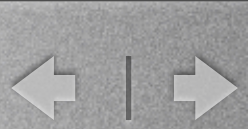
Understanding Data Complexity through Models & Computation

Stephen J Merrill
Department MSCS
Marquette University
Milwaukee

ISL Workshop

Wednesday July 12, 2006

Four Examples

1. HRV (heart rate variability) in pediatric patients
2. Spatial variation in the cDNA microarray
3. Early HIV infection dynamics
4. Dynamics of engraftment in hematopoietic stem cell transplants

2


Common Themes

- Complicated dynamics apparent in data
- Data needs to be seen through the lens of a model to reveal important features
- Model form dictated by the nature of the data and the process
- Questions of interest dictated by a discipline outside of mathematics

3


1. HRV Dynamics

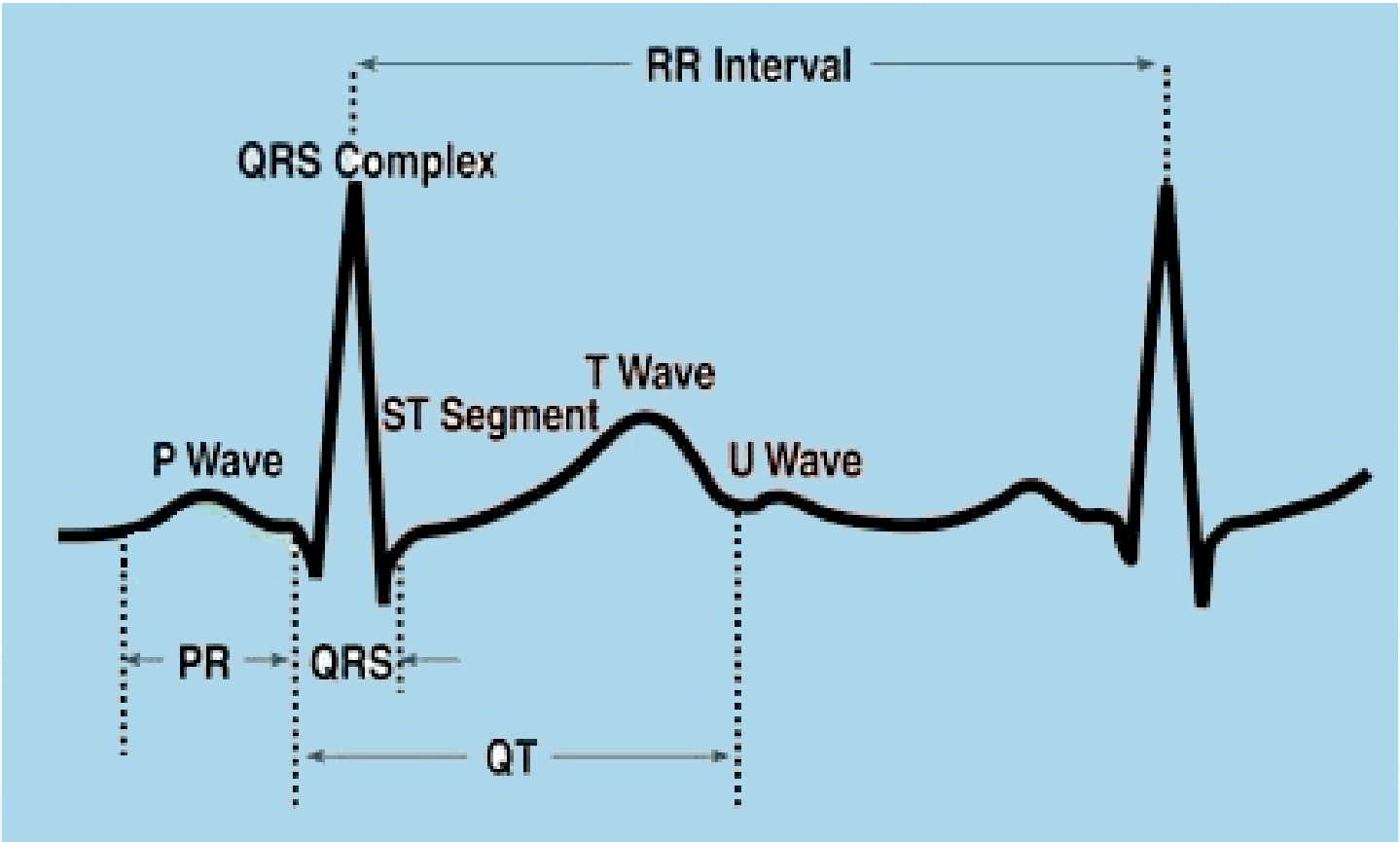
- Heart Rate Variability (HRV) describes the beat-to-beat variation in the time interval between beats as seen on ECG. It is described by many different indices.
- The variability is due to several different control mechanisms in the systems
- Operation of the controls are affected by drugs (specifically here, anesthesia)

4


Application and Background

- Pediatric patients undergoing surgery
- Goal was to design a real-time monitor as sudden cardiac arrest is an issue
- Data had been collected on several patients and indices did not behave well as measures of HRV in several patients

5



ECG illustrating RR interval

6

The Data

2.5 year
old girl

7.5 year old boy --
<u>early phase</u> with
halothane
<u>late phase</u> with the
addition of atropine

