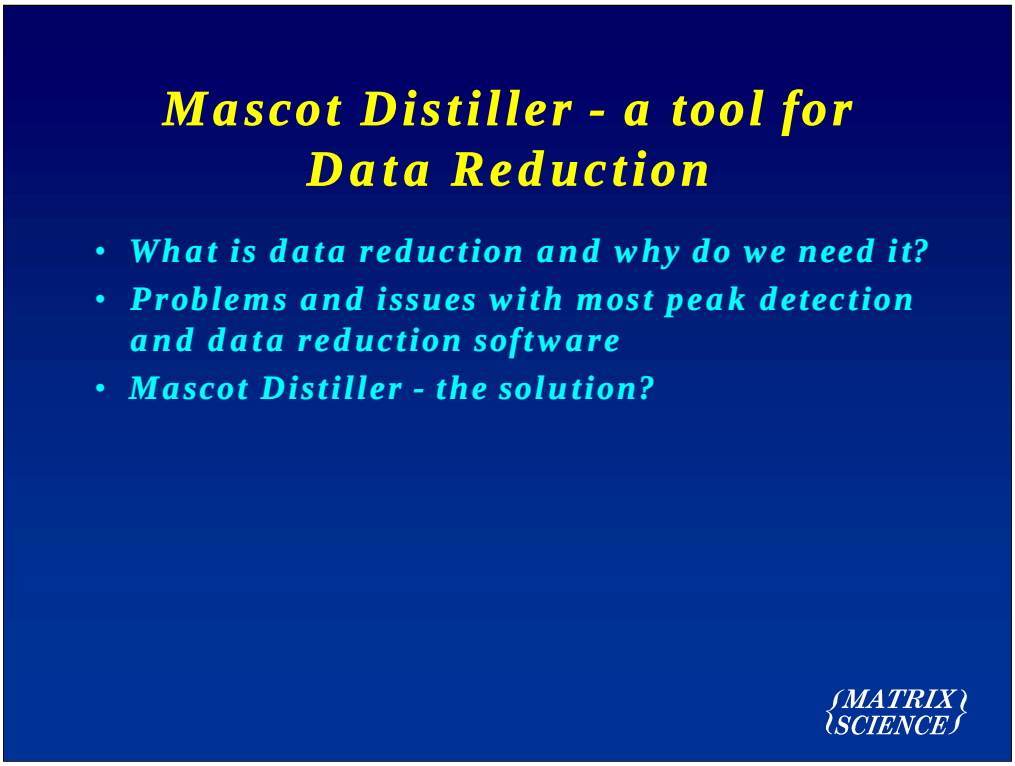
Mascot Distiller

a new tool for data reduction

MATRIX SCIENCE



For this final session, I would like to spend a little time describing a new product that we are developing - Mascot Distiller.

