# The Cytometry Game: Past and Future

C. Bruce Bagwell MD, Ph.D.

Verity Software House

OK, cytometry is a pretty cool technology, but a game??? I'm not kidding you. This lecture is design to show that cytometry is every bit as fun and challenging as a good game of Sudoku or a New York times crossword puzzle. I'm also going to show you that the cytometry game we have been playing for the last 30 years has been very powerful, but there is a future new game that is even more powerful. Not convinced that cytometry is a game…well prepare yourself.
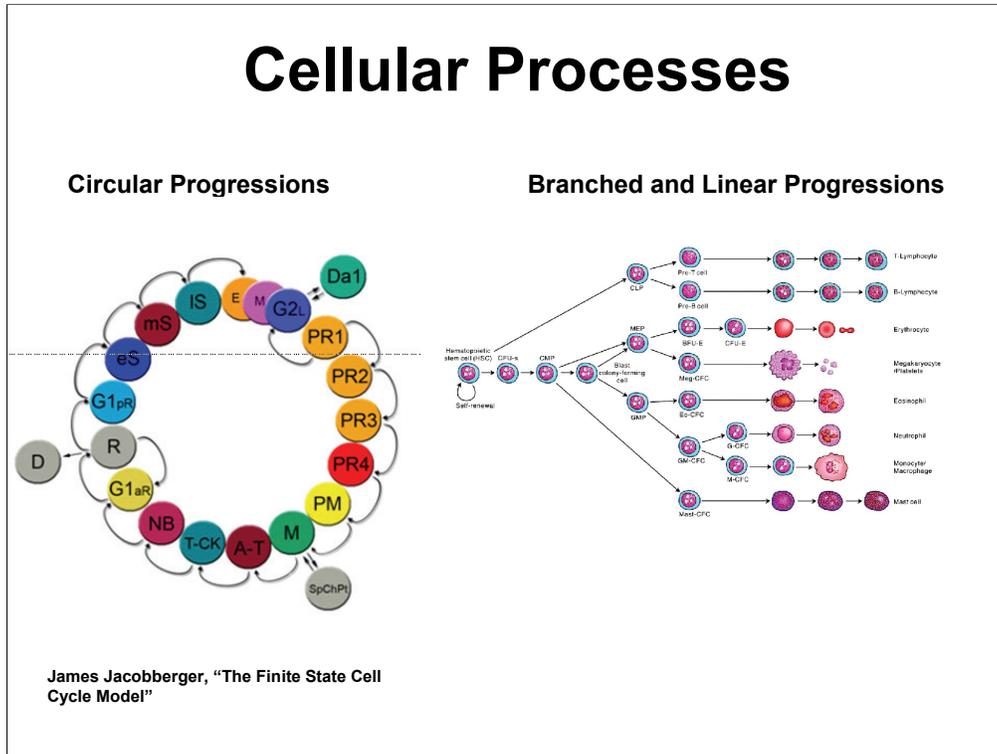
Cells are amazing. The more we learn about how cells do what they do, the more amazing they become. Living cells involve thousands of biochemical processes. These processes are essentially chemical machines that move from state to state along thermodynamic gradients. Some processes or progressions are circular in nature such as found in cell division. Eukaryotic division involves the interplay of all kinds of proteins. In human bone marrow, there are numerous progressions that manufacture our red cells as well as cells involved in immunity. Everywhere you look, cells do what they do in distinct progression steps which can be discerned by the technology of cytometry.
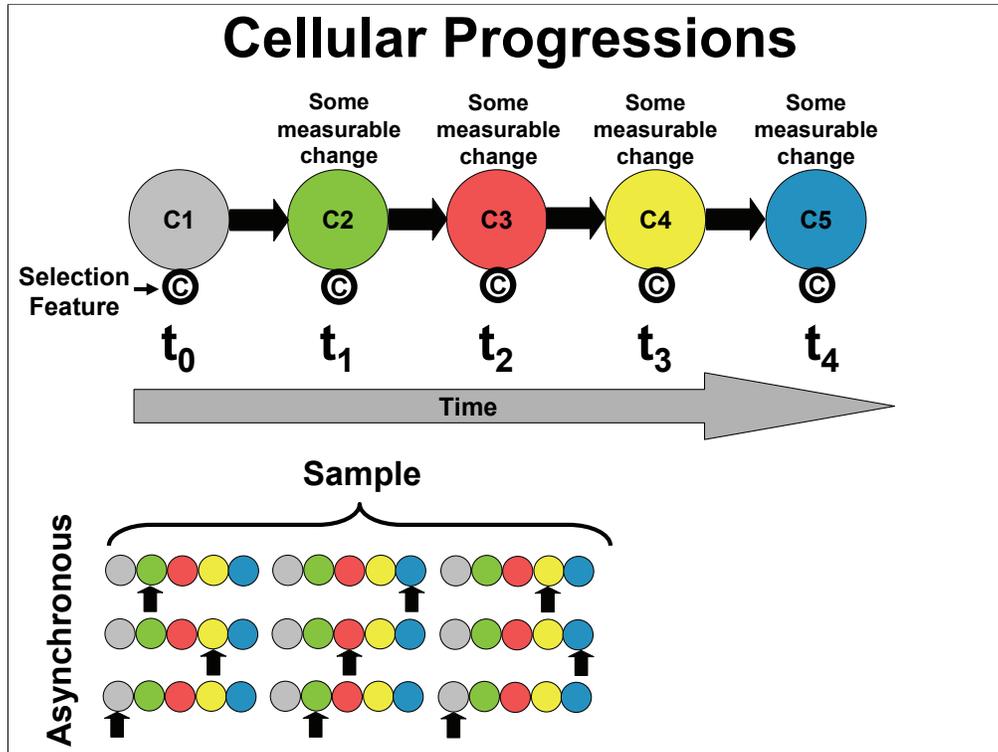
Cytometry plays an essential role in the understanding of many of these processes or progressions. Let's take a simple linear progression apart to better understand cytometry's critical role in developing our understanding of the order of progression steps. We'll keep things general at first so we can talk about all progressions before narrowing in on a few examples.

Progressions are kinetic processes that involve time. In our example, we start the progression at time t_0 with a cell labeled as C1. We can select for this progression because these cells have some unique signature of measurable features. In this particular example, we will assume the cell has a lot of protein C on its surface which can be leveraged to investigate just this progression in a sea of other progressions. Because of some signal, something that is measurable on or in this cell changes at a later time, t_1. It could be the up- or down-regulation of one or more proteins or some change in an intrinsic feature such as cell size. When we detect this change with cytometry, we can assign this cell to a new state or stage and label it C2. As the cell moves along its progression C3->C4->C5, these measurable features change, which can be detected by cytometry.