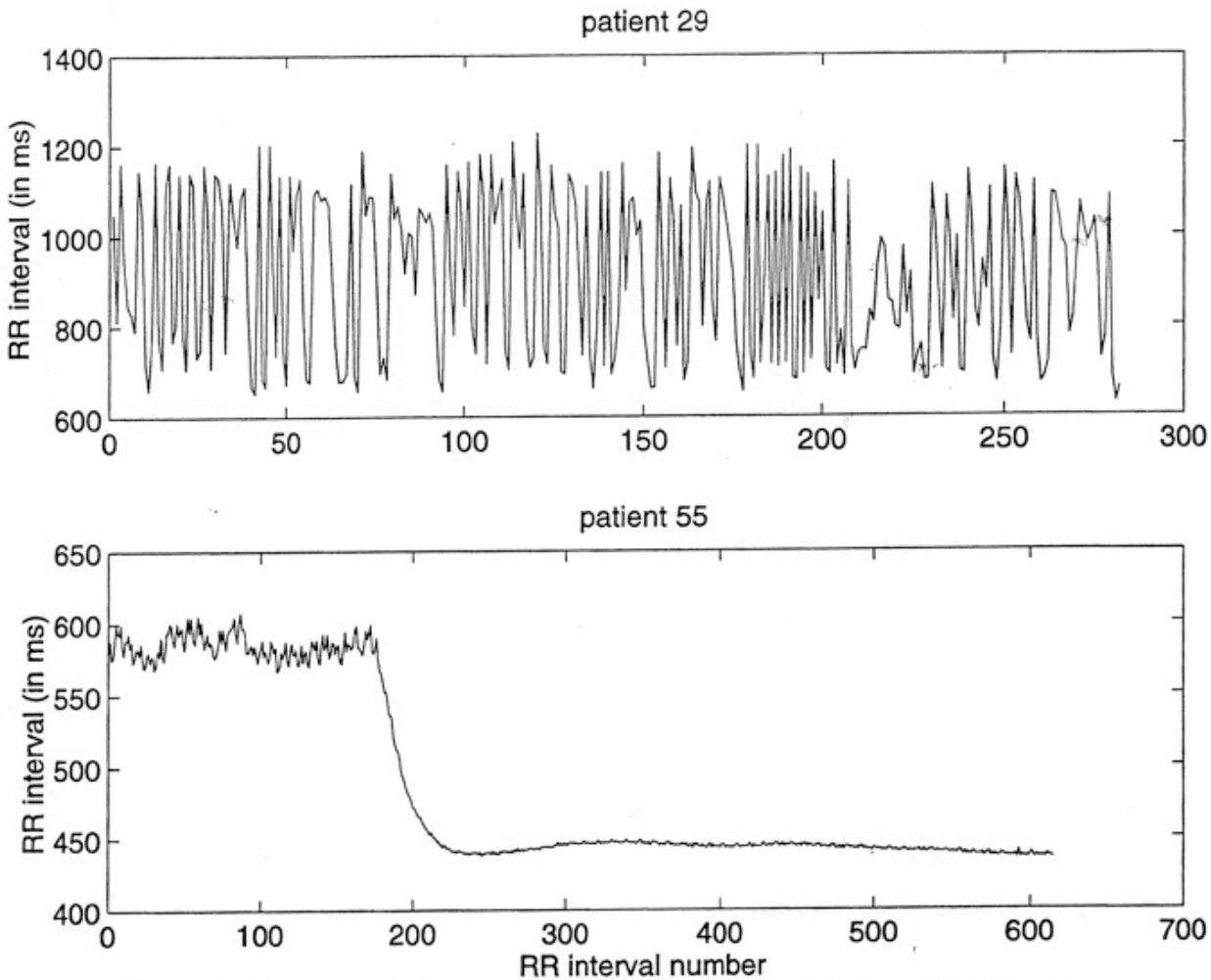


Figure 1 RR interval data from Patient 29 and Patient 55. In the analysis, the second set will be divided into an early and late phase.

7

Lag 1 maps

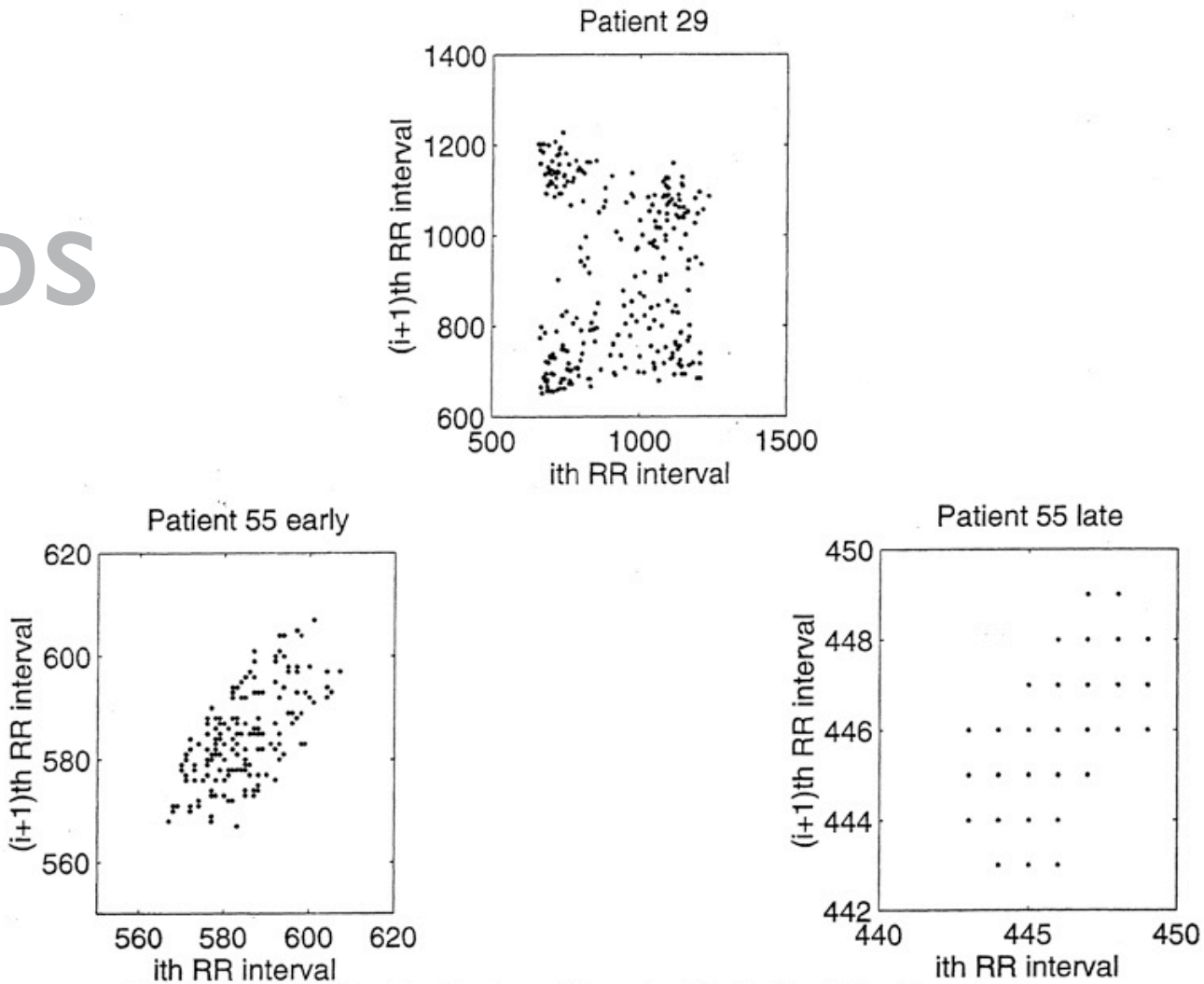


Figure 3 Lag 1 plots for the three data sets. The Patient 29 pattern has been described previously as a "complex pattern" (Woo et al. [1992]). Patient 55 early and late data plots would be classified as "torpedo patterns" by Woo et al. [1992].

8


Model of RR interval data

- Create an empirical Markov chain. Data is in the form of sequence of numbers and lag 1 maps indicate first order structure.
- Need to define the bin size corresponding to the length of the data set (usual number of bins used was 10). Then estimating transition probabilities to get a transition matrix. Note that many possible transitions are not observed.
- Transient aspects of the chain are of interest (not asymptotic behavior). Characterization of the dynamics (or the resulting matrix) is desired.
- Basic idea is to use properties of the matrix (such as eigenvalues) to distinguish between cases.