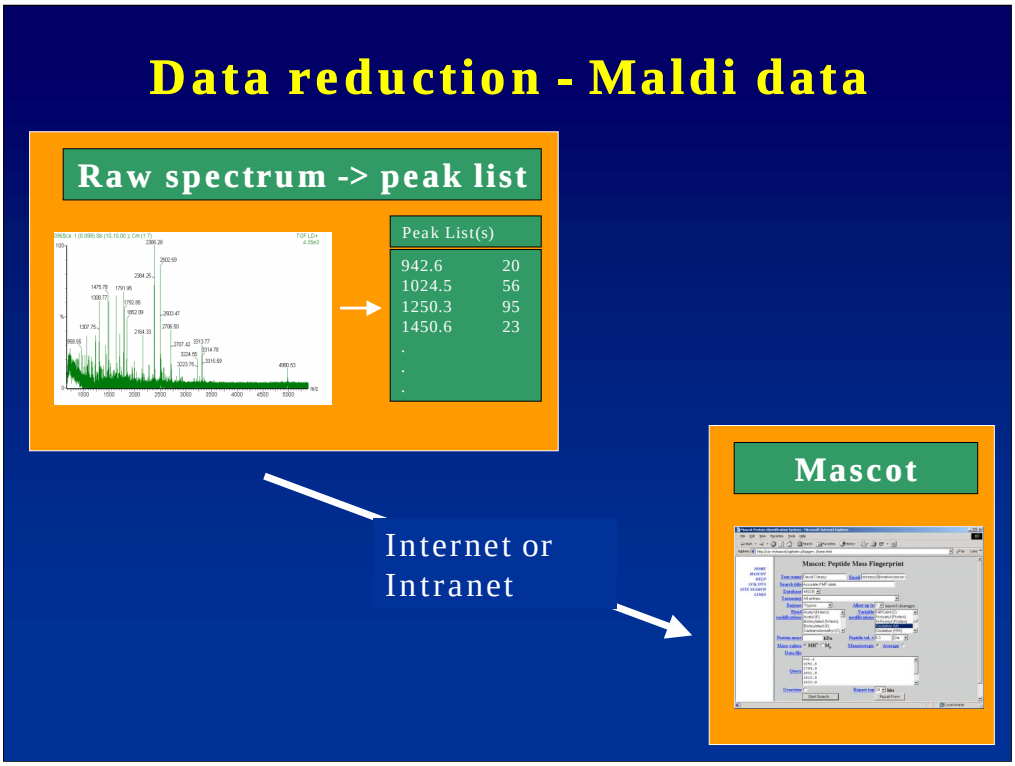
Mascot Distiller is used for 'data reduction' and in this session I want to describe:

what we mean by data reduction,

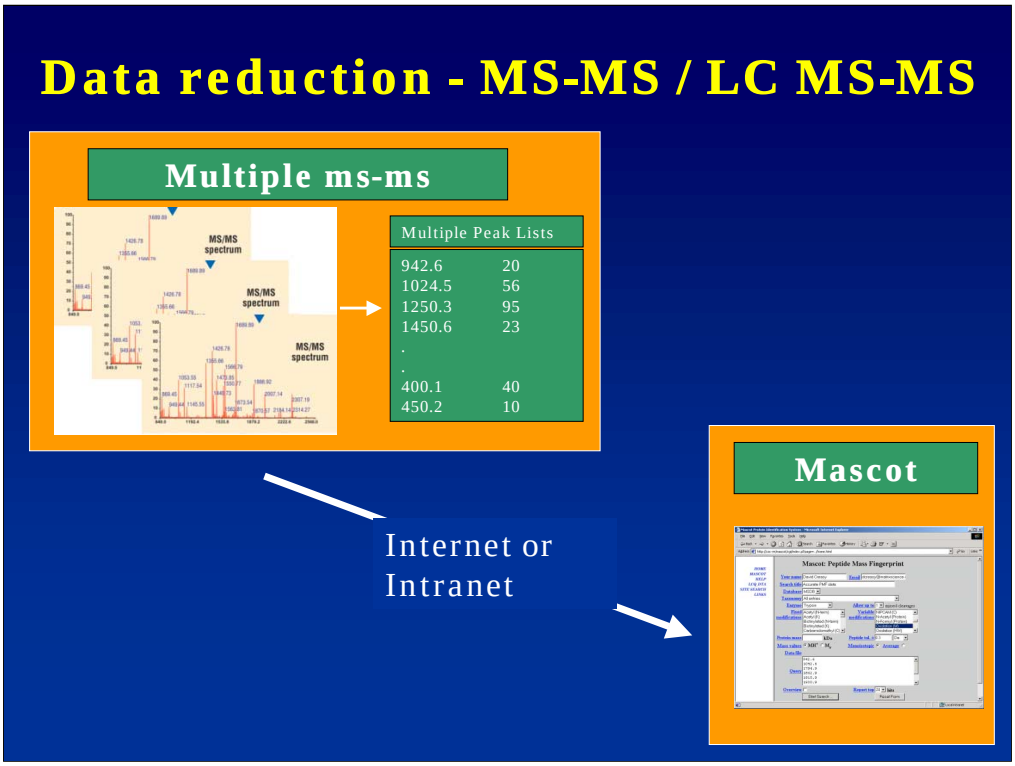
what are some of the issues with peak detection and data reduction software.