When we examine a sample with a cytometer, we catch all these cells at some point along their progressions. This asynchronous distribution of cells in all stages of a progression is one of the reasons we can work out steps in a biological progression. The next few slides will explain how we go about figuring out the order of the progression. As you might imagine, this is where the game comes in.
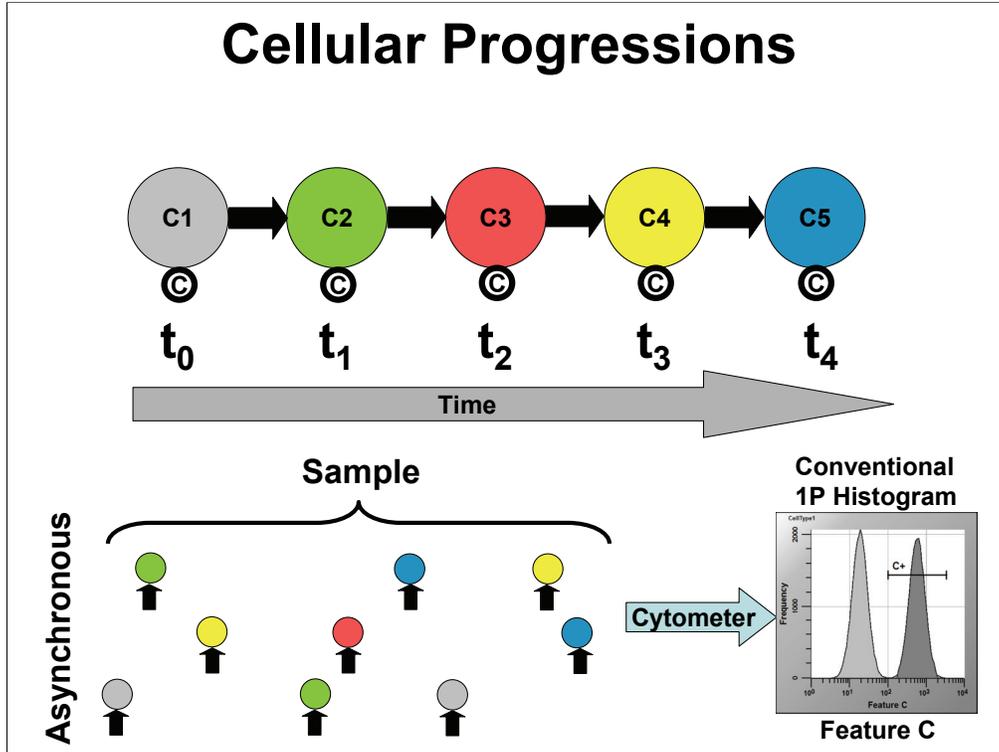
The cytometer makes all the measurements on each cell and stores the information in special kind of database called a listmode file. If the operator puts a gating range around the C+ events, we can view and analyze just the cells that are involved in our C+ progression (above).
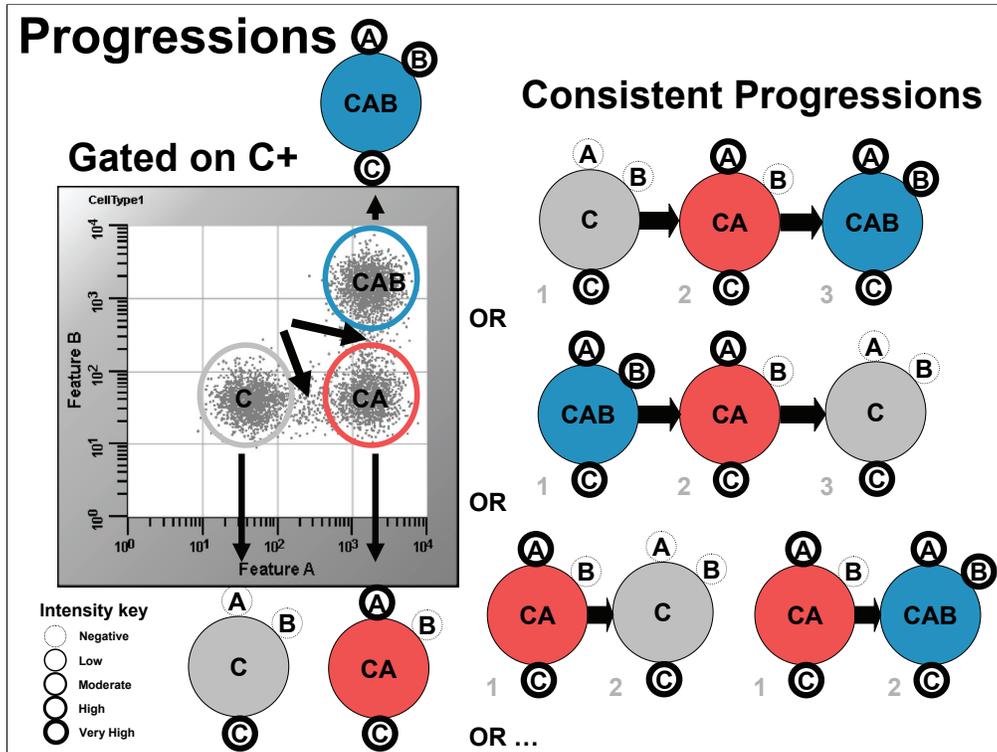
Let's now move from the general case to something a bit more specific. Let's assume that we are measuring proteins A, B, C, D, and E on the cell. We use C to select for the cells of interest and we'll use the rest of the markers to determine the steps of the progression. Once gated, we can inspect a dot plot of feature A vs. B. Assume we see the pattern of dots shown above. We can see three clusters of dots which we can label as C, CA, and CAB. We can also see a continuum between C and CA and another one between CA and CAB. We don't see any continuum between C and CAB. These continuums are important because they are like tracks in the snow and normally show where a progression path is.

What can we deduce about the progression with this data? One possibility is that C is the first state of the progression, then CA, and then CAB (top-right). But, if you think about it, we would see the same exact data if the progression started at CAB, went to CA, and finished with C (middle-right). In fact, we could have a branched progression that started at CA and some events went to C and others went to CAB (bottom-right). The progression could be more complicated. It could be C to CA to CAB and then back to CA again. The more you think about it, the more you realize that there are quite a large number of progressions that are consistent with this data pattern.

How can that be? If there are an infinite number of progressions, how can we use cytometry to better understand progressions?

Progression Order and *A Priori* Information
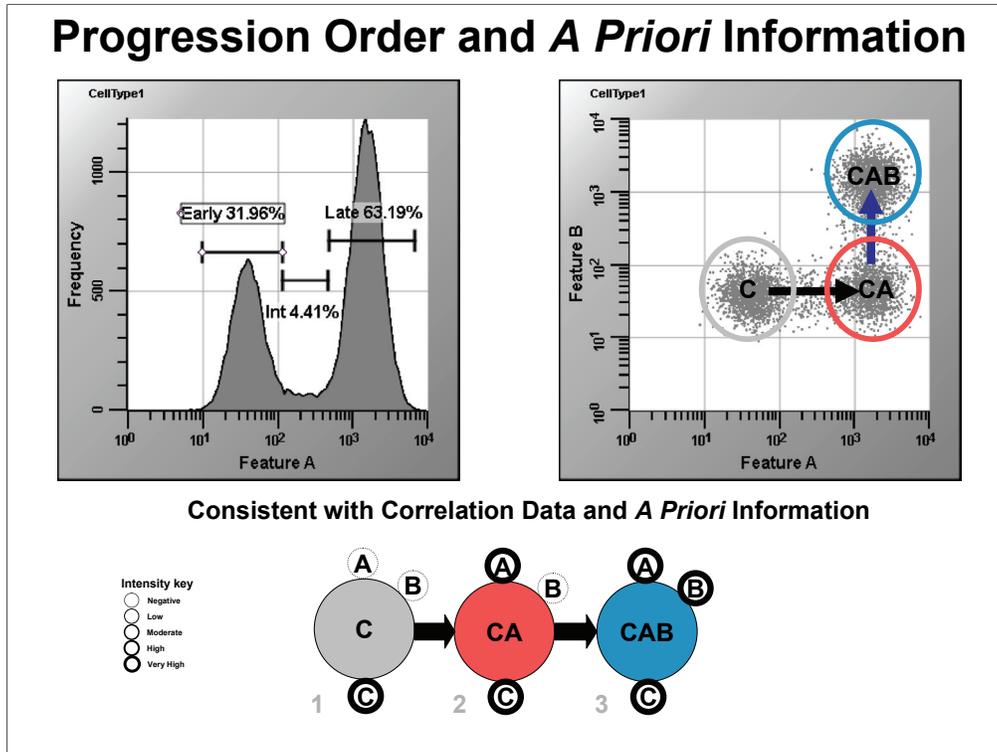
In order to figure out a temporal progression, it is necessary that experimentation be performed to determine the directionality of one or more of features in a progression. These experiments can include pulse labeling, cell synchronization, sorting with morphological and/or biochemical testing, sterile sorting and subsequent tissue culture, and many others. In other words, for us to interpret our data, we absolutely need at least one piece of information that tells us the directionality of changes in one measurement along the progression.

In the above example (top-left), assume this experimentation was done and it was found that the first peak in a feature A histogram represents the beginning of a cellular progression and the last peak represents the end, and the continuum represents intermediate cell states between the early and late stages. If the bivariate dot plot is now examined, which correlates feature A to B, the progression of C to CA to CAB (bottom picture of sequence) can be deduced that would be consistent with the data. Correlation by itself does not infer progression order, but when a directional relationship is determined for one or more features by experimentation, progression order can be inferred by correlation to these features. It should be noted that many of the selection criteria that are used to enrich for the populations of interest also must also be initially determined by experimentation. The above progression is consistent with all the data, but it is not the only progression that can fit the data. It's always possible that the progression can be more complicated. For example, it's possible that some of CA becomes CAB and some becomes another population that has yet to be defined (e.g. CAZ ??). Or, it is possible that CA becomes CAB and then it later becomes CA again. Don't go on until you fully understand the above points.