9

All eigenvalues of a Markov chain transition matrix lie in (or on) the unit circle. Uniqueness of a modulus 1 eigenvalue means the presence of a limit distribution

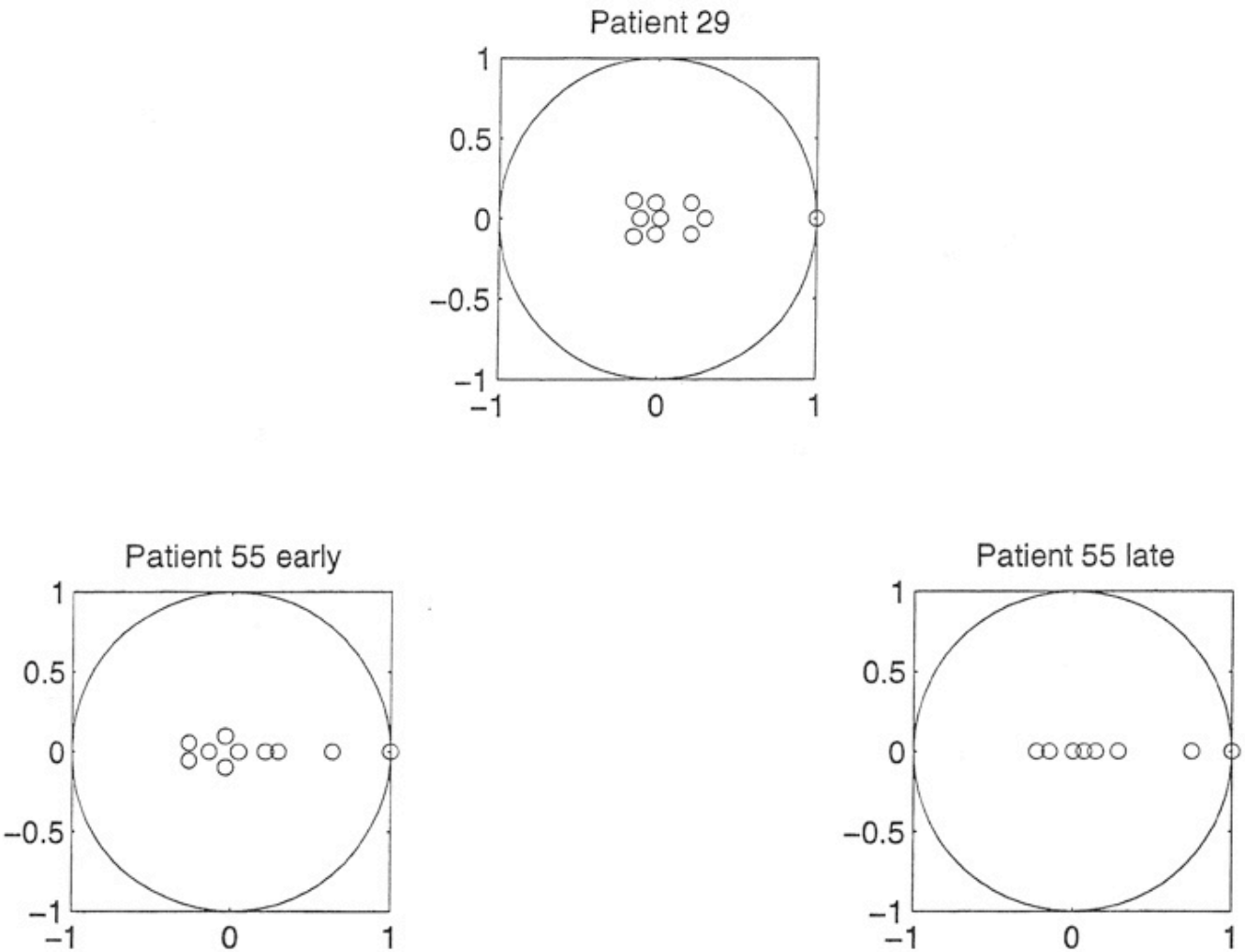
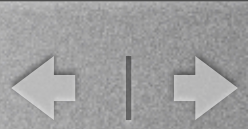


Figure 5 Eigenvalues for each of the three transition matrices are shown in relation to the unit circle. Besides the nearness of the "non-1" eigenvalues to the unit circle, notice the differences in the number of complex eigenvalues.

10

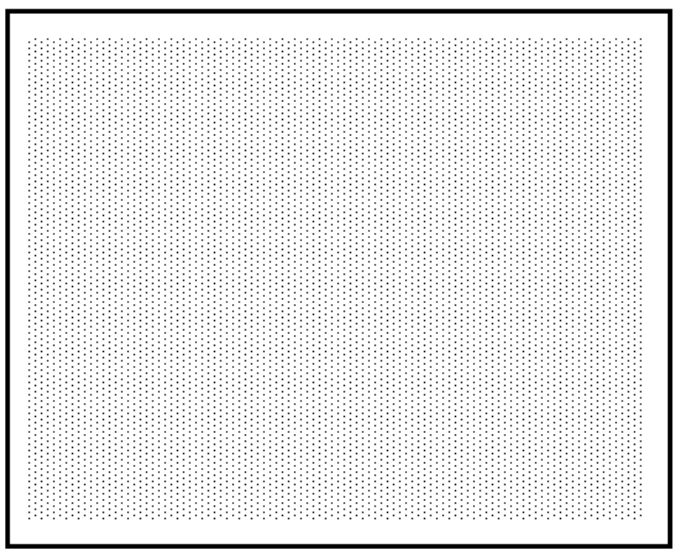



Computational Issues

- Estimation of transition probabilities
- "Sensitivity" of matrix properties on the estimates
- Eigenvalue estimates and the sensitivity of eigenvalue estimates on the transition probabilities
- Tracking eigenvalues as in bifurcation

2. Spatial variation in the cDNA microarray

- cDNA microarray used to identify genes that are differentially under or over expressed in a sample (as seen through mRNA).



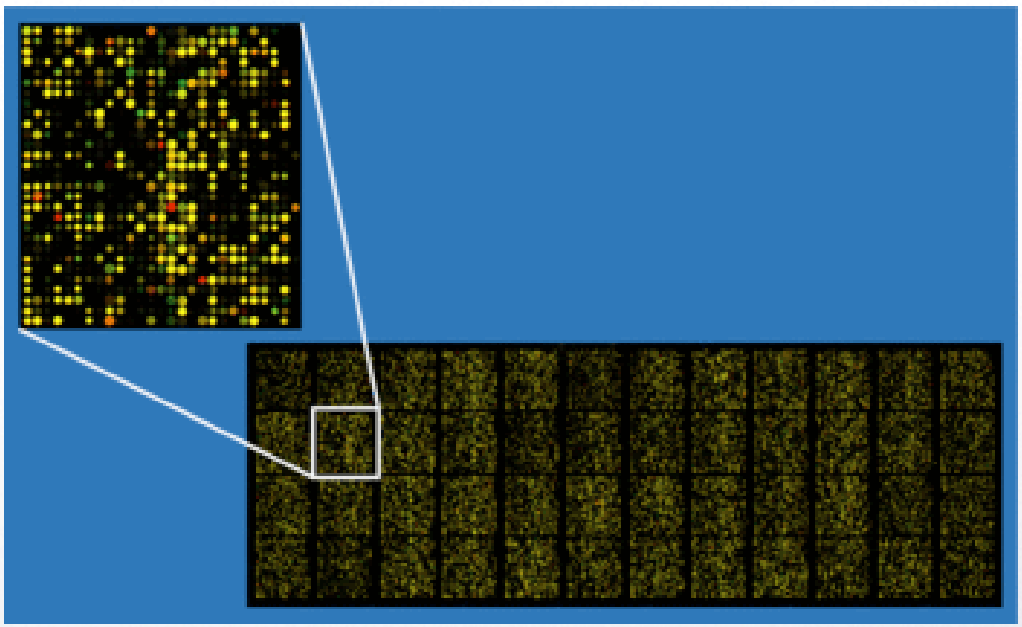
A bit of the process

- the slide or chip is printed with a library of genes including those of special interest 
- collect mRNA under two different conditions. Using RT and two different fluorescent dies, samples of labeled ("red" and "green") DNA are produced.
- incubate the samples with the slide under a cover slip.
- scan the result to measure the amount of red and green fluorescence at each spot to measure the relative amount of mRNA present in the two samples.

