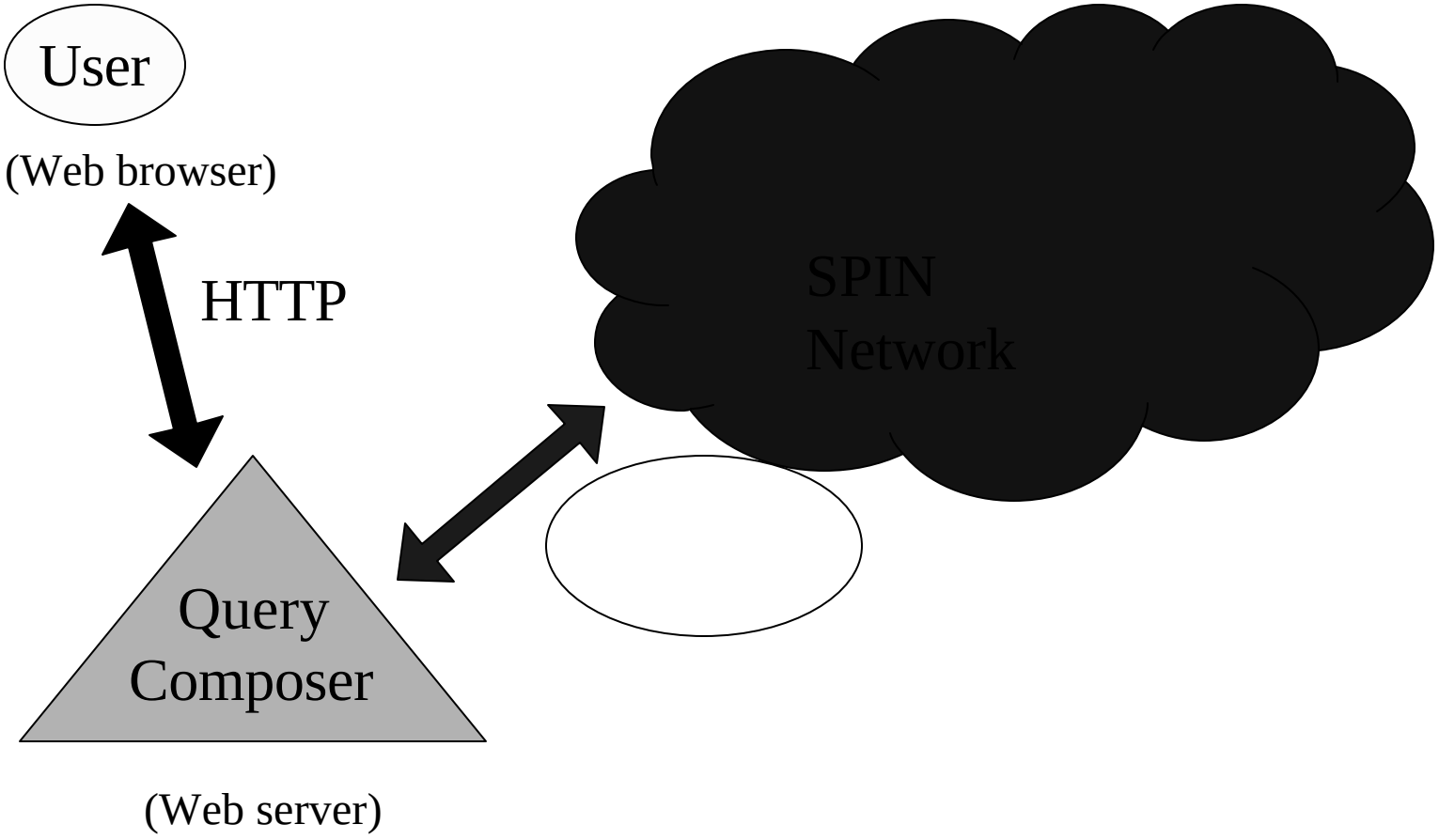
Relation between tumor and day of development