During the session , I will show some examples from Mascot Distiller.

So, first of all, what do we mean by data reduction



This is the simplest example of data reduction. This is a typical maldi spectrum which contains about 50,000 data points. We need to reduce this to somewhere between about 30 and 150 peaks. Unfortunately, this is not as trivial as it seems, and we are still seeing people perform this process manually for some spectra.



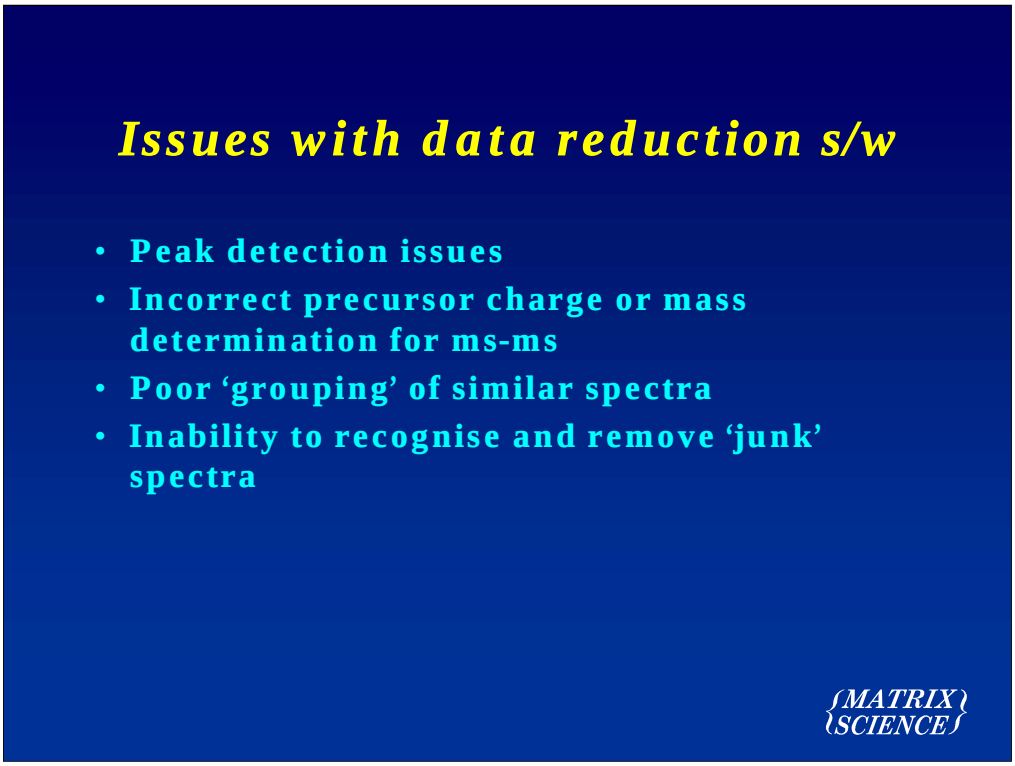
More complex examples are for LC MS-MS data or nano spray ms-ms data. Here we have four issues:

• Firstly, as with the ms data, we need to get a peak list from each spectrum

•Secondly, we need to understand the structure of the file. There is no value in detecting peaks in a zoom scan for example and sending this data to Mascot. Clearly the data from a zoom scan needs to be used for accurate precursor mass and charge determination