13

Application and Background

- High variability -- both between replicates and within the same slide (with duplicated specificities in dots).
- Spatial variation in the brightness observed ("bright edges")
- Need to understand the proper normalization for this process

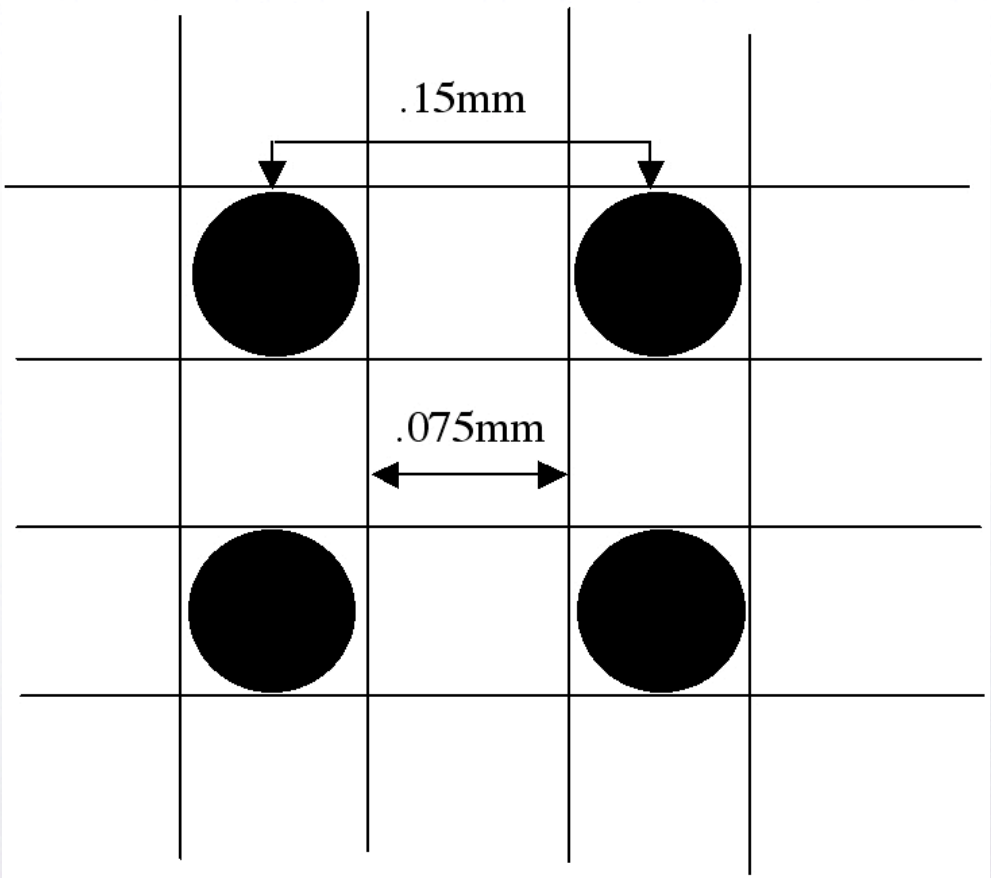


14

Model of microarray hybridization

- Using the natural grid of positions on a slide, a Markov corresponding to each of the 16,000 dots is constructed. The goal being to compute the probability of absorption as a function of the transition number.

- The transition probabilities are based on the "taxi-cab" metric on the grid.