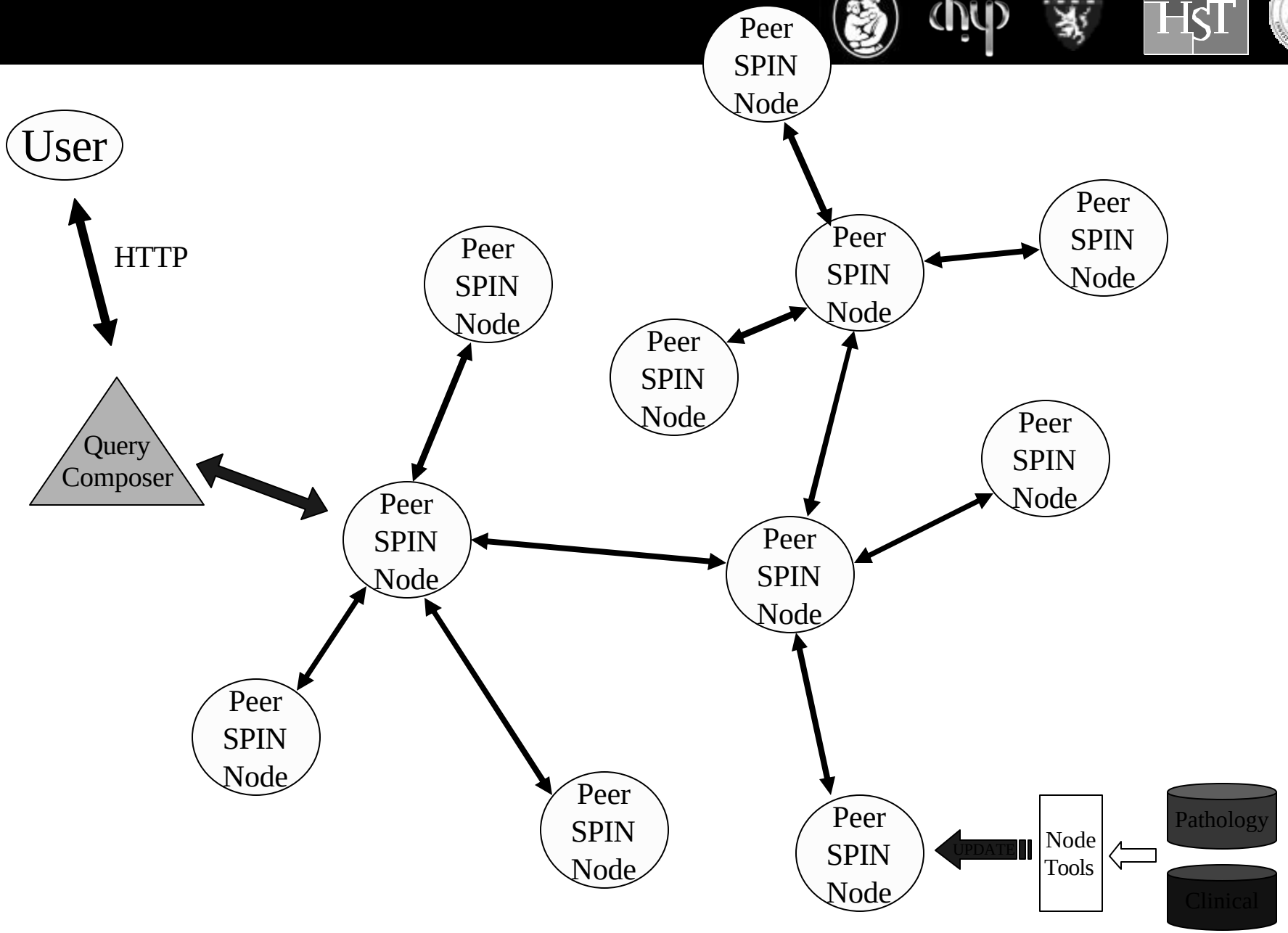
Interim Summary

- Computing to identify pathways
- Computing to provide natural classification of disease
  - ✓ With further insight into mechanism