Progression Order by Extension
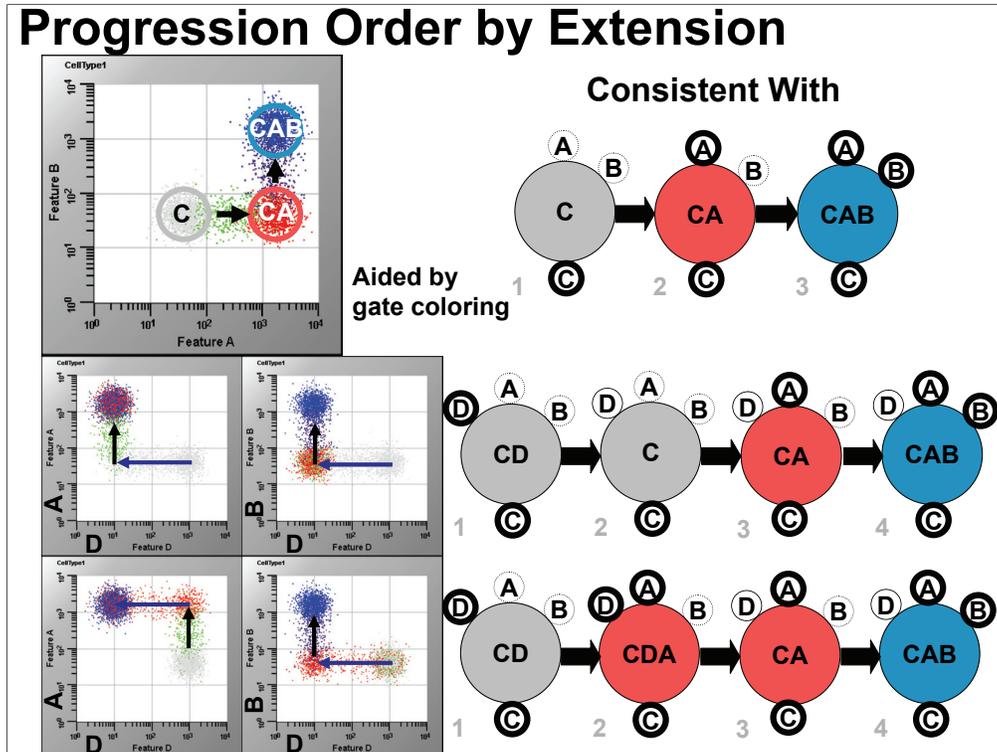
The summary of the progression defined by features A, B, and C and some *a priori* experimental data is shown at the top. In order to better see subsequent correlations, unique colors are normally assigned to the events in regions: C, CA, CAB, and perhaps the continuums between them. Suppose the data has an additional correlated feature named D. In order to determine how a change in D is temporarily related to A and B, the bivariate dot plots of D vs. A and D vs. B (middle-left panels) may need to be examined. The colors are helpful to properly interpret the data, but they are not required since the correlation is already determined in the dot plots. Both the D vs. A and D vs. B plots show that feature D is lost in the progression. Stop here for a moment and really look at those middle two histograms. Make sure you see the pattern that D is lost before both A and B are gained. The D vs. A plot shows that D is completely lost before A is gained which suggests the sequence to the right since we already knew that A was gained before B. In many ways, the interpretation of the correlated dot plots is a kind of puzzle that's fun to solve. If by chance, the D vs. A and D vs. B dot plots on the bottom row were observed, the progression that would be consistent with the data is shown to the right. In this case, feature A is up-regulated before D is lost but D is lost before B is gained. As usual, the caveats about the progression possibly being more complicated than it appears applies. Make sure you understand this logic before going on.

