Imagine that you have received a call from a medical research scientist who has just developed the perfect implant to control diabetes.

It monitors blood sugar, it secretes insulin, it does everything necessary to let a diabetic patient live a life free from daily injections or pills.

The only problem is: due to certain constraints, the implant has to be made of a synthetic polymer, and if he were to place it directly into the body, it would most likely be rejected.

So, the researcher says to you, "Well, the body is made of cells and cells have lipid exteriors. I'm thinking that if I coat the polymer with a lipid membrane, the body may think the implant is an ordinary organ."

But, of course, to best disguise the implant he will need to understand the interactions between the polymer and lipid in order to engineer something that most resembles an organ.

So, he has called on you to point him toward the proper instrumentation and techniques for this study.

One of the properties he wants to study is diffusion, and for that you could tell him about confocal microscopy.

In confocal microscopy, molecules with a fluorescent label are placed throughout the sample to be studied.

Then a laser is focused into a very small volume or region of the sample.

Light emitted by the fluorescent tags is collected and focused with a second set of optics onto an avalanche photon detector.

Due to the finely controlled detection space, confocal microscopy is a particularly useful technique because it can detect interactions at the single molecular level.

The molecules forming a membrane diffuse randomly due to the ambient heat of the sample.
And, each time a tagged molecule enters the laser detection circle, a burst of light is measured.

For our study of methods to analyze confocal microscopy data, we wrote a simulation of surface diffusion so that we could easily control conditions such as diffusion, density, fluorescence intensity, and so forth in a reproducible manner.
You see here a mock up of the original simulation written in LabView.
As "molecules" symbolized by the green dots move into the detection circle, a burst of light is calculated based on the fluorescence of the species and is registered here.
Recently we rewrote the simulation in C, which allowed us to collect data at some 50-60 times faster.
The simulation time can be further reduced using parallelization. Twenty computer working together can reduce simulation time from 1.5 hours to 3 minutes, for 4000 sec of simulated time, 500ns dt.
This has given us much greater freedom in studying a wide variety of conditions.

Now, typically diffusion information is extracted from the real-time data via Fluorescence Correlation Spectroscopy.

The data are autocorrelated and fit to an equation such as this one, giving the diffusion constant, D, and the time-averaged number of molecules in the detection circle, N.

But, FCS is an averaging technique, so it doesn't always perform reliably under certain heterogeneous conditions.

These are data collected from a phospholipid mono-layer on a polystyrene substrate.

Most of the events are very brief and very bright; however, sometimes an event will be dim and last much longer.

These long events show that there has been some differentiation among the molecules: Some molecules are diffusing quickly across the surface, and others are diffusing slowly.

FCS has difficulty distinguishing the two, but rather averages them together.

So, we thought that since the difference between the species is apparent from the length of the molecule crossing time, perhaps the relative number of sort events and long events might also allow us to distinguish the two differently diffusing species.

So, we began working with the Single Event Duration Histogram.

We set a threshold value and calculate the amount of time the light level spikes above the threshold.

Then, the lengths of each event, or molecule crossing, are histogrammed.

As you can see, there are many very short events and fewer longer events.

Since the SEDH is a discrete technique, it has a potential advantage over FCS in discerning information about heterogeneous systems.

And, qualitatively, we see clearly from the SEDH that two species are present.

The experimental SEDH looks like a combination of two differently diffusing species.

For short events, the curve resembles that of a quickly diffusing species (the green curve); and for long events, the curve resembles a slowly diffusing species (the blue curve).

Now all we need is a quantitative method for extracting diffusion information from the SEDH.

Here, briefly, we can see how Diffusion constant affects the shape of the SEDH.

For quickly diffusing species (the red curve), there are very few long events; and for a slowly diffusing species (the blue curve), there is a much higher probability that long events occur.

This makes sense because a slow molecule will take longer moving across the detection circle.

But, there are many other parameters that can affect the SEDH.

Some of them we can control: detection bin width, duration, laser spot size, background (known). These, then, can be factored out of our mathematical model.

This leaves diffusion constant, particle density, maximum fluorescence intensity, and the ratio of any multiple

components. These are the parameters *ideally* we would like our method to handle.

But this quickly becomes a very complex problem, and it may be necessary to scale back to predicting diffusion constant alone, relying on other data for educated guesses of the remaining parameters. Then we may be able gradually to add the extra complexity back into the model.

So far, we have begun working on three approaches to extracting the desired information: analytical geometry, multivariate analysis, and trial-and-error prediction.