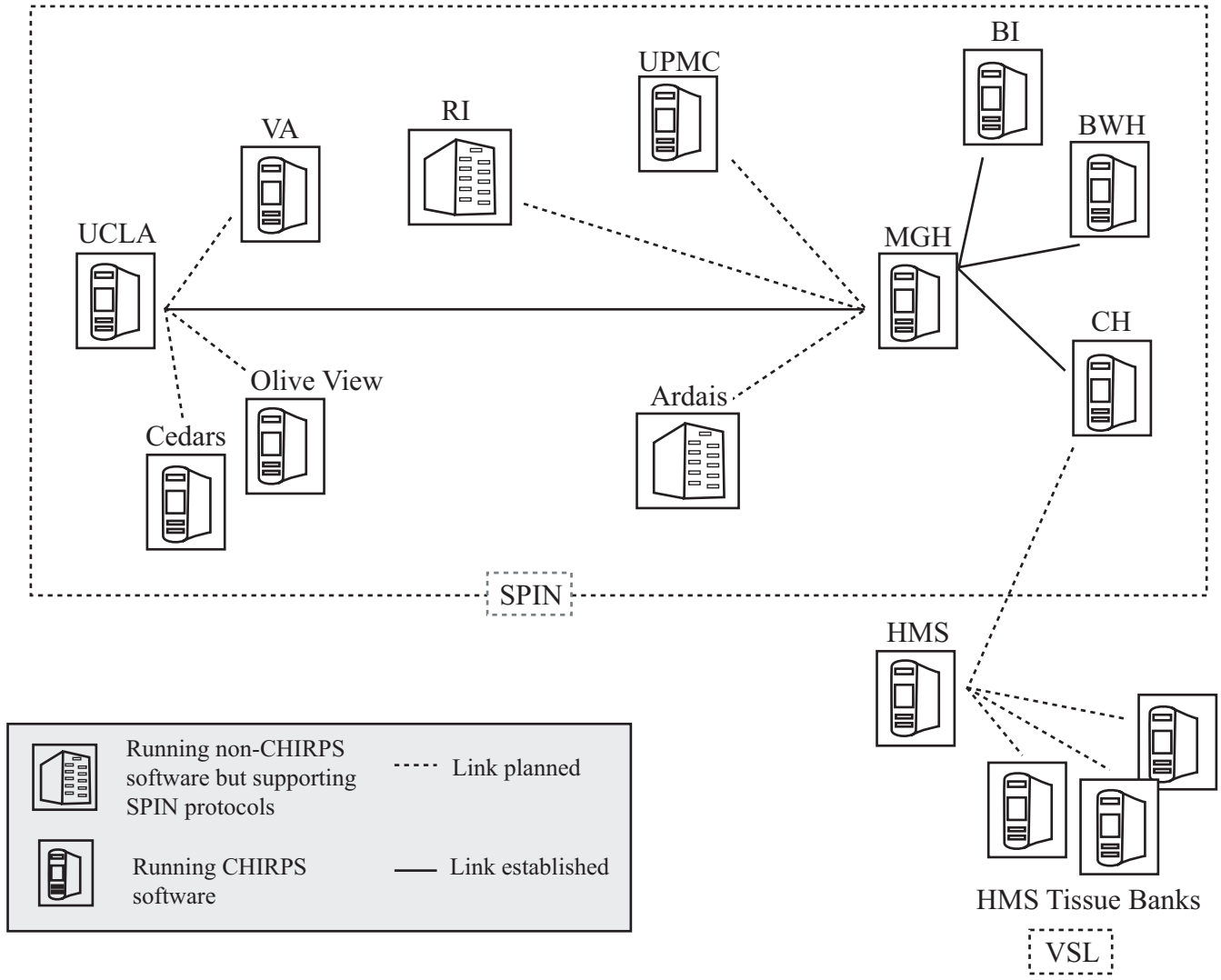

patient
Family history
Data Analysis
Clinic
Annotated Database of patient data
Tissue Bank
Genomics - microarray

Birds-eye view of SPIN




User
HTTP
Query Composer
Peer SPIN Node
Peer SPIN Node
Peer SPIN Node
Peer SPIN Node
Peer SPIN Node
Peer SPIN Node
Peer SPIN Node
Peer SPIN Node
Peer SPIN Node
Peer SPIN Node
Peer SPIN Node
UPDATE
Node Tools
Pathology
Clinical