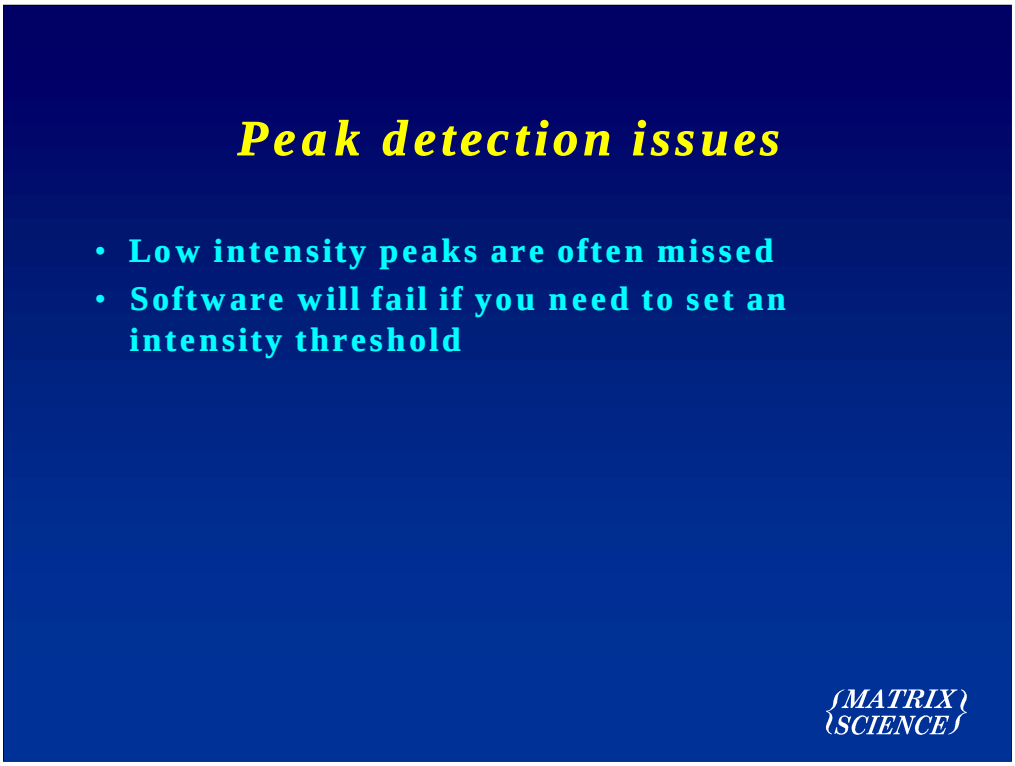
•Thirdly, we need to average together spectra from the same peptide

•Fourthly, in any lc ms-ms run, there will be many 'junk' spectra. Ideally, we should discard these at this stage rather than submitting them to Mascot. You would be amazed to see how many ms-ms spectra with a single peak get submitted to the Mascot public web server



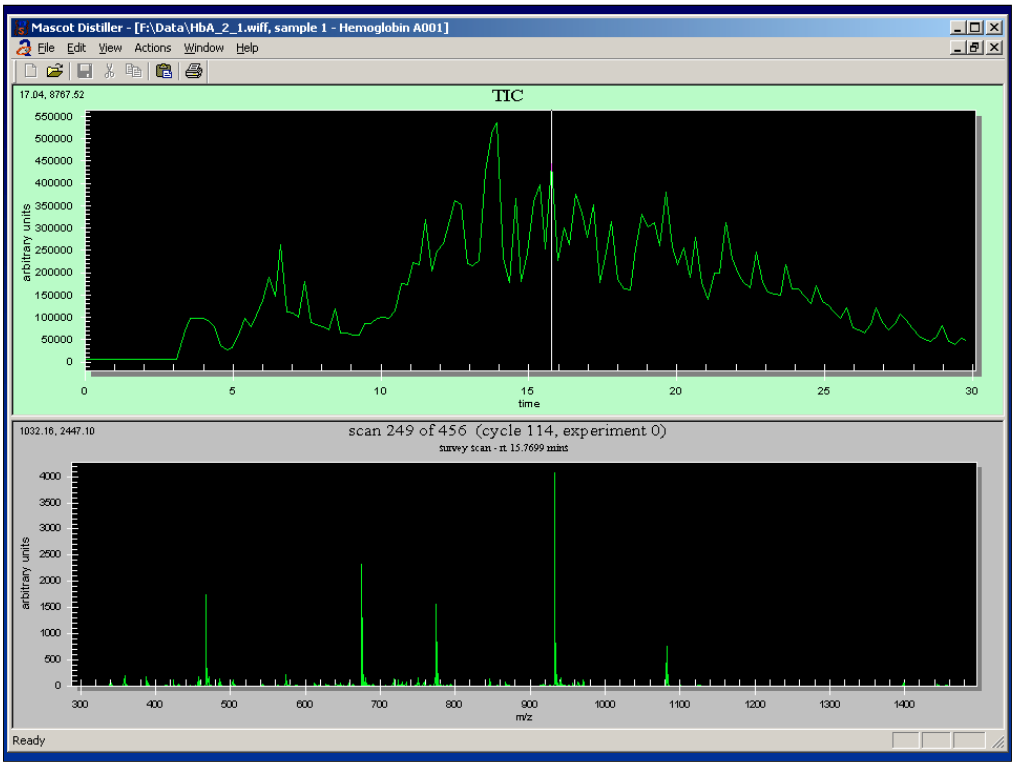
The issues with most data reduction software can be broken down as shown here