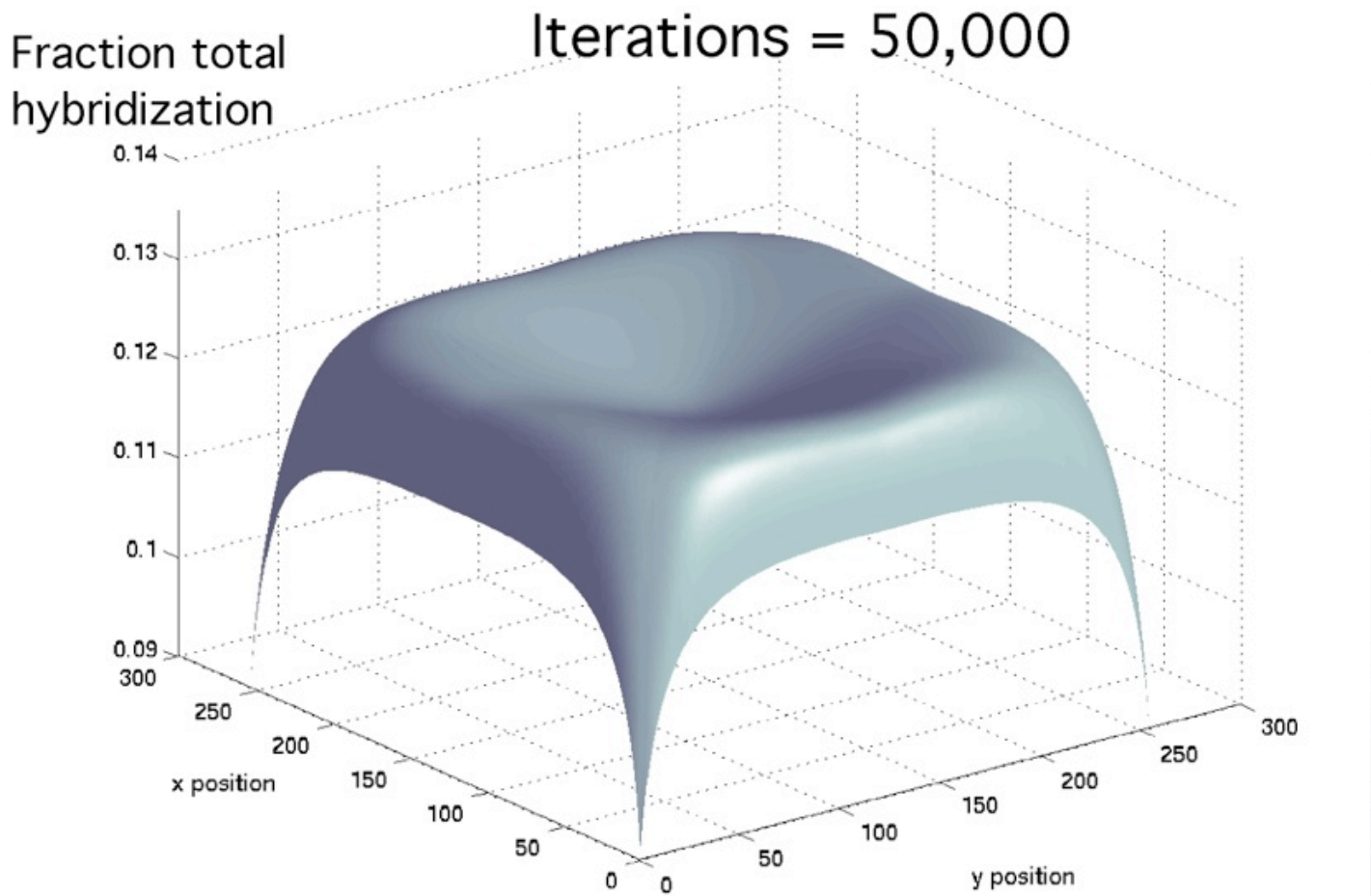
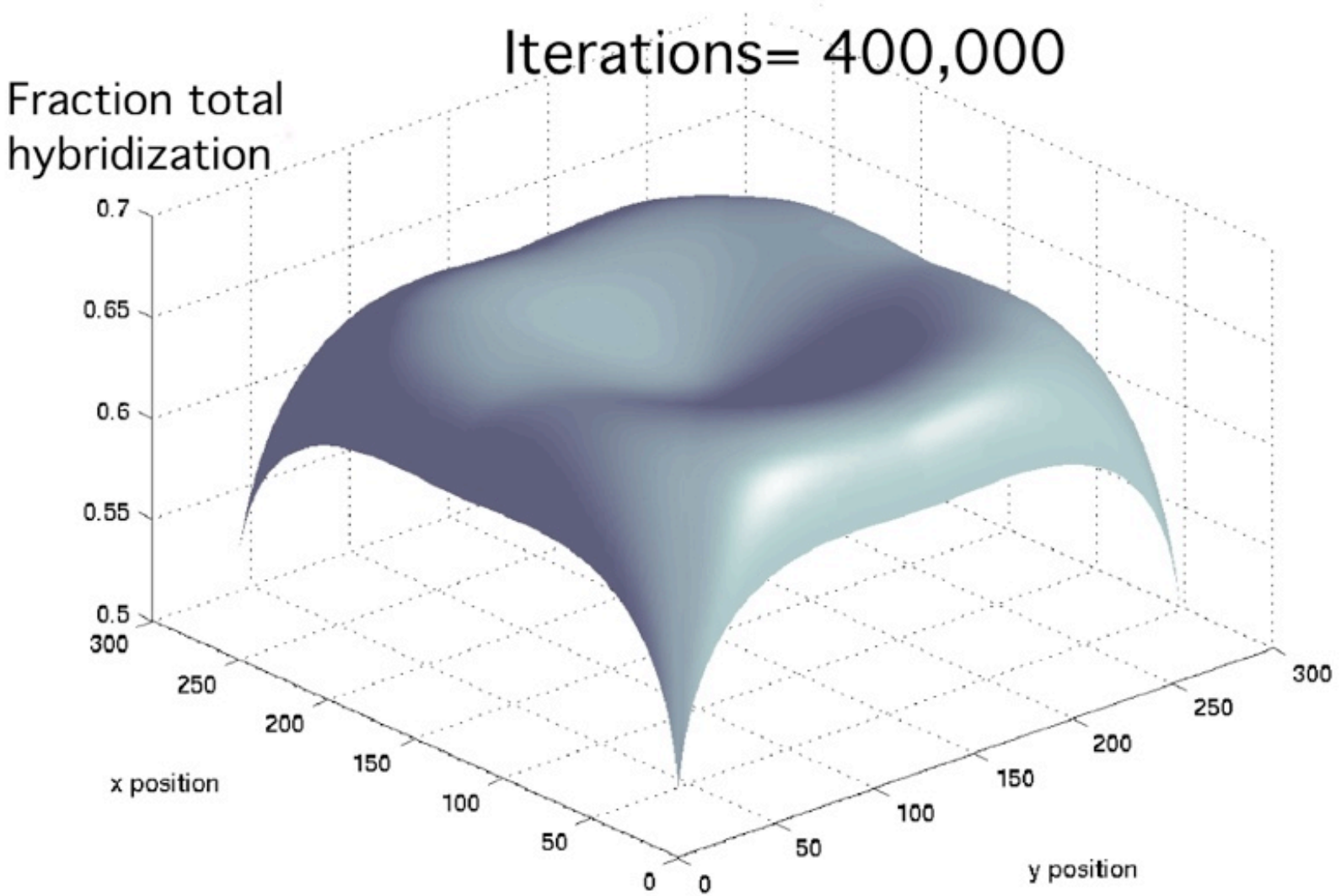


Figure 6. Results from 50,00 and 400,000 iterations of 252 x 252 different Markov chains simulating the fraction of possible hybridization for spot positions in that grid after a simulated 1.5 hour and 12 hour incubation, respectively. Note the "bright edges" and the lower hybridization in the center. Simulations and graphs from Matlab 6.5 (Natick, MA USA).

16


Iterations= 400,000
Fraction total hybridization
0.7
0.65
0.6
0.55
0.5
300
250
200
150
100
50
0
x position
0
50
100
150
200
250
300
y position