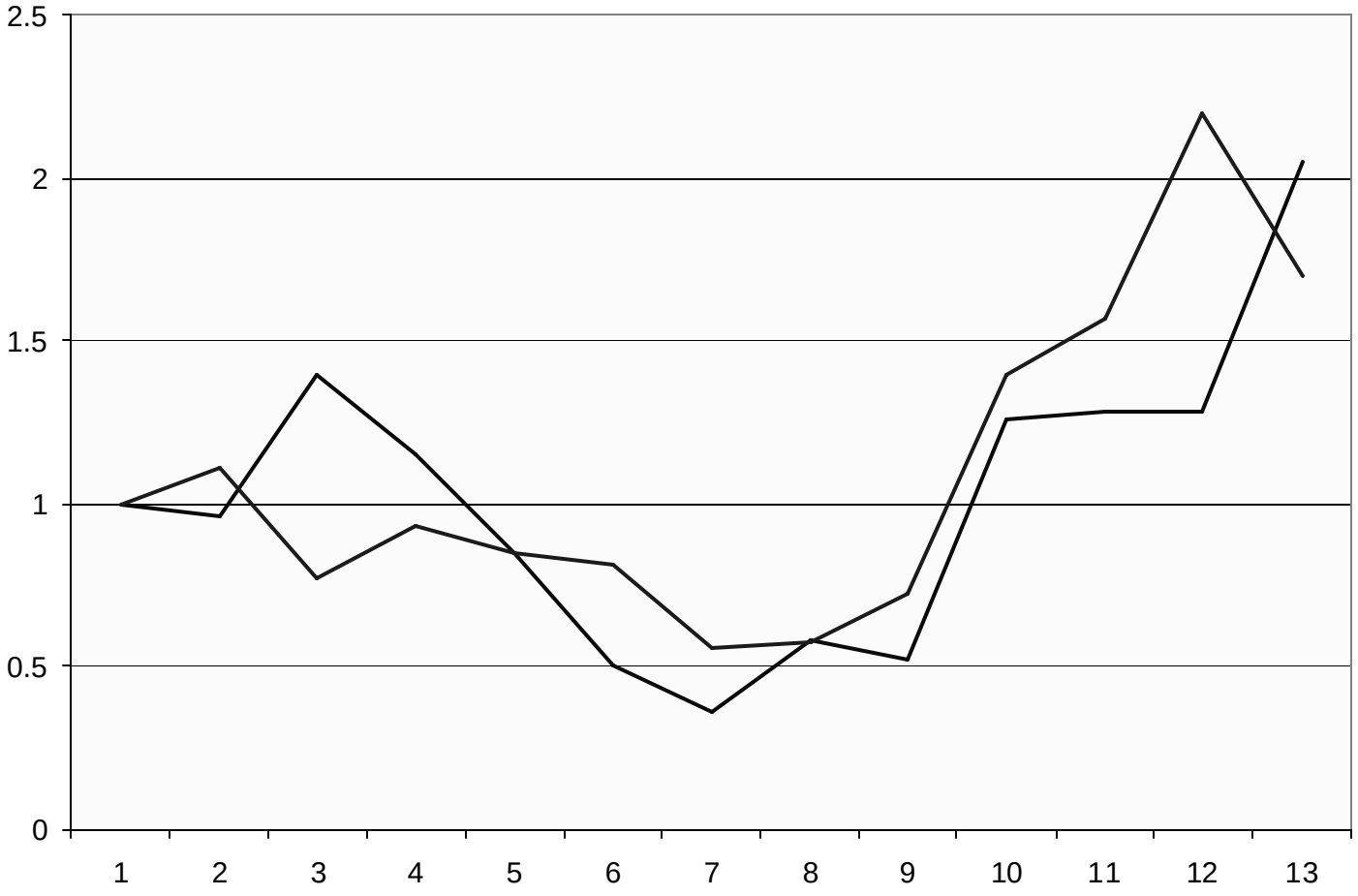
Boston-UCLA: Network view



Please refer to the Shared Pathology Informatics Network homepage at http://www.sharedpath.org/

Dynamics: Another Case History of How Bioinformatics Can Make For Different Biology

Are They Similar?



Two Gene Expression Courses? Are They Similar? Does order Matter?

Please see Curr Opin Mol Ther. 1999 Jun; 1(3): 344-58.

Modified oligonucleotides -- synthesis, properties and applications.

Lyer RP, Rolanf A, Zhou W, Ghosh K.

Autoregressive Models

- Take a time series, of dependent observations:

$$x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \ldots$$

- Approximate with an autoregressive model:

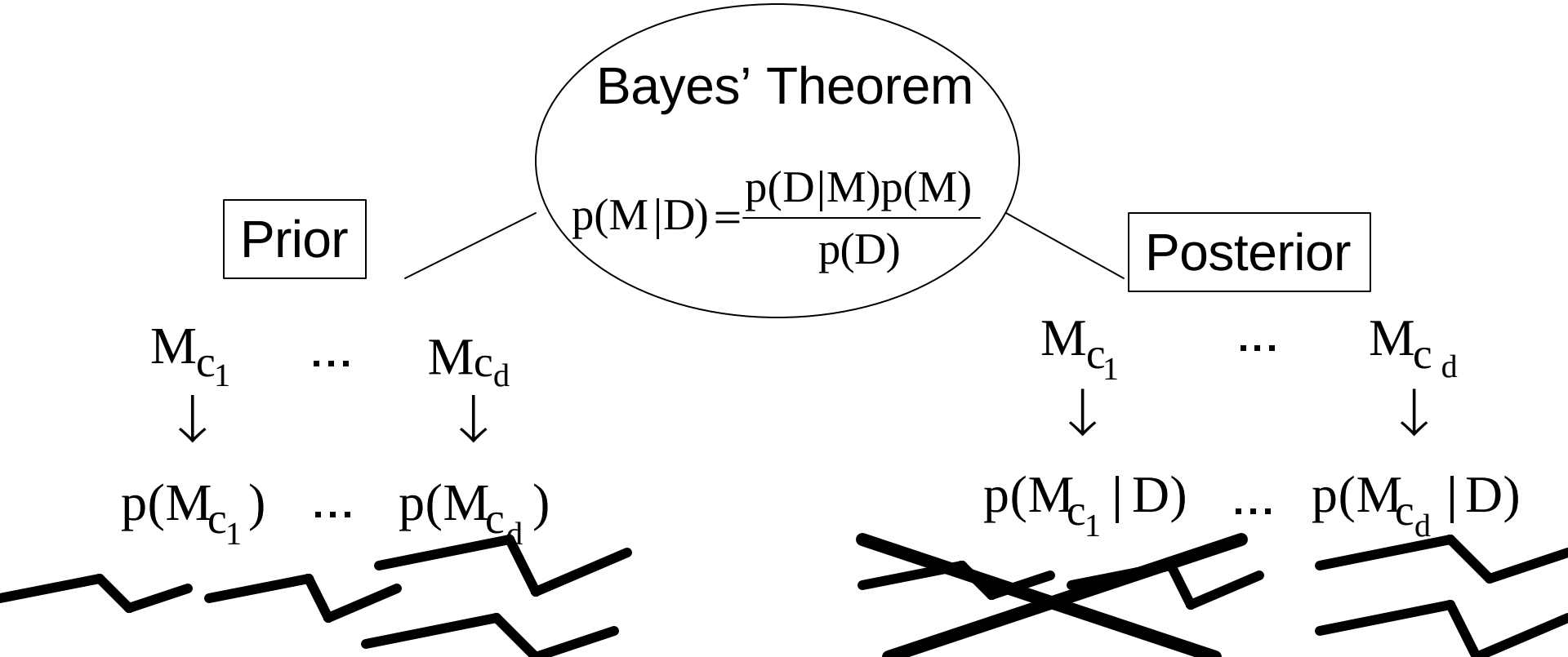
$$P(x_t \mid x_0, x_1, \ldots, x_{t-1}) \quad \text{——} \quad P(x_t \mid x_{t-p}, \ldots, x_{t-1})$$

  the basic assumption is that t_0 is independent of the remote past given the recent past.