Now, in working toward an analytical description of the SEDH, two factors become very important: binning and the Gaussian profile of the laser spot.

It's tempting to represent the detection circle as a mathematical circle.

But, because the particle exhibiting Brownian motion tends to "waver" in any given location, namely the edge of the circle, several analytical "events" (moving in and out of the mathematical circle) are coalesced together from the binning during data acquisition.

Furthermore, the laser spot is not in actuality a circle with a hard and well-defined boundary. Rather, the intensity of the laser decreases in a Gaussian fashion as the particle moves radially away from the center.

There is no definite line past which you can say that the particle no longer fluoresces. Instead the "line" that marks the beginning or end of an event depends on how quickly the particle is moving inward or outward.

Therefore, an analytical model for the SEDH must include more than a simple circle.

Our second approach was through multivariate analysis, which is frequently used in chemometrics to model complex systems, such as the IR Spectrometric detection of protein and moisture in wheat, e.g.

One of the staples of MVA is the Multiple Linear Regression, based on the simple line, $y=mx+b$.

But, instead of having one y and one x, there are many ys and many xes.

MLR uses matrix algebra to reduce this complexity into elegant and tidy equations.

For our purposes, then, each point of the SEDH becomes a y-variable and must have a linear response to the input parameters (diffusion constant, density, maximum fluorescence intensity).

But, MLR performs well only for well-behaved systems. If there is too much information in the response matrix or if there is a critical amount of noise, MLR fails miserably.

So, many times, Principal Component Regression is used instead. PCR uses Principal Component Analysis to reduce the noise and extra information in the dataset.

This is done by first computing the eigenvectors of the response (eigenvectors represent the direction of greatest variation in the data), projecting the response onto the most significant eigenvectors, then regressing the rotated response matrix with the input parameters using MLR.

An important question in PCR is thus how many eigenvectors should be used. This is usually determined with a Scree Plot.

These plots get their name from geology: Scree is related to the stony slope of a mountain side, which these plots normally resemble.

The amount of variation described by each eigenvector normally decreases greatly after the first few then levels off.

So, in the case shown here, one might decide to keep either three or four components.

Now, sometimes cross validation is used to determine the number of eigenvectors instead of using merely the amount of variation each eigenvector represents.

In cross validation, a training set is used to develop a model with one eigenvector, then the model is tested against some other known set of data.

Then the process is repeated with two eigenvectors, and three, and so on.

The error of the prediction for successively more eigenvectors should, like a scree plot, either level off or come to a minimum.

And, under some conditions this is what we observed. But, at other times, the percent error varied wildly with the number of eigenvectors.

Leave one out cross validation uses one set of training datasets from which one has been "left out" or reserved for the testing/prediction step.

LOOCV should yield approximately the same best number of eigenvectors no matter which sample is left out.

But, this was hardly the case.

These results led us to question how linearly the SEDH responds to the input parameters. And we may need to turn to non-linear forms of MVA.

Our third approach to analyzing the SEDH was Trial-and-Error prediction. This method is based of fitting data to an equation.

An experimental curve is compared with a curve determined by simulation.

Then the parameters of the "known" curve are adjusted in order to minimize the figure of comparison.

We have used two optimization methods so far: the simplex algorithm in multiple dimensions, and Brent's algorithm when we were considering only diffusion coefficient.

Now, unfortunately we have to contend with multiple minima in the merit function.

These minima result when different parameters cancel the effects of other parameters.

Here the black curve was generated with a density of one molecule per square micron. For the blue curve, the density was increased to two molecules per square micron, and the curve shifted to the right.

But, when the density remains the same and the maximum intensity is decreased, the curve shifts back to the left, farther then it started.

At some point, the curves "cross," and two different sets of parameters generate indistinguishable curves, as far as our figure of comparison is concerned.

These are undesired minima, since our trial-and-error algorithm can't discern which set of parameters truly represents the experimental conditions.

Furthermore, the trial-and-error method can get confused by random noise in the simulated trials.

We're measuring Brownian motion, which is a stochastic process, so at different times any given simulation may have a more or fewer events, even though the input parameters may remain the same.

When this occurs, there is a chance increase in the merit function, creating a false minimum into which the algorithm can be trapped.

So, where is this research headed?

- Take into account binning and Gaussian intensity profile
- GOA specifically designed for noisy and expensive functions (three minutes for one iteration)
- Smart merit functions, smoothing
- More sophisticated MVA
- Novel techniques: NN, GA used in pattern recognition