17

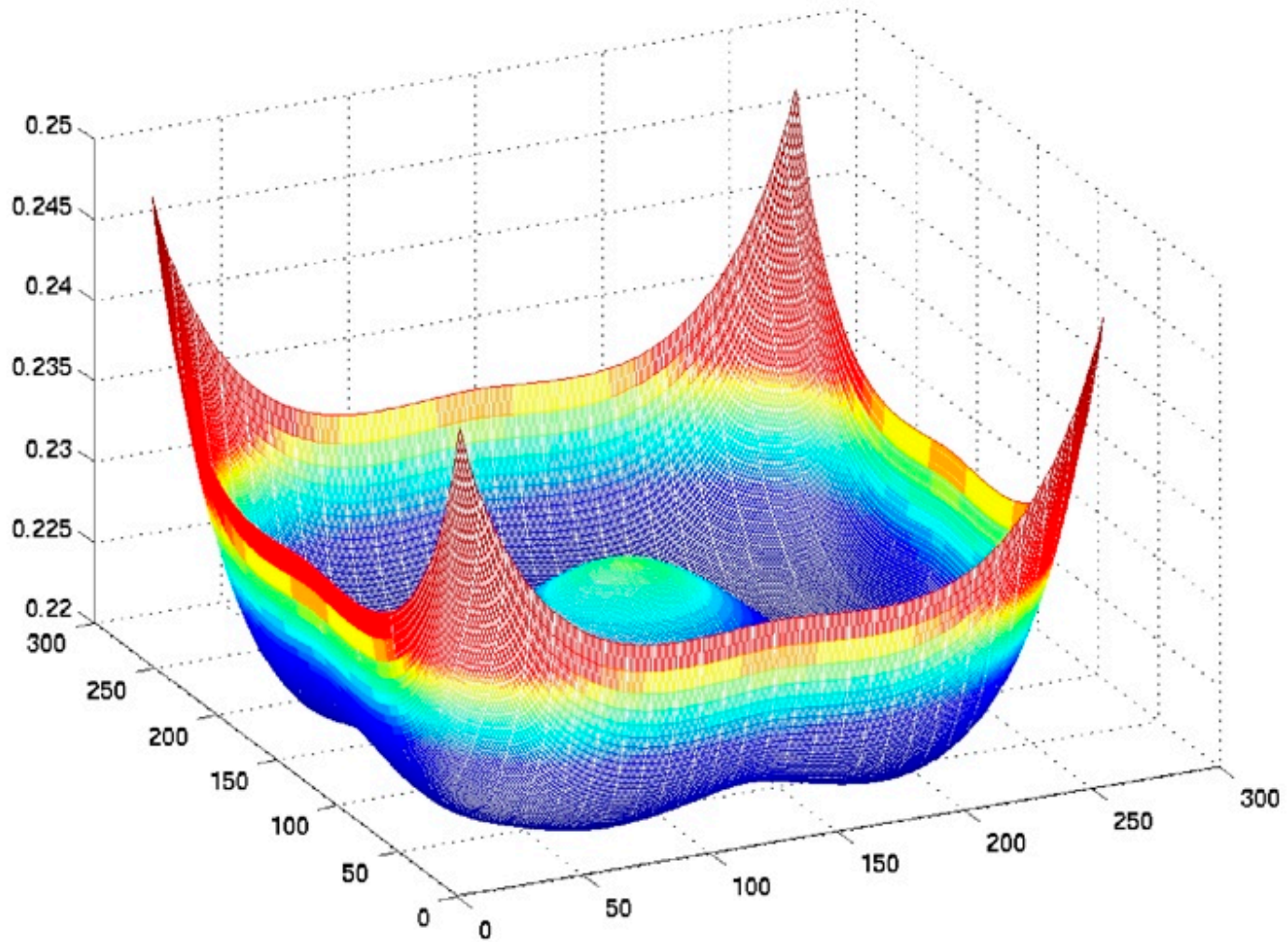


Figure 7. The variance in the hybridized fraction at 400,000 iterations as a function of position.


Computational Issues

- 252 x 252 different Markov chains, each with hundreds of states (depending on the position of the dot corresponding to the chain).
- Modeling the possibility of replicates on the same slide and dealing with how the data should be analyzed.

19

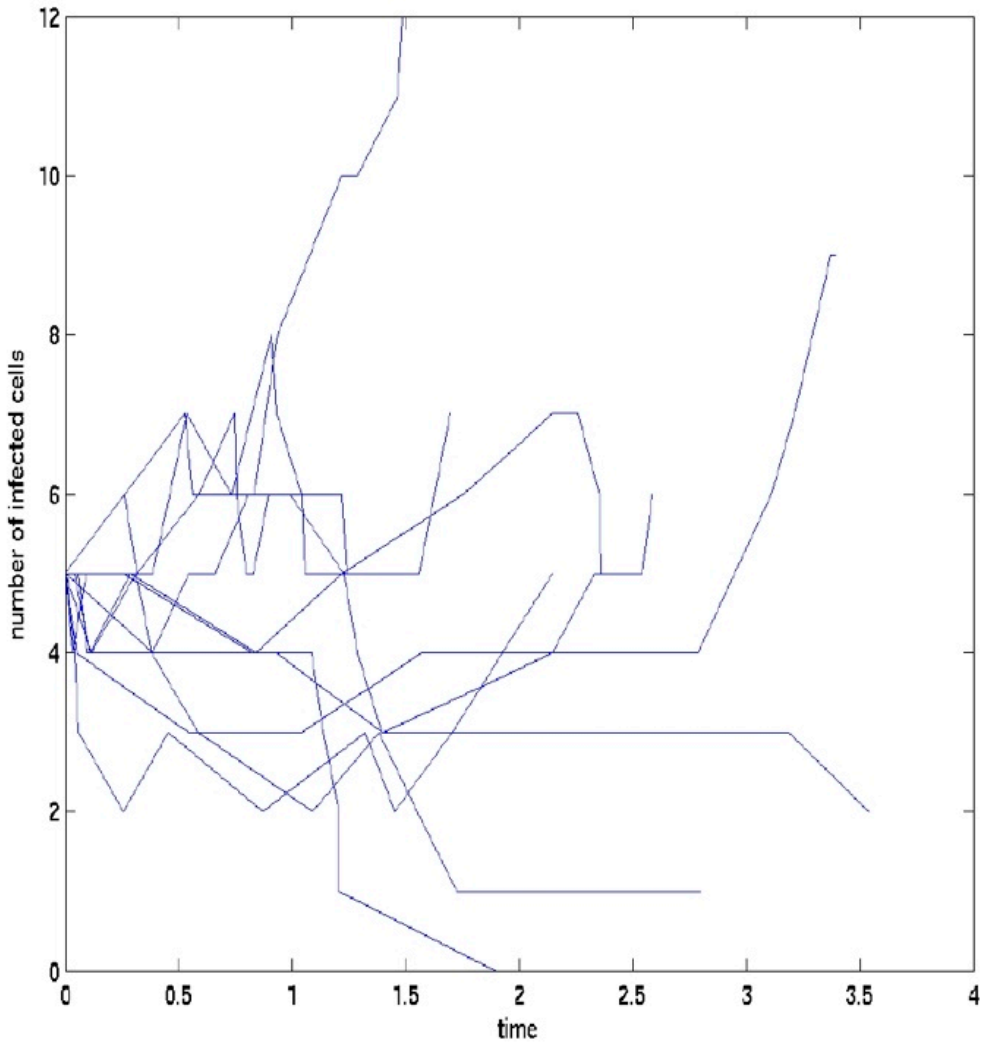

3. Early HIV infection

- Long term time-course of the infection depends on the "set point" -- related to the state of the infection at the time the immune response controls the initial acute infection.
- Interested in computing the incubation-time distribution (defined as the time from infection to a fixed clinical marker such as seroconversion -- the appearance of anti-HIV antibodies).

20

The model

- Branching process with immigration. The basic quantity tracked is the number of infected T cells
- Only data that exists is th distributional information
- Model designed to see what happens in the initial stages.