I will go into some detail on each of the issues, and describe how Mascot Distiller attempts to resolve them



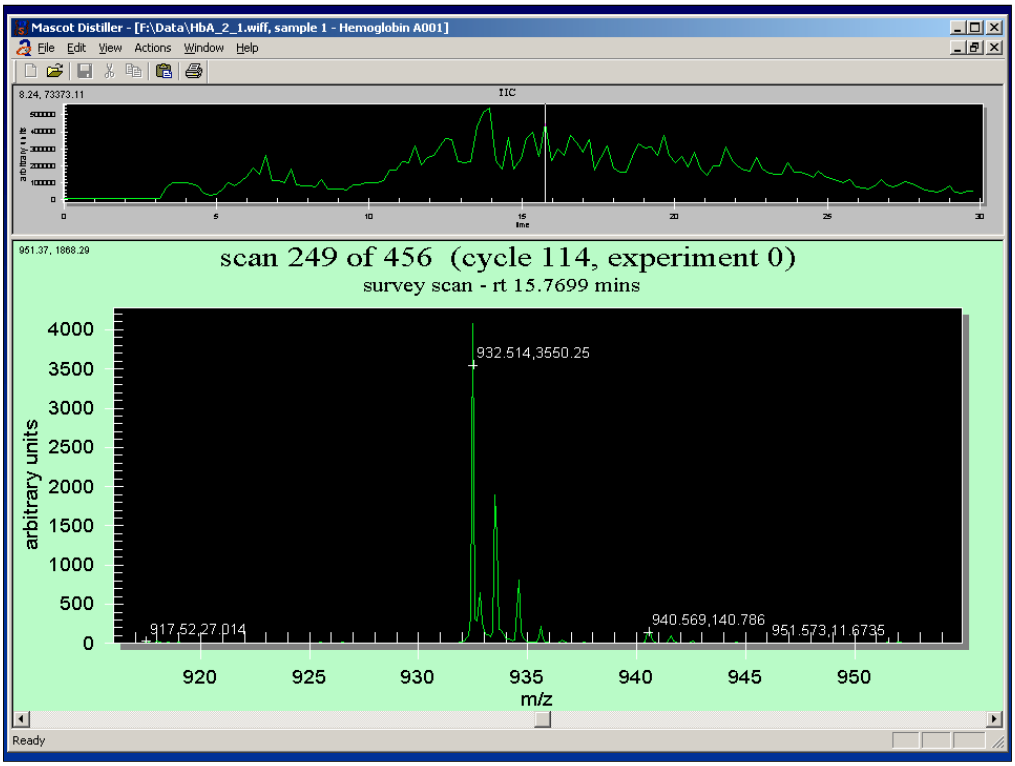
Firstly then, I'm going to describe some peak detection issues.