- The length of the recent past is the Markov Order *p*.

Bayesian Model Selection

- Clustering is a statistical model selection problem.
- We are looking for the model with maximum posterior probability given the data.



Marginal Likelihood

- We want the most probable model given the data:

$$p(M_i \mid \mathrm{D}) = \frac{p(M_i, \mathrm{D})}{p(\mathrm{D})} = \frac{p(\mathrm{D} \mid M_i)\, p(M_i)}{p(\mathrm{D})}$$

- But we use the same data for all models:

$$p(M_i|D) \propto p(D\,|M_i)p(M_i).$$

- We assume all models are a priori equally likely:

$$p(M_i|D) \propto p(D\,|M_i).$$

- This is the marginal likelihood, which gives the most probable model generating Δ.

Four Clusters

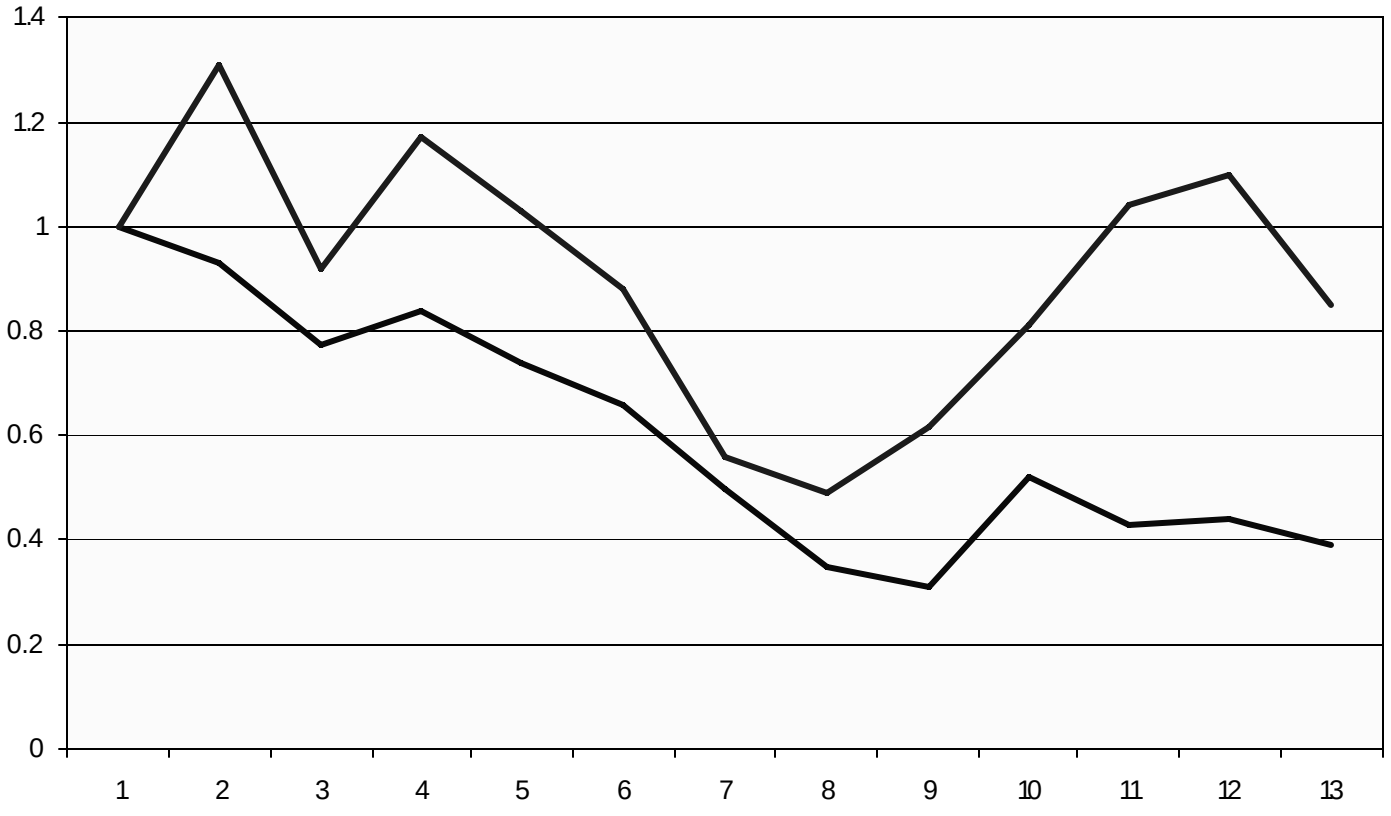
Cluster Members

Cytokine Cluster	*Cluster 2*	*Cluster 3*	*Apoptosis Cluster*
Interleukin 8	D	A	Tyrosine kinase-like orphan receptor 2
Interleukin 6 (interferon beta 2)	E	B	TRAF-binding protein domain
Prostaglandin-endoperoxide synth 2	F	C	Death-associated protein kinase
	G		Transcription termination factor-like protein
	H		DKFZP586G1122 protein
	I		
	J		