Mental Progressions
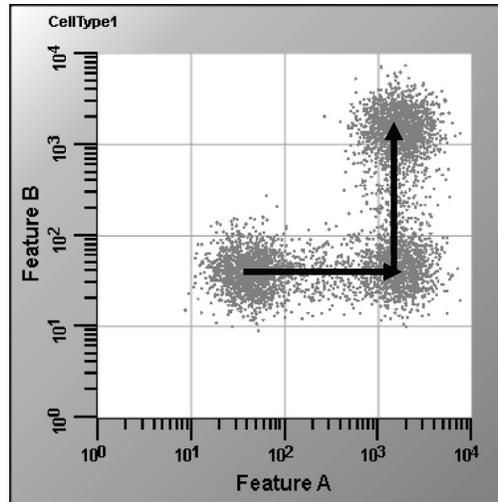
Because cytometry makes measurements on a per cell basis, all measurements are correlated. This correlation allows us to deduce or hypothesize the steps in cellular progressions. What is amazing about this is that most experienced cytometrists don't even know they are playing this game. When they look at a progression such as shown above, their mind connects the dots with imaginary arrows if they know that A is up-regulated. If this progression is new to them, all they see are three clusters connected by continuums. So, given this discussion, let's play the Cytometry Game. We'll start slowly and work up to a frenzy.
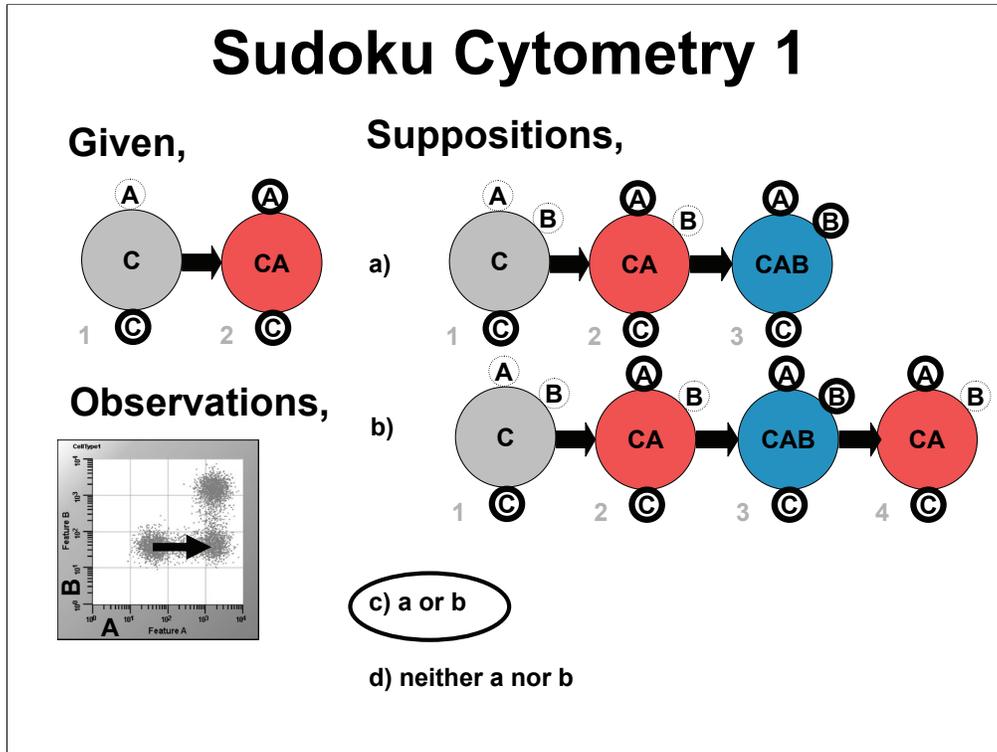
Suppose we know that A is low first and then it is high for cells expressing C (upper-left). We observe the pattern shown in the lower-left. I've put a black arrow on what we know to be true. Figure out which of the following statements are true: a) this sequence is consistent with the data; b) this more complicated sequence is consistent with the data; c) a or b; and d) neither a nor b. If you have been paying attention to the preceding slides, you should know that it is 'c'. 'a' is probably the most probable progression but it is always possible that the progression can double-back on itself as 'b' does. If the progression does in fact double-back, it will become obvious as more correlated measurements are investigated. The rule of thumb is to assume the simplest progression until you need to make it more complicated to be consistent with all the data. As will be discussed in the laboratory, this is known as Occam's Razor and ends up being an important way we can find simple progressions that are consistent with the data. OK, that was a review and it was easy. Let's try something a little harder.
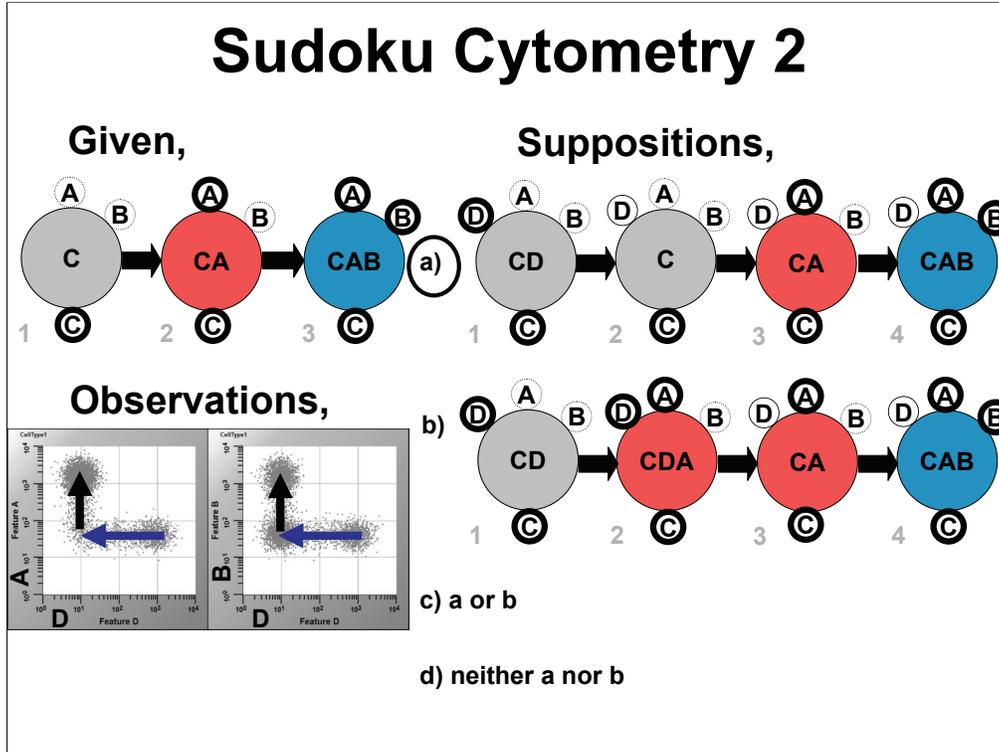
From the preceding logic, we know that the progression is C->CA->CAB. We observe the lower-left patterns of two dot plots: D vs. A and D vs. B. Again, I've put black arrows on what we know and a blue arrows on our conjectures.   As before, select the choice that is true. You should have chosen 'a'. If you didn't, go back and make sure you understand the preceding slides.
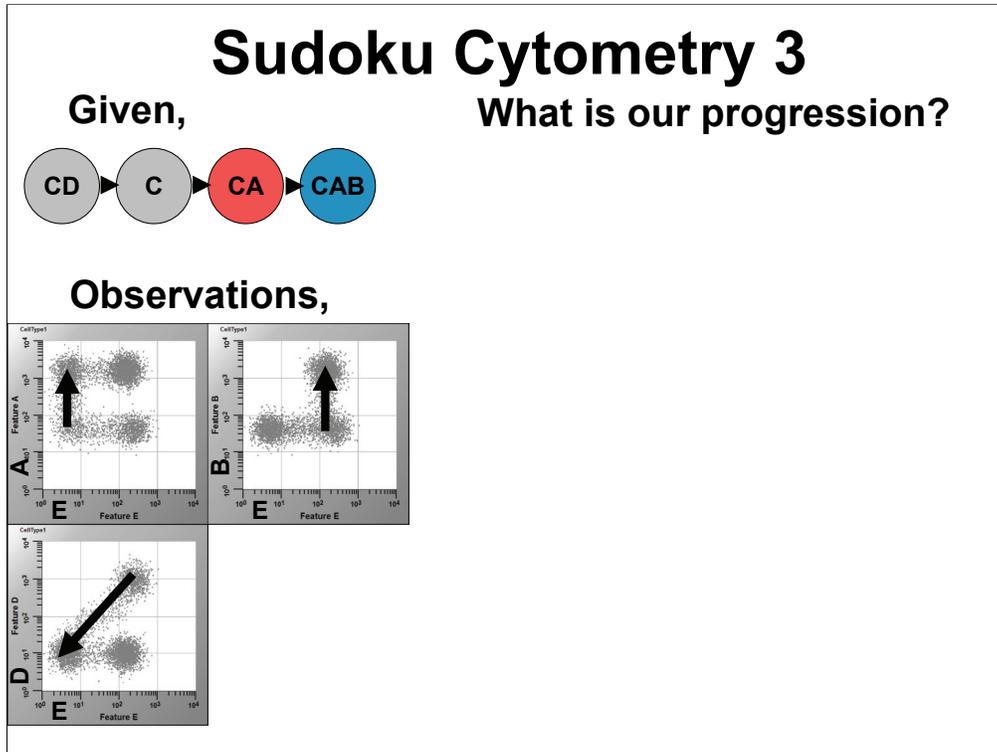
We now know that the progression most likely CD->C->CA->CAB. We observe the patterns shown in the lower-left dot plots when we look at E vs. A, E vs. B, and E vs. D. The black arrows represent what we know. Now for the game, figure out the progression. Go ahead and give it a try. The answer and explanation is on the next slide.
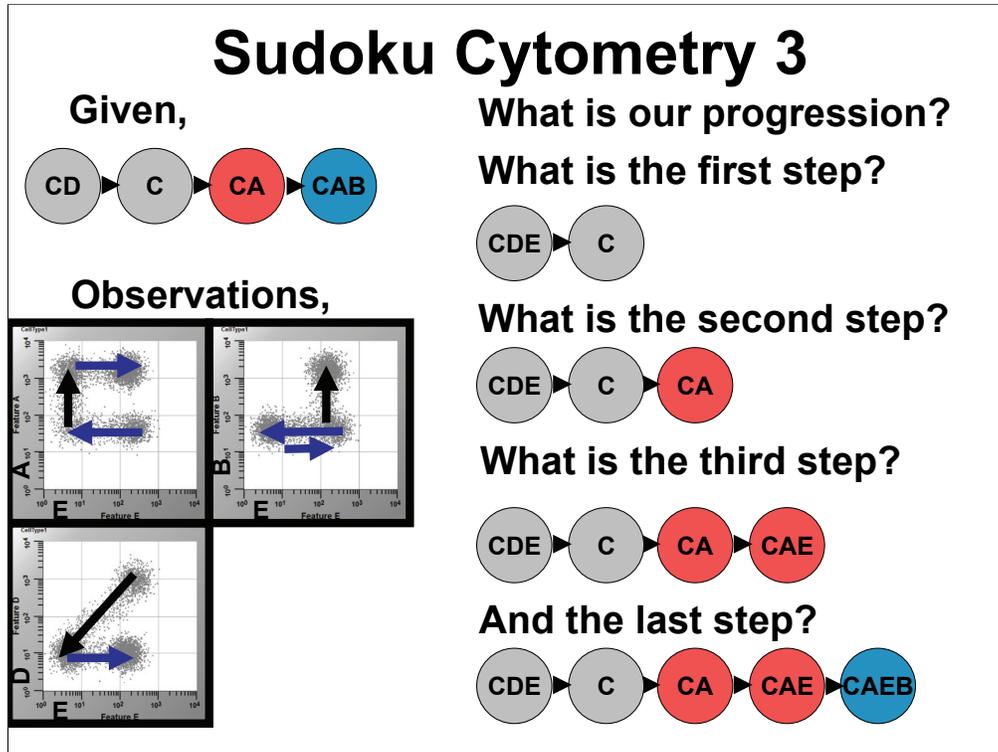
Let's tackle the first dot plot of E vs. A. We know that A is up-regulated so the black arrow represents what we know. If that is true, then E must down-regulate before A up-regulates and then up-regulate later (blue arrows). Notice that the head of one arrow meets the tail of another in a progression. How about the next plot of E vs. B? Given what we just found out, E must down-regulate and then up-regulate before B is up-regulated. The last plot, E vs. D is tricky. We know that D down-regulates so the black arrow represents the known progression. Why is it on a diagonal? When you see the diagonal pattern it generally means that both markers are changing together or are co-regulated. At some time later, E must up-regulate. If we take all this information, what is the first step in the progression? It must be CDE->C. The third plot gives us this information. What is the second step? It must be C->CA because of the first plot. What is the third step? It has to be CA->CAE because the second plot tells us that E is up-regulated before B. Finally, what is the last step? It must be CAE->CAEB because of the E vs. B plot. As I promised you, cytometry is a game you can play to find out the steps in any biologic progressions. It's a lot more fun than Sudoku because your logic is uncovering a mystery of how cells do what they do. When you arrive at your solution, realize that this is not proof that the sequence is as shown. As mentioned earlier, it can always be more complicated. In order to prove this sequence, you would need to do more experimentation to make sure that this model predicts the correct sequence. This, of course, is the scientific method and how science progresses.