One key issue that we see is that high quality, but low intensity peaks are often missed. With many peak detection routines, you are required to set a threshold for peak detection. However, low intensity peaks are often real, and disregarding these will give a poorer match. I'll just show you an example:



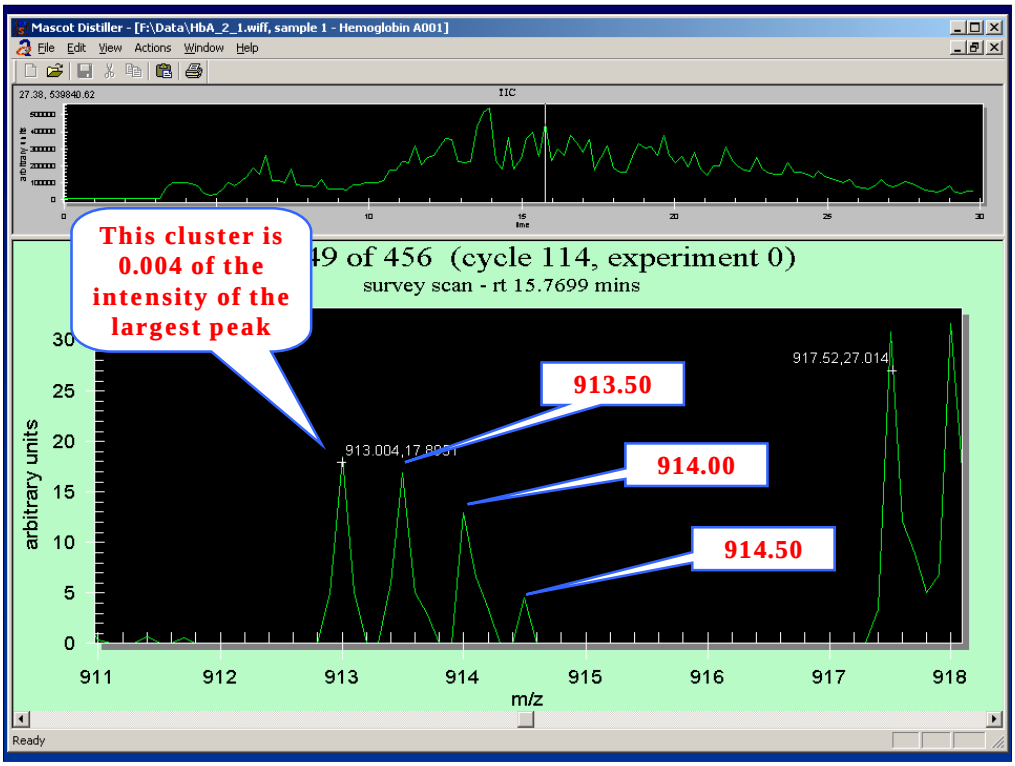
Here we can see some data from a Sciex QStar. This is very nice data, and from this survey scan we can clearly see some large peaks that any respectable software will detect.