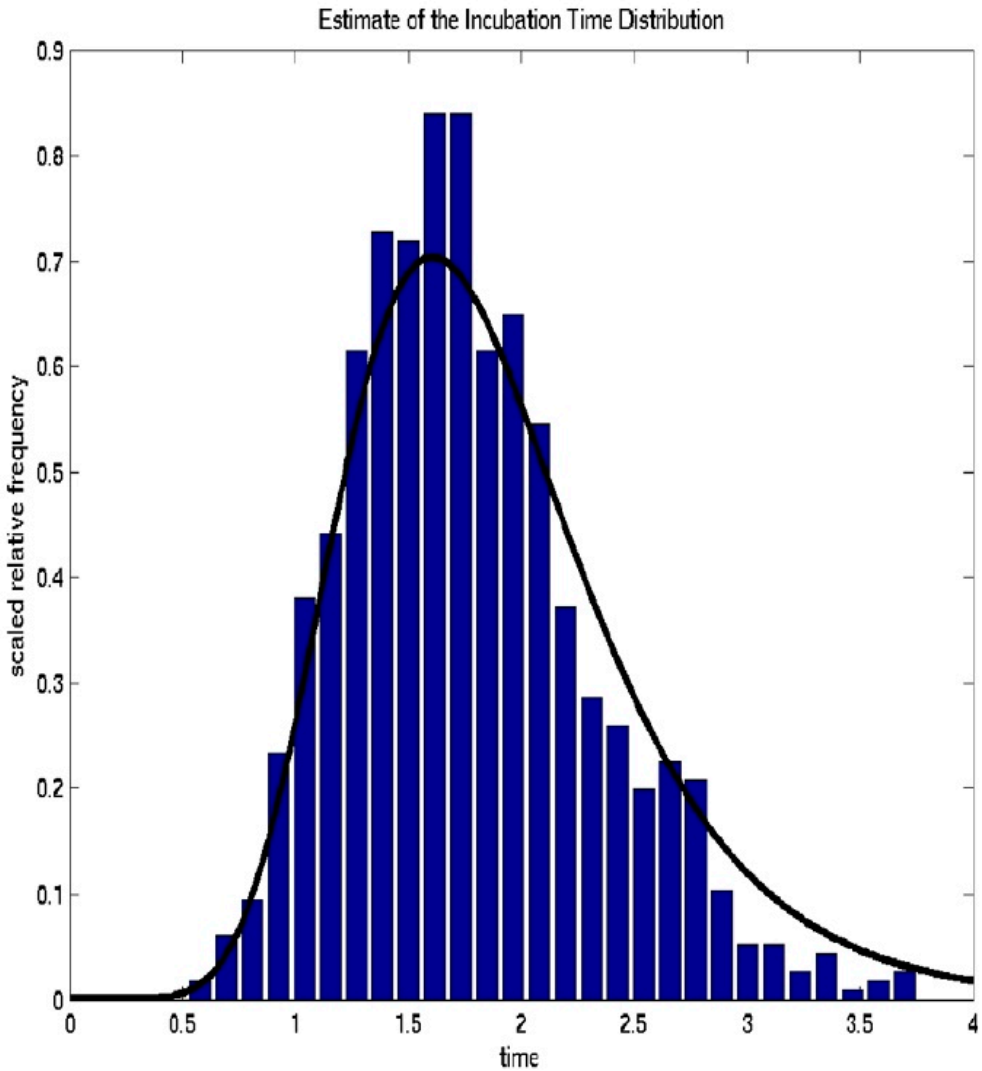


Sample Paths of branching model -- no immigration

21

Simulations

- 1000 simulations of model, recording the time at which the number of infected cells reach some fixed value

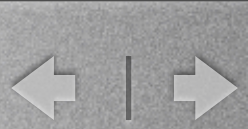


22


Computational Issues

- Intensive computation required to obtain distributional information.
- Parameter-based results (such as sensitivity) were difficult.
- Combination of following intervals of values, instead of sample paths along with probabilities -- or densities may have been useful.

23

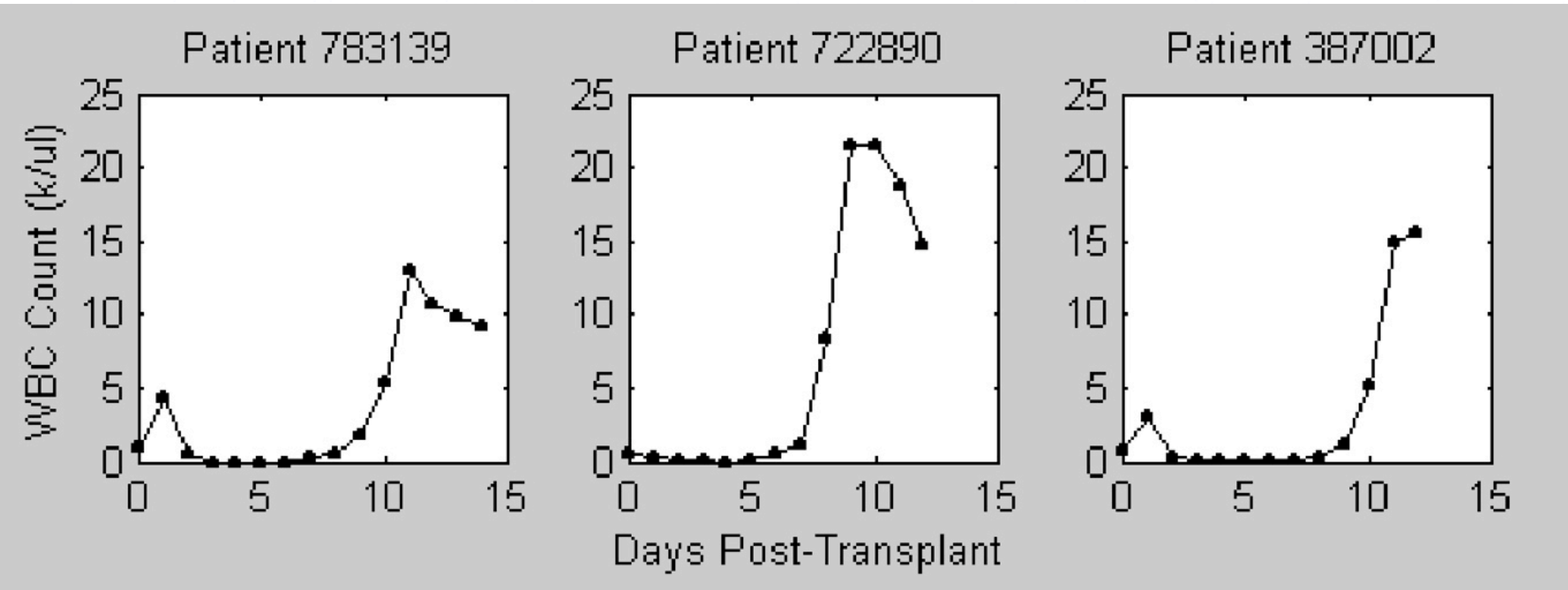



4. Dynamics of Engraftment

- Hematopoietic stem cells can be collected from blood (or bone marrow) for later infusion (transplantation) after high-dose chemotherapy.
- In autologous transplants, no rejection is present.
- Interested in monitoring engraftment (return to normal levels) of each cell type -- primarily leukocytes (WBC in early counts), lymphocytes, and platelets, and red cells.