Early Identification of Bone Marrow Progressions

Loken MR, Shah VP, Dattilio KL, Civin CI (1987a): Flow cytometric analysis of human bone marrow. I. Normal erythroid development. Blood 69: 255-263.

Loken MR, Shah VP, Dattilio KL, Civin CI (1987b): Flow cytometric analysis of human bone marrow. II. Normal B lymphocyte development. Blood 70: 1316-1324.

Loken M, Terstappen I, Civin C, Fackler M (1992) Flow cytometric characterization of erythroid, lymphoid, and monomyeloid lineages in normal human bone marrow. In Laerum., Bjerksnes R (eds). Flow Cytometry in Hematology. San Diego: Academic Press, pp 31-42.

B-Cell Development
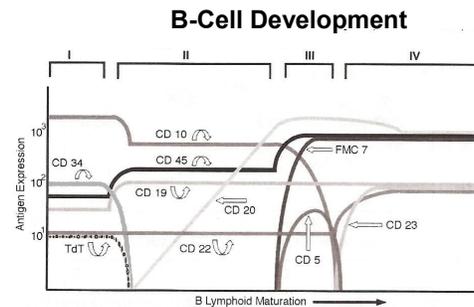
As soon as cytometry matured to where we could measure four simultaneous colors and light scatter, an interesting evolution began to occur in our understanding of normal bone marrow lineages. Michael Loken's three publications in the late 80's and early 90's represented a huge step in our understanding of the coordinated steps in the development of all these lineages. Probably the best text on this is still Carl Stewart's Immunophenotyping book. Mike realized that the end stage of these bone marrow progressions is typically found in peripheral blood and by knowing where these cells were in the bone marrow, he could obtain that initial important directional information that enables cytometry to deduce complex progression patterns. An example of months of work for the B-cell lineage is shown at the bottom of the slide. Once he worked out the B-cell lineage progression, he was stunned by the beautiful coordinated changes in markers as B-cells marched through their stages of development. Over time, this progression has been confirmed by numerous experiments. These lineage charts are routinely used by hematopathologists to understand normal dot plot patterns so they can find aberrant populations. This is how the cytometry game has been played for over thirty years. Let's now go back to our example, review what we have learned, and see if there is a better way to play this game.
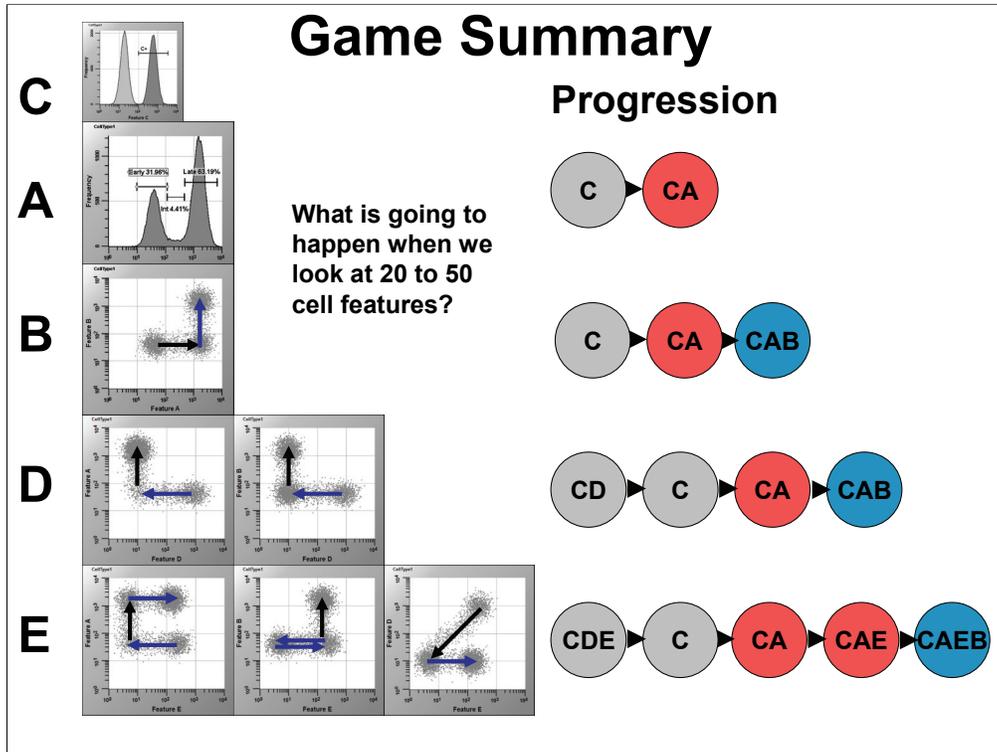
Our process of understanding the cellular progression defined by C, A, B, D, and E began with knowing that we could select the progression of interest with C, and that A was low early in the progression. Once we knew this information about C and A and we observed the A vs. B dot plot patterns, we could deduce how B adds to our progression. Once we knew this progression and observed the D vs. A and D vs. B dot plot patterns, we could deduce how D added to our progression. Once we knew this progression and observed the patterns of the dot plots of E vs. A, E vs. B, and E vs. D, we could deduce the final progression. There are a couple of salient points to make about this exercise. Notice that with remarkably little initial information, how much we can learn about a cellular progression with cytometry data. The bad news is that as the number of measured features increases, the number of necessary dot plot interpretations increases geometrically. Also, the logic to interpret these dot plot patterns becomes more difficult with more and more features. Imagine what would happen if we tried to play this game with 20 to 50 features. That was the past, now let's see how we will play this game in the future.