If we just zoom in on this peak now



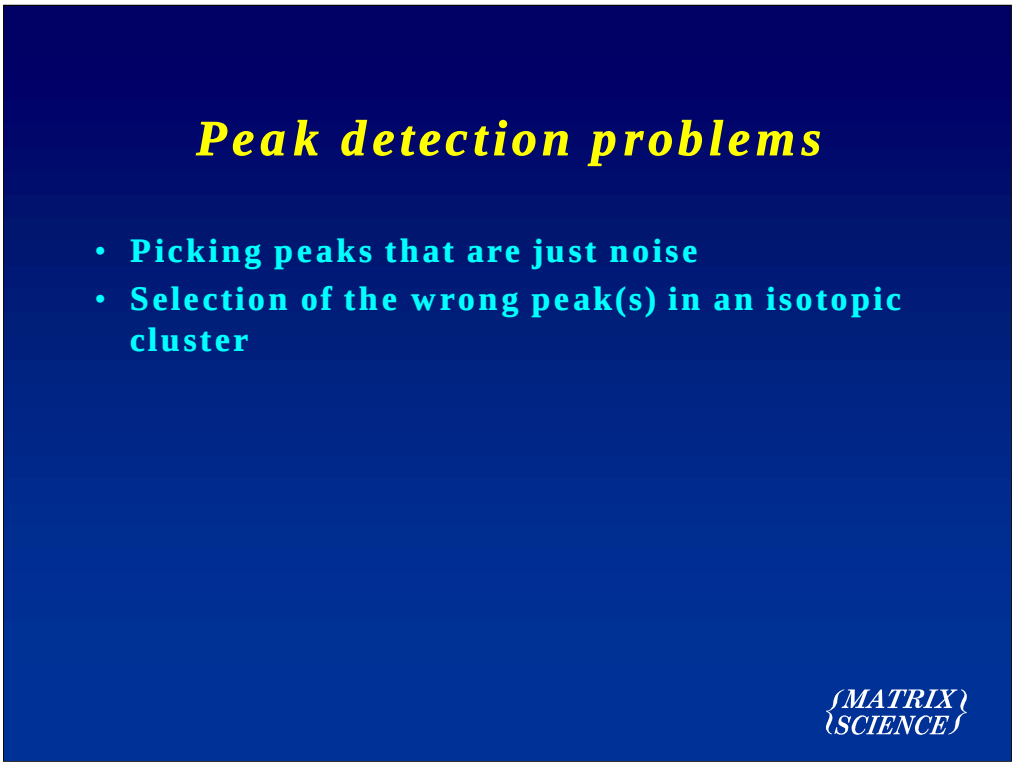
… we can see a nice looking isotopic cluster that any peak detection software would detect - although some of the poorer algorithms will unfortunately detect and label the carbon 13 and carbon 14 peaks as well. I would just like you to note that the intensity of this peak is about 4000.

Lets now zoom into the low intensity area down over here:



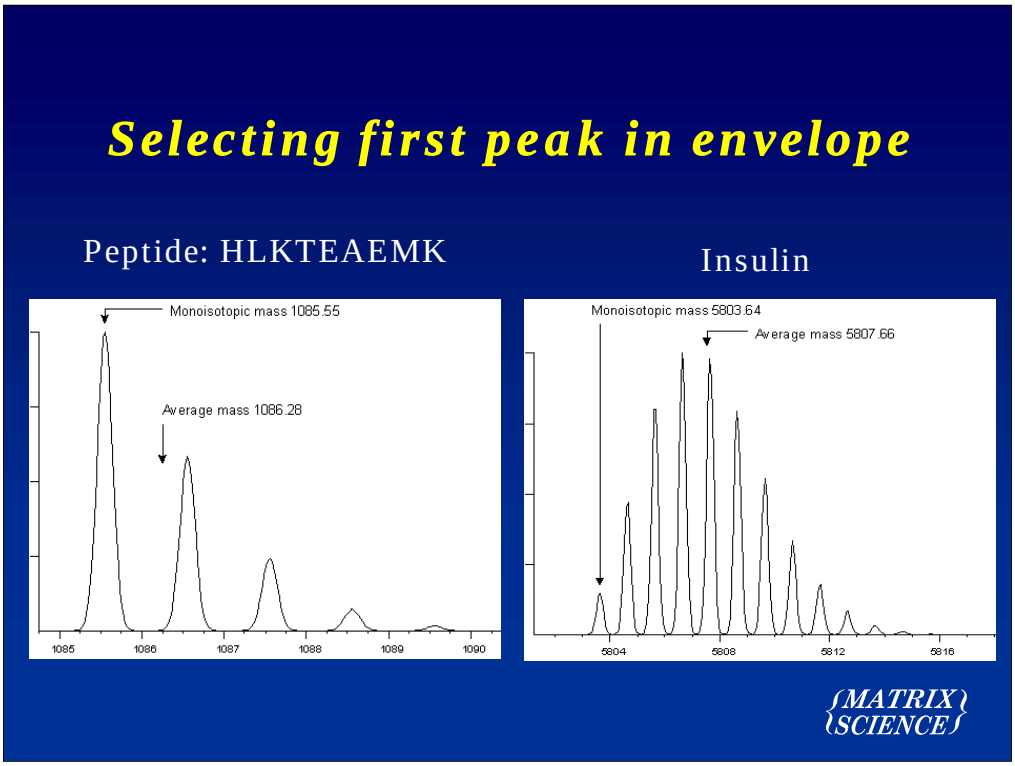
So, if we look carefully at this cluster, we can see that the peaks are exactly 0.5 Da apart. Also, note the shape of the cluster - it is almost an ideal shape - the individual peaks are not an ideal shape, but we shouldn't expect this at this low intensity. However, I am sure that you will agree with me that this is almost certainly a real peak from a peptide