Validation

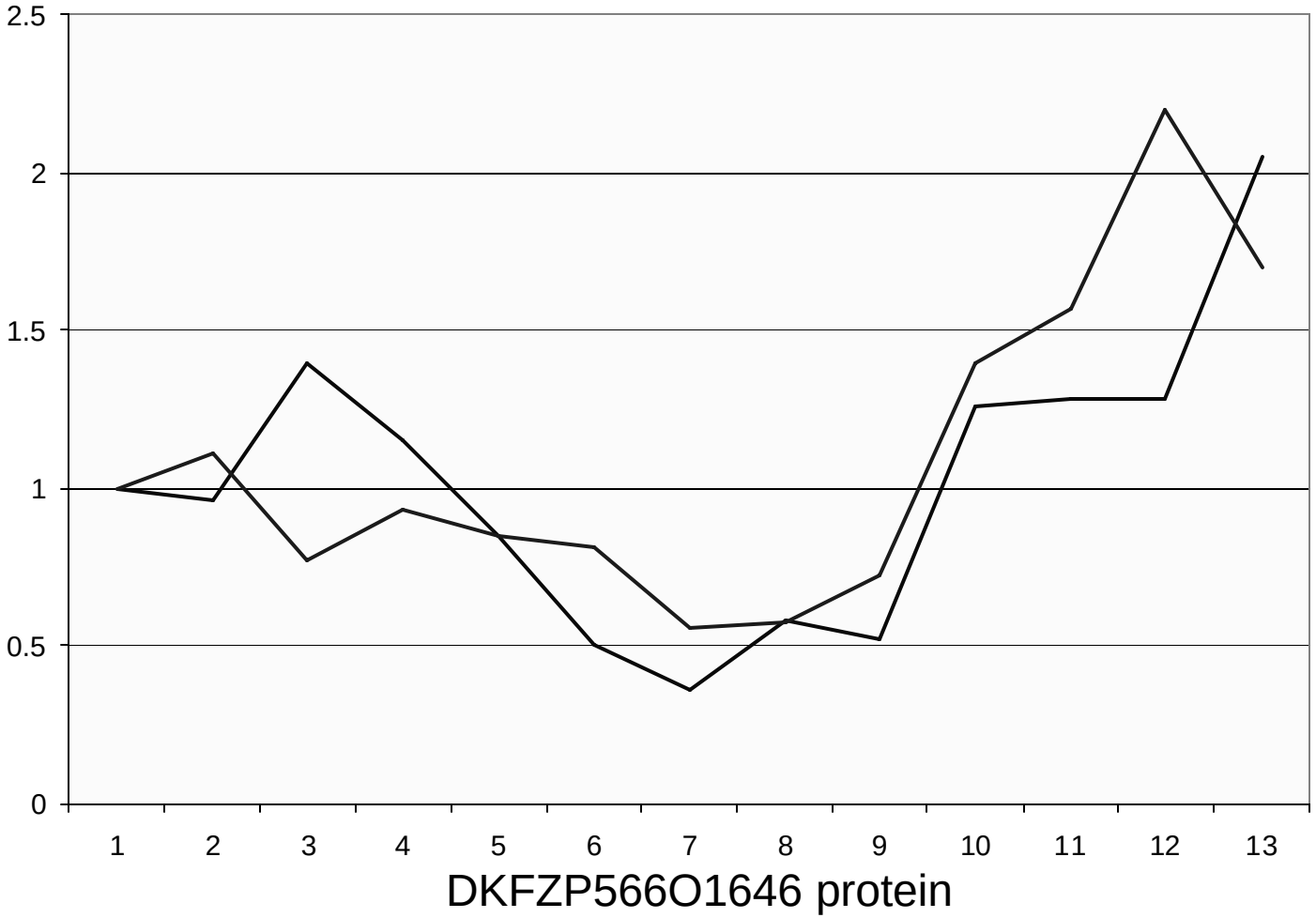
- Back in 1999, 238 out of 517 genes were unknown.
- We relabeled the genes according to the current state of the art and less than 20 are left unknown.
- There are 19 *repeated genes* in the dataset:
  - ✓ Original clustering puts 4 of these in different clusters;
  - ✓ We put 1 of these in two different clusters.
- Interestingly enough, if we run the clustering with Markov order 0 (assuming uncorrelated iid data), we get 4 “misplacements” as well, albeit of other genes.
- Conclusion:
  - ✓ Temporal order provides more accurate insights into concerted behavior of gene expression.
  - ✓ Sound statistical basis for models rather than the “looks right” test
  - ✓ Now being applied to several time-series in heart disease model, brain development, transplant rejection, insulin effect