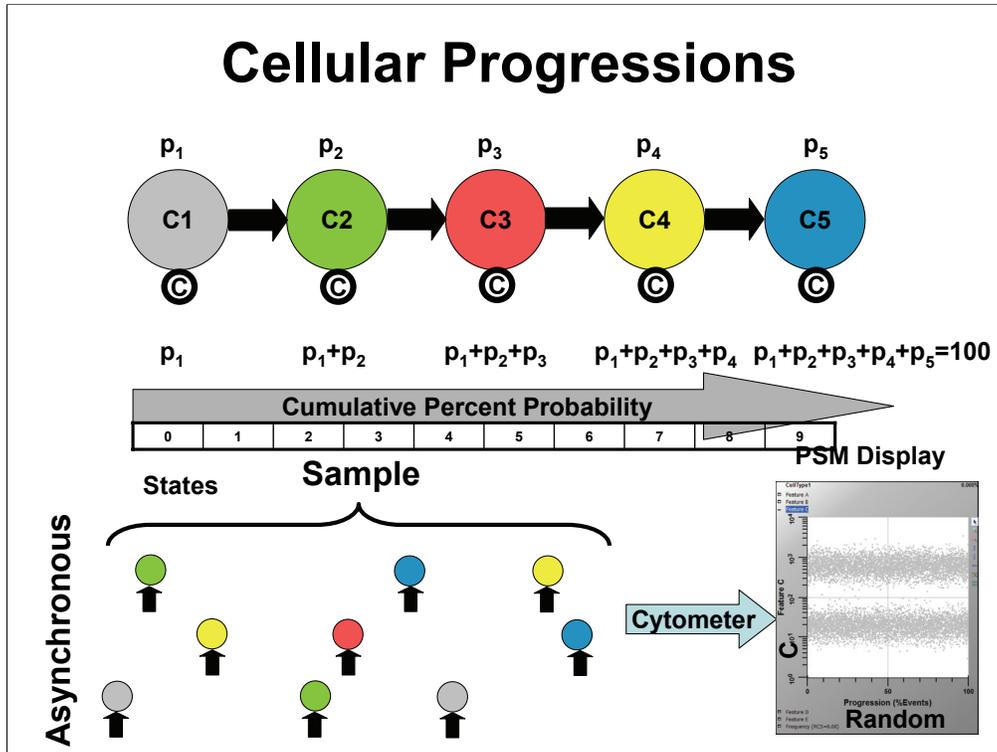
Let's go back to one of our earlier slides and imagine a world where we don't need to look at one measurement versus another. That's right, no dot plots. This new system is called Probability State Modeling. In this new world, we plot each measurement versus progression order. The units of progression order are in cumulative percent probability. The location of the dots along this axis represents where the cell or event is in the progression. Events to the far left are early in the progression and events to the far right are late. When we begin, the progression order is random since there is no information to stratify the events along this axis. We begin this analysis with C vs. Progression Order since C will be used to select the cells of interest.
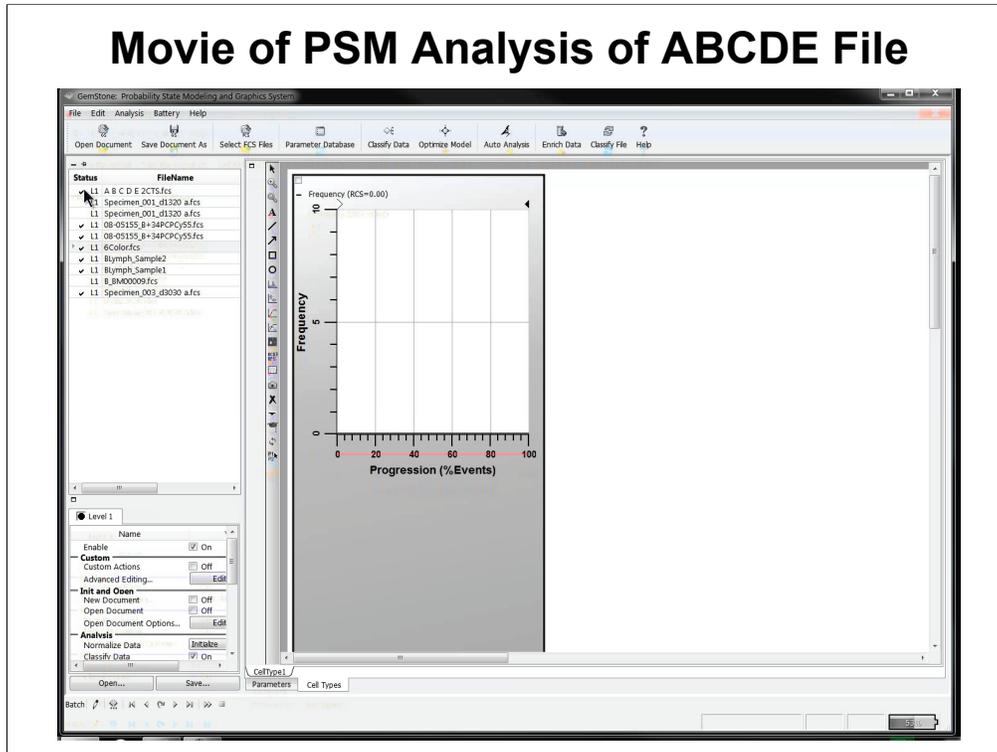
We first select for C+ events. The selection process is automatic and reproducible. Once we have selected the events of interest, C+ events in this example, we provide the initial directionality information into the system by fitting the A feature with a step-up parameter profile. At this point, the system can automatically find all the other feature progressions and create all the graphics that show the complete progression.
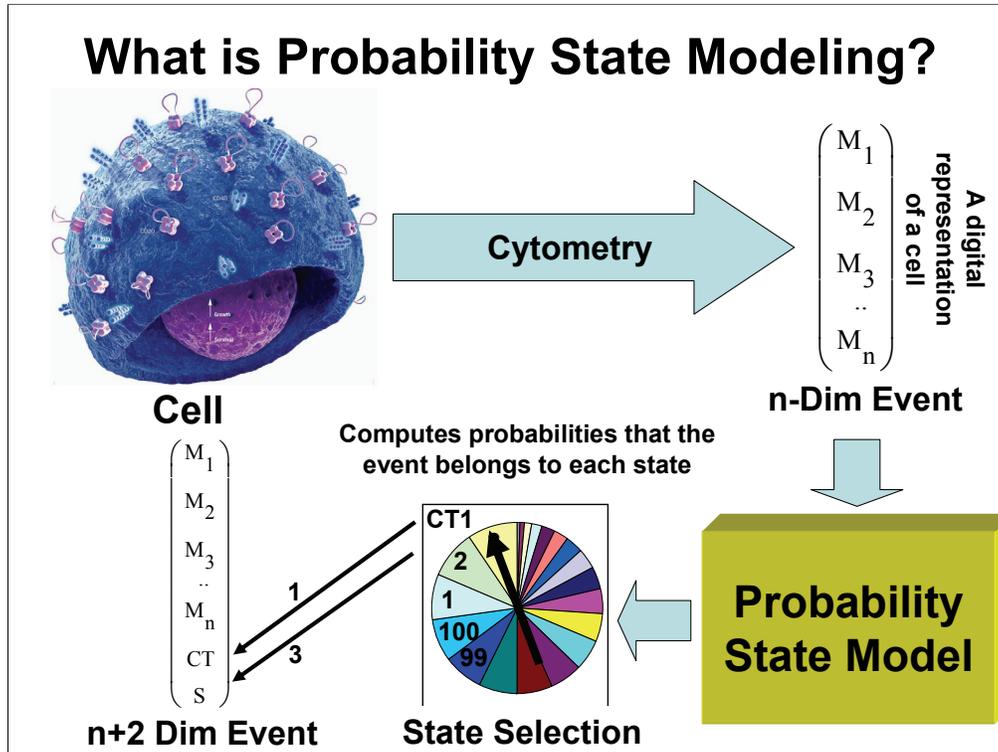
What is Probability State Modeling? The figures that follow are probably the simplest way of understanding what it is. We begin with a cell or particle. On or within this cell are a set of entities that we wish to measure (top-left). We use the technology of cytometry to make measurements of these entities, which forms a kind of digital signature of the measured cell (top-right). The number of measurements has been increasing rather rapidly in the last decade which creates problems for using conventional gating to analyze the data which is why this new method was invented in 2006. With Probability State Modeling (PSM), after we process the event, we end up with a set of probabilities that give the chance that the event belongs to a certain state within a progression or Cell Type. The system uses these probabilities to choose a specific state. This selection process is analogous to contestants on the Wheel-Of-Fortune show spinning the wheel with different size segments. In the example above, the method chooses state #3 with a process called stochastic selection. Stochastic is just another word for probabilistic. I'll explain how the system comes up with these probabilities in just a moment.