However, look at the intensity level - this is about 20 - I.e. about 0.4% of the intensity of the most intense peak. If your software requires that you set an intensity threshold, then you are going to miss this peak cluster unless you set the threshold to a very low value.



Another classic problem with peak detection is that either the wrong peak or too many peaks are selected from an isotopic cluster.

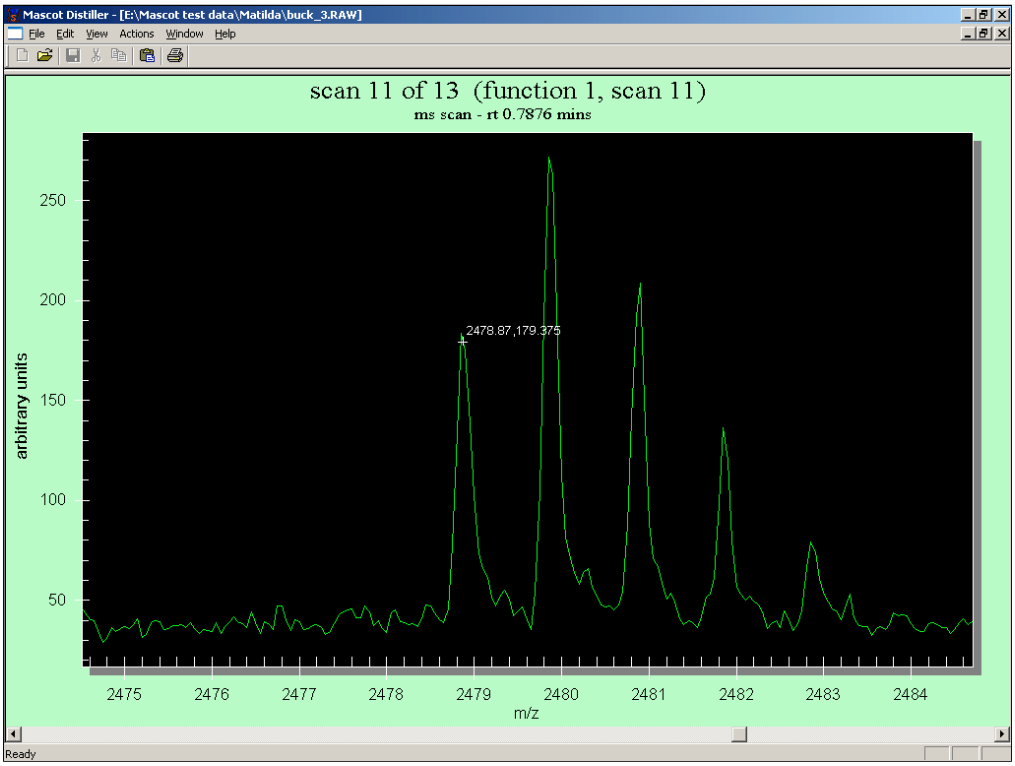
As I am sure that many of you are aware, the cluster shape changes depending on mass and resolution.