They are the same



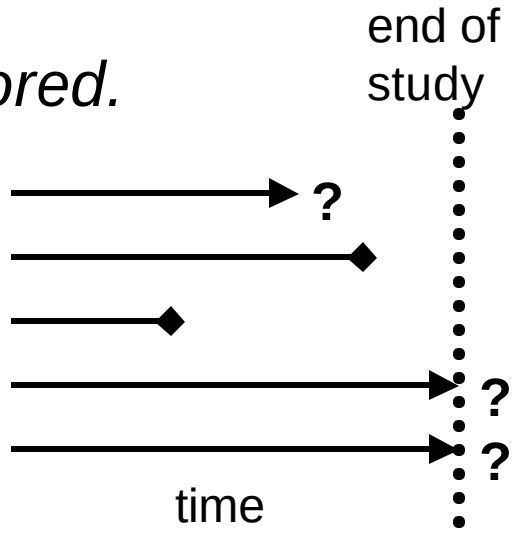
Tax1 (human T-cell leukemia virus type I) binding protein 1

Are they really the same?



Using patient survival times

- Patient survival times are often *censored.*
  - a study is terminated before patients die
  - a patient drops out of a study



- If we exclude these patients from the study or treat them as uncensored, we obtain substantially biased results.
- The phenotype can denote time to some specific event, e.g., reoccurrence of a tumor.

Previous studies

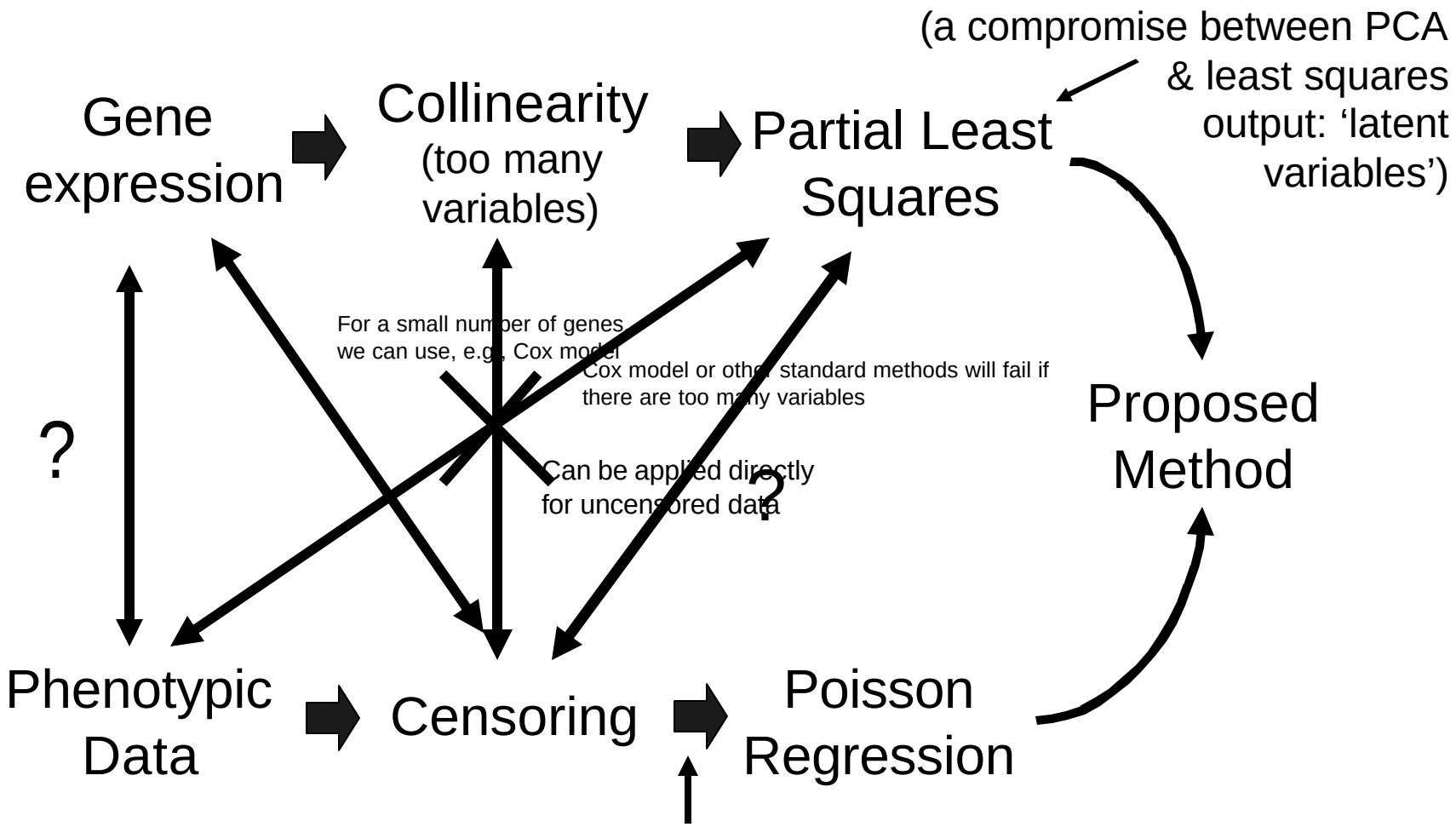
Please see Nature. 2000 Feb 3; 403(6769): 503-11.
Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.
Alizadeh AA, et al.