It's important to realize that this process can be extended to any number of Cell Types. Once the system has calculated the Cell Type (e.g. 1) and state (e.g. 3) it adds those numbers to the list of measurements for the event. This process essentially converts an n-dimensional problem to a n+2 dimensional problem. It turns out that by adding these two extra dimensions to our event, we can visualize and quantify any number of measurements with potentially no loss of information. As will be shown later, this method's accuracy is theoretically only bounded by counting error.

What is Probability State Modeling?
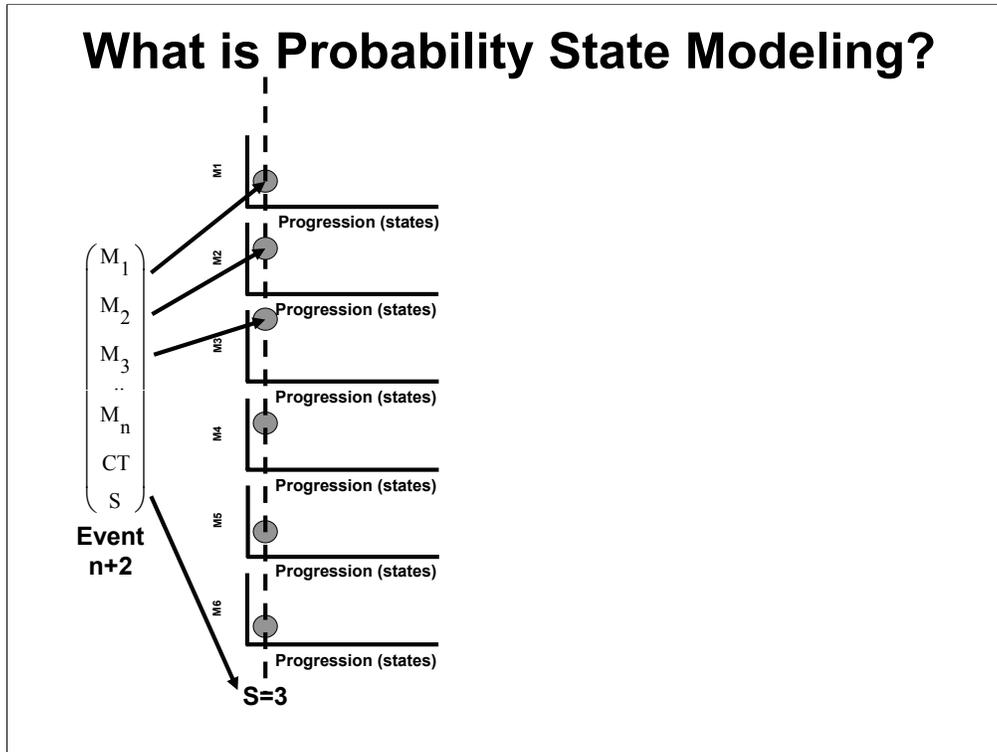
Let's assume for a moment that we have six correlated measurements for this event. What the system does next is to find the location of this event on six state plots as shown above. The vertical axis represents the intensity of the measurement and the horizontal axis is progression defined in units of cumulative %fraction. In this example, there are a total of 100 states along the progression axis so an event with state=3 determines the event's location on the progression axis (dotted line) and the individual measurements determine its positions along the vertical measurement axes. Thus, for one event we end up with six dots on the six state plots.

**What is Probability State Modeling?**

Many Events
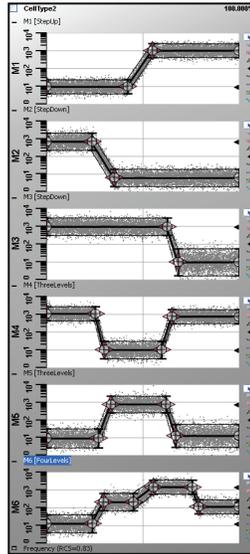
Multiple Parametric Plots
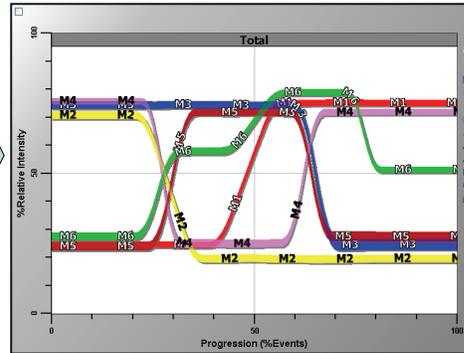
Single Overlay Plot

Shows most all the correlation between the measurements

After many events are processed in this manner, we end up with tracks of events in each state plot as shown above. Each one of these tracks has statistically determined 95% confidence limits for each state, forming a data envelop.   These data envelopes can be scaled-down and co-plotted on an overlay as shown to the right where the common vertical axis is %Relative Intensity and the common horizontal axis is Progression. This one graph summarizes all the correlations between the measurements as well as all the important stage statistics.
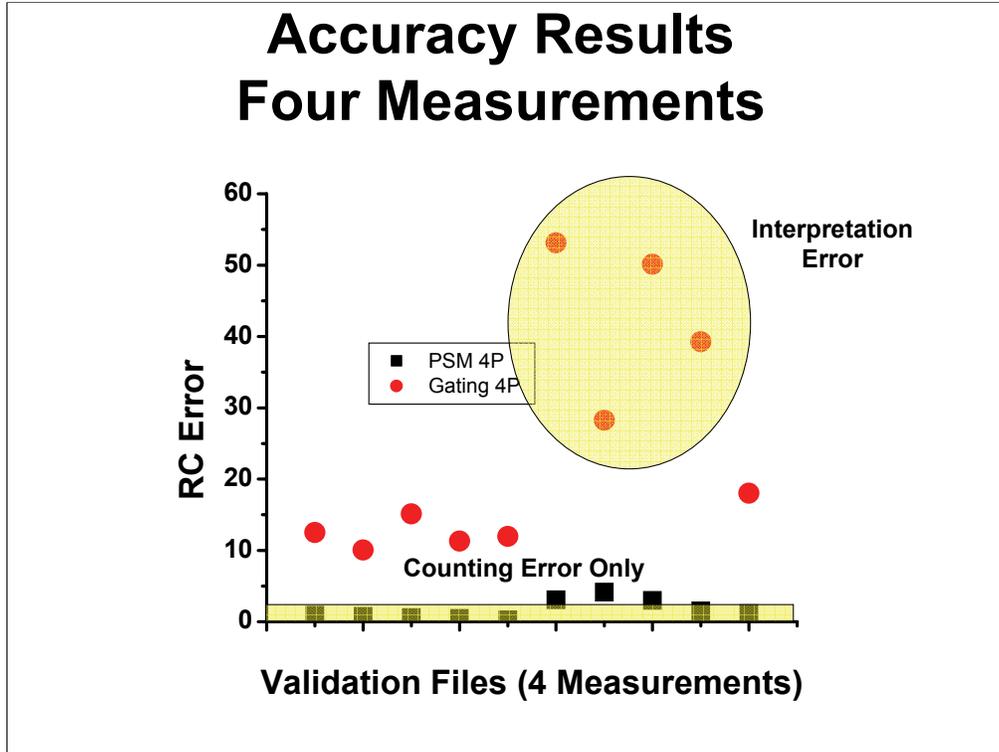
If we now examine the errors for four-measurements we again find that PSM yields estimates that are very close to the limit of accuracy while gating approaches have significantly higher errors.

Movie of PSM B-Cell Analysis

11 Parameter Data From BDIS LSRII, Brent Wood, Seatle WA
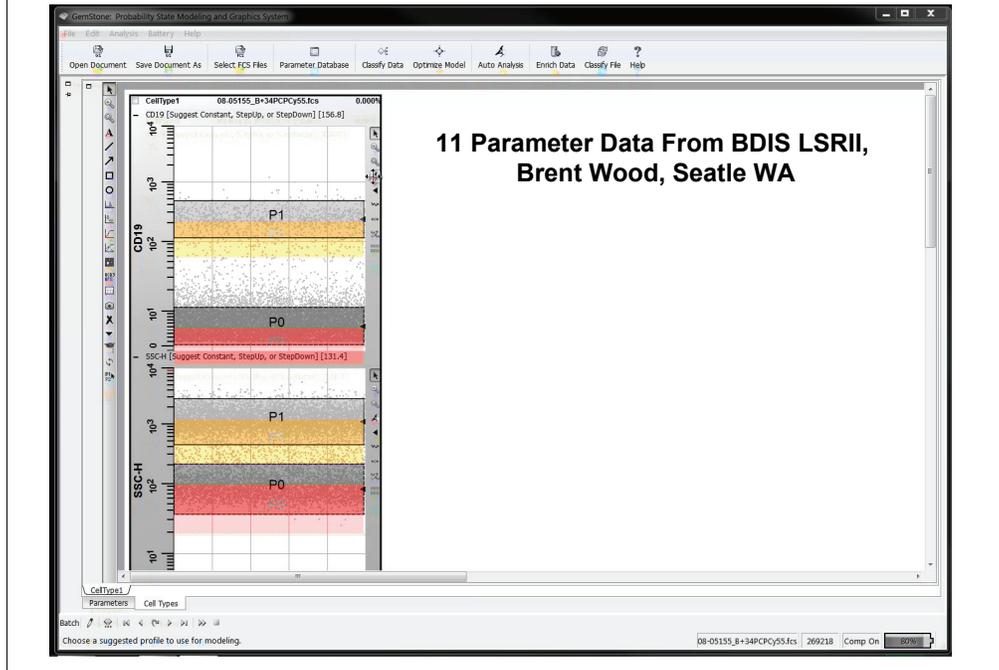
This is another movie showing an automatic analysis of some bone marrow B-cells. It really doesn't matter to the modeling system that the complexity of the sample has increased. Since it tackles each measurement separately, all that happens when we increase the number of correlated measurements is that the model becomes more refined.