24

The Data

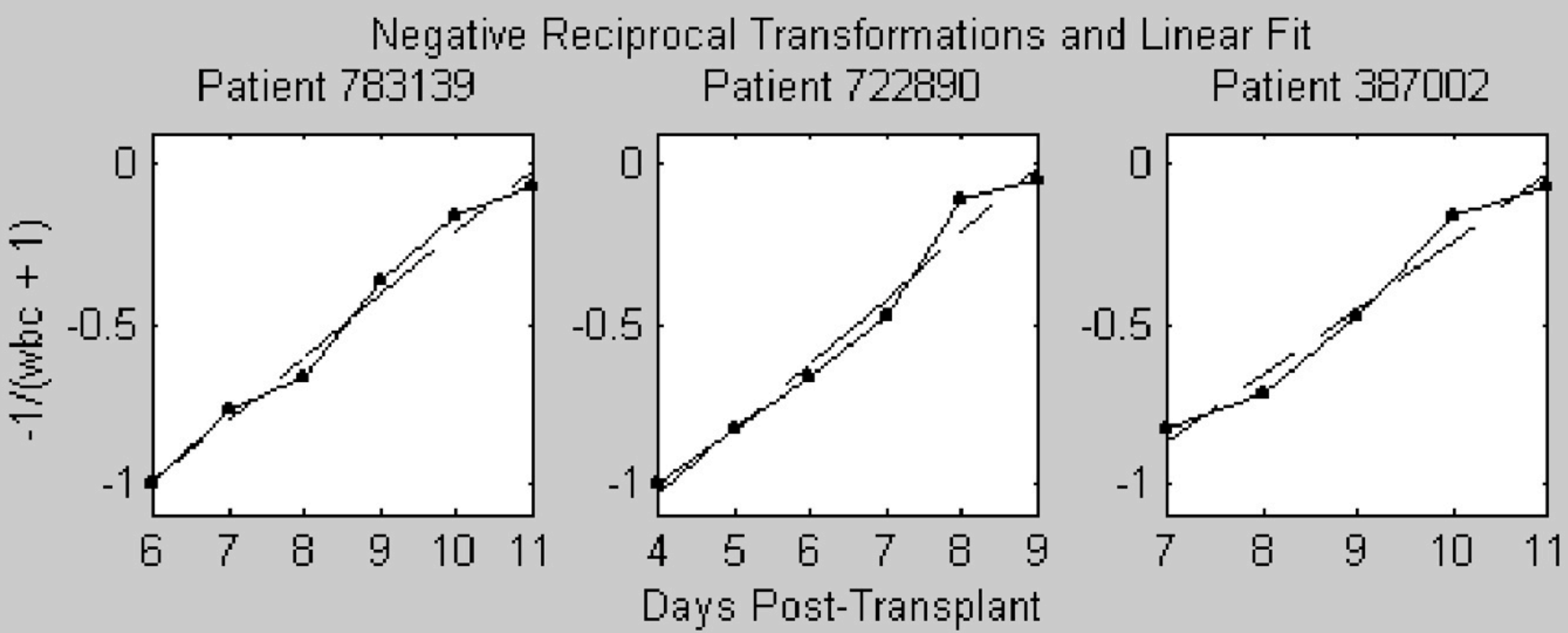


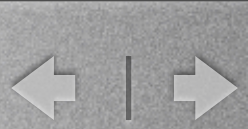
Results of Daily White Blood Cell counts for 3 patients post transplant

25

The Model

- Reciprocal plot shows hyperbolic growth $r^2 = .94$

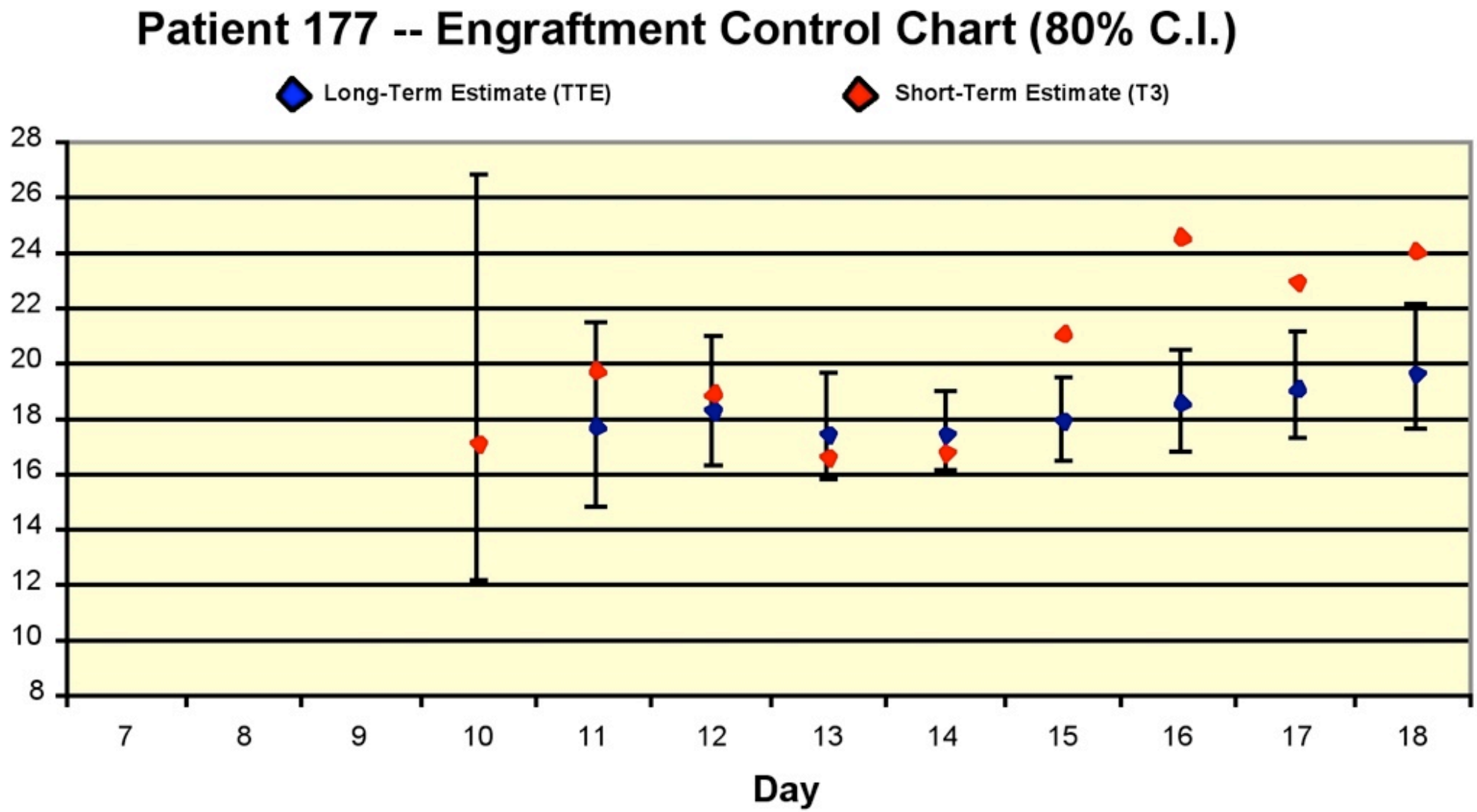


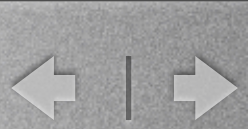
Results

- Estimating the position of the asymptote (and as it changes with each days data) allows the estimation of the time to engraftment -- and resulting release.
- Changes in the estimates indicate problems, represented in a control chart.
- Similar results for lymphocytes. For platelets, polynomial growth was observed.

27



B.M. Murphy, Modeling the time to engraftment ..., 2001.

Computational Issues

- The need to generate the non-symmetric confidence intervals using simulation -- using distributions of slopes and intercept estimates from the regression.
- The hyperbolic growth $(x' = x^2)$ phase of the process caused issues with the de solvers.

29


Common Themes

- Complicated dynamics apparent in data
- Data needs to be seen through the lens of a model to reveal important features
- Model form dictated by the nature of the data and the process
- Questions of interest dictated by a discipline outside of mathematics

30


References

- S.J. Merrill & J.R. Cochran (1997) *Markov chain methods in the analysis of heart rate variability*, Fields Institute Comm. 11:241-252.
- S.J. Merrill, S. Nelson, & C.A. Struble (2003) *Spatial dependence of hybridization in the cDNA microarray*, Can. Appl. Math. Quart. 11: 321-337.
- S.J. Merrill (2005) *The stochastic dance of early HIV infection,* J. Comp. Appl. Math. 184:242-257.
- S.J. Merrill & B.M. Murphy (2002) *Detection of autocatalytic dynamics in data modeled by a compartmental model,* Math. Biosci. 180:255-262.

31