Most studies so far used survival curves as a way to verify the results of unsupervised clustering.

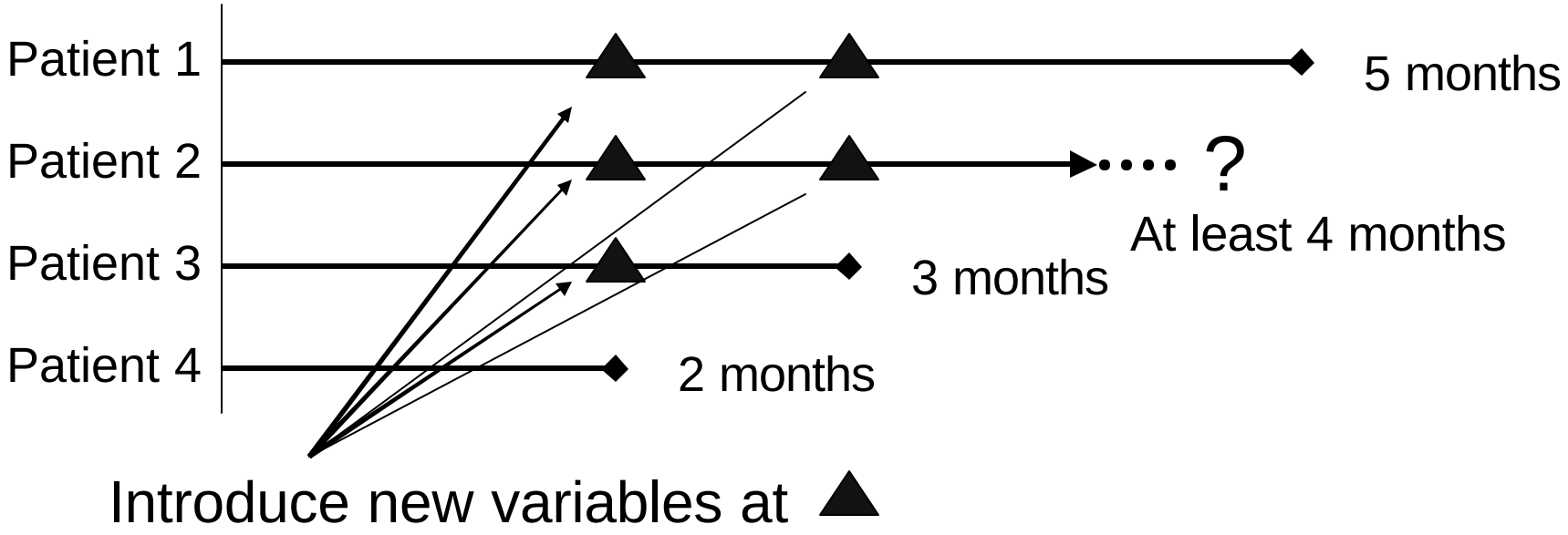
The Question

- Another way to deal with the censoring: turn survival times into a binary indicator, e.g., 5-year survival rate. → loss of information
- Question: Can we directly find genes or linear combinations of genes that are highly correlated with the survival times?
- For example, (gene A + .5 * gene B + 2 * gene C) may be highly predictive of the survival time.
- How can we use the survival times directly to find good predictors?

The Big Picture:



Intuition behind the transformation



Just before each observed failure time, patients who are alive have the same chance of failure. We can also consider the variables at each failure time to be independent (Cox model)

(Whitehead, 1980)

Putting it together: Example

Bhattacharjee, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, *PNAS* 98:13790–13795, 2001.

Total of 186 lung carcinoma and 17 normal specimens.

125 adenocarcinoma samples were associated with clinical data and with histological slides from adjacent sections.

The authors reduced the data to few hundred reliably measured genes (using replicates).

Patient	Survival	Censor
1	25.1	1
2	62.6	0
3	7.3	1
4	22.3	1
5	41.2	1
6	66.8	1
7	75.4	0
8	50.1	0
9	60.5	0
.....		
.....		

Example: Results

One way of evaluating significance of latent variables is through the Cox proportional hazards model

Latent variables using Partial Least Squares

p-values

Principal Components

1	9.3E-007
2	1.8E-007
3	5.2E-005
4	3.6E-007
5	2.1E-008

Low p-values indicates high correlation between the latent variable and the survival time

1	.0029
2	.0170
3	.4400
4	.2600
5	.0930

Cross-validation is a bit tricky with censored data!

Example: Results

Kaplan-Meier survival curve based on the first latent variable:

Please see Peter J. Park, Atul J. Butte, and Isaac S. Kohane
Comparing expression profiles of genes with similar promoter regions
Bioinformatics 2002 18:1576-1584

The p-value is .00002 for the null hypothesis of no difference between the two groups.