Movie of PSM T-Cell Analysis

16 Parameter Data From BDIS LSRII, Margaret Inokuma, BD Biosciences
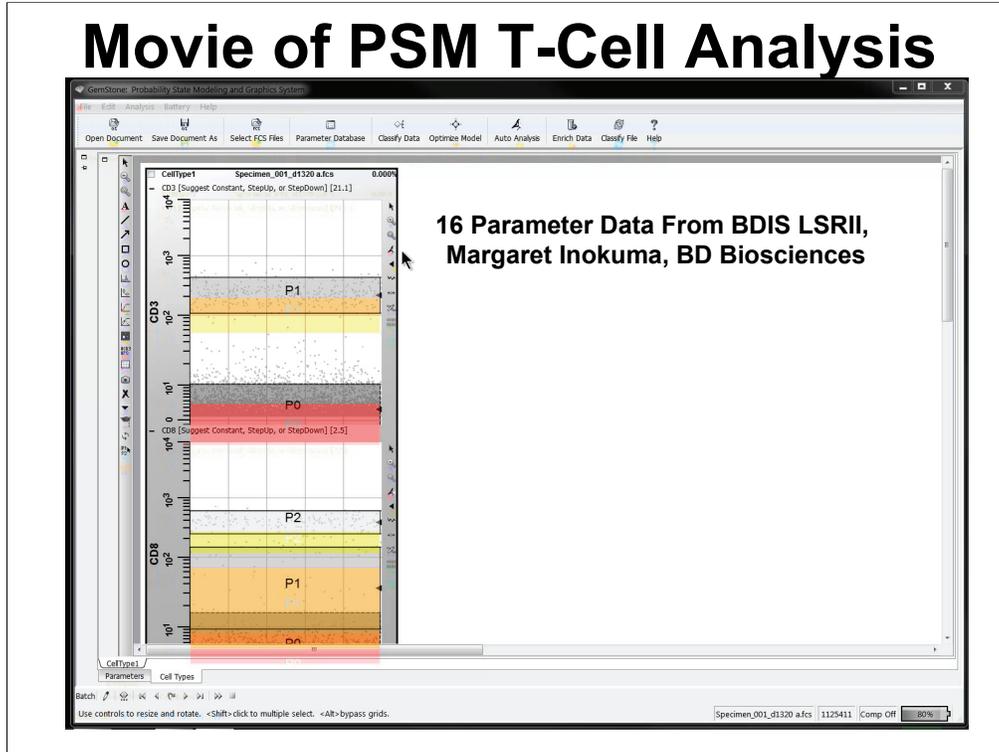
This is another example of PSM modeling. In this case we are modeling T-cells in 16-dimensional space. The number of bivariates that would show the same information would be 16*15/2 or 120. Data like this really shows that the old method of looking at high-dimensional data with bivariates just can't handle this kind of dimensionality.

Progression from Very High-Dimensional Data

Inokuma, M et al. Definition of CD8+ and CD4+ T-cell effector and memory subpopulations using GemStone, a novel modeling system. Abstract/Poster, Cyto 2010, Seattle WA.

7 16 Parameter Data Files From BDIS LSRII, Margaret Inokuma, BD Biosciences
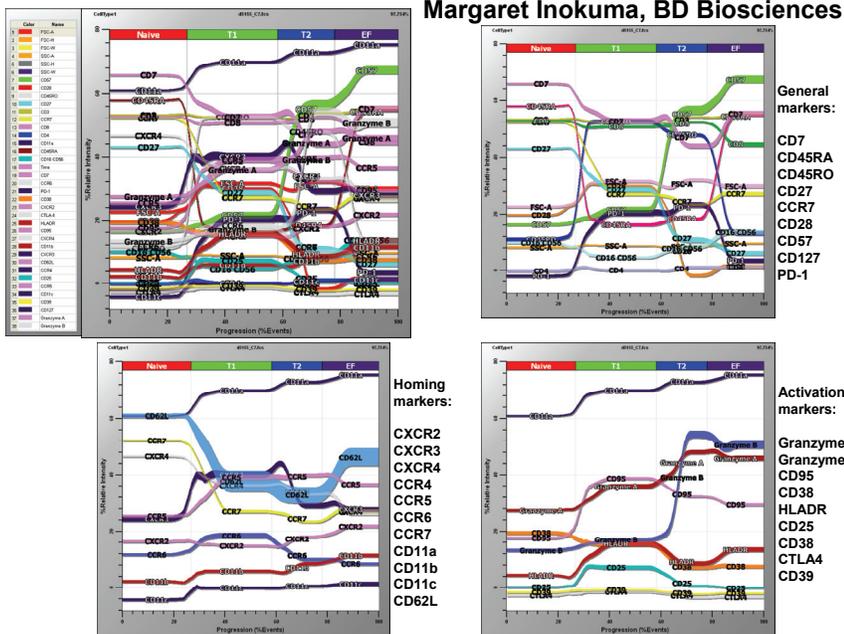
Because PSM involves one measurement at a time and the patterns become simpler with additional correlated measurements, there really is no limit to the number of measurements that can be investigated. Above is some recent work by Margaret Inokuma at BD Biosciences showing 38 correlated measurements for CD3+ CD8+ antigen-dependent progression. These 38 correlated measurements can be more easily appreciated by grouping them into three overlays representing general, homing, and activation marker changes. These data would need almost 700 dot plots to represent all the displayed correlations.

Summary

- **The asynchronous nature of cell samples and feature correlation on a per cell basis makes cytometry particularly suitable for discovering biological progressions.**

- **In order to deduce a progression, at least one feature's progression must be known and there must be one or more known selection features.**

- **Bivariate dot plot patterns allow the deduction of progression order but become more complicated with a larger number of cell features.**

- **Probability State Modeling (PSM) calculates progressions from any number of cellular features accurately and presents the feature correlations in an easy-to-understand overlay graph.**
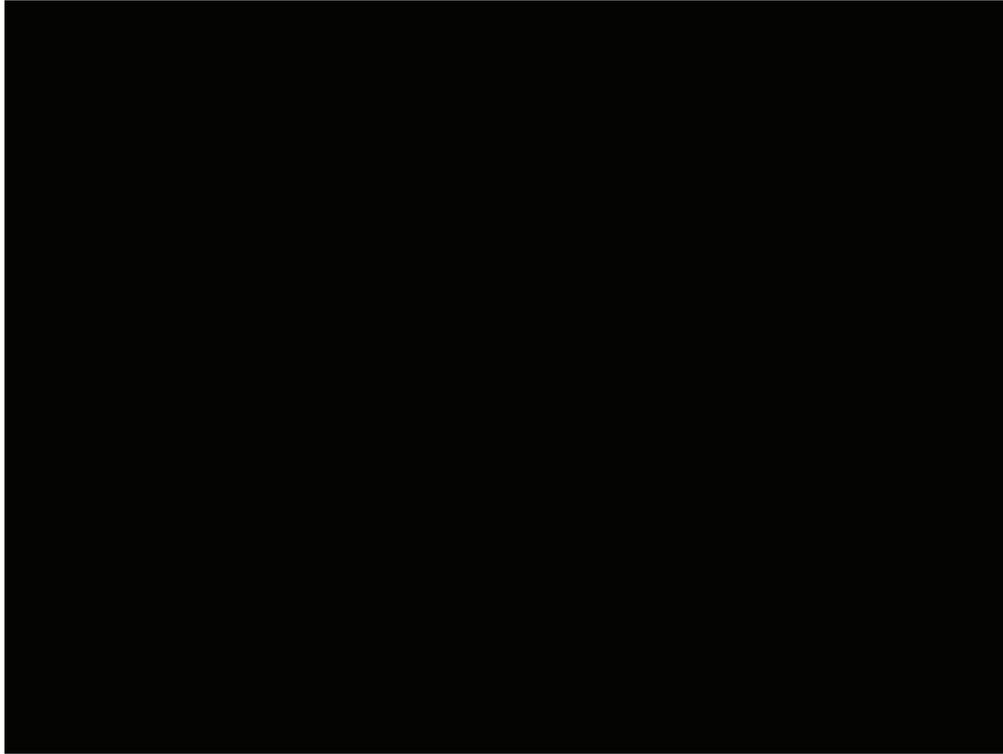
The Verity team made the creation of Probability State Modeling possible. Thanks guys!

What if there is no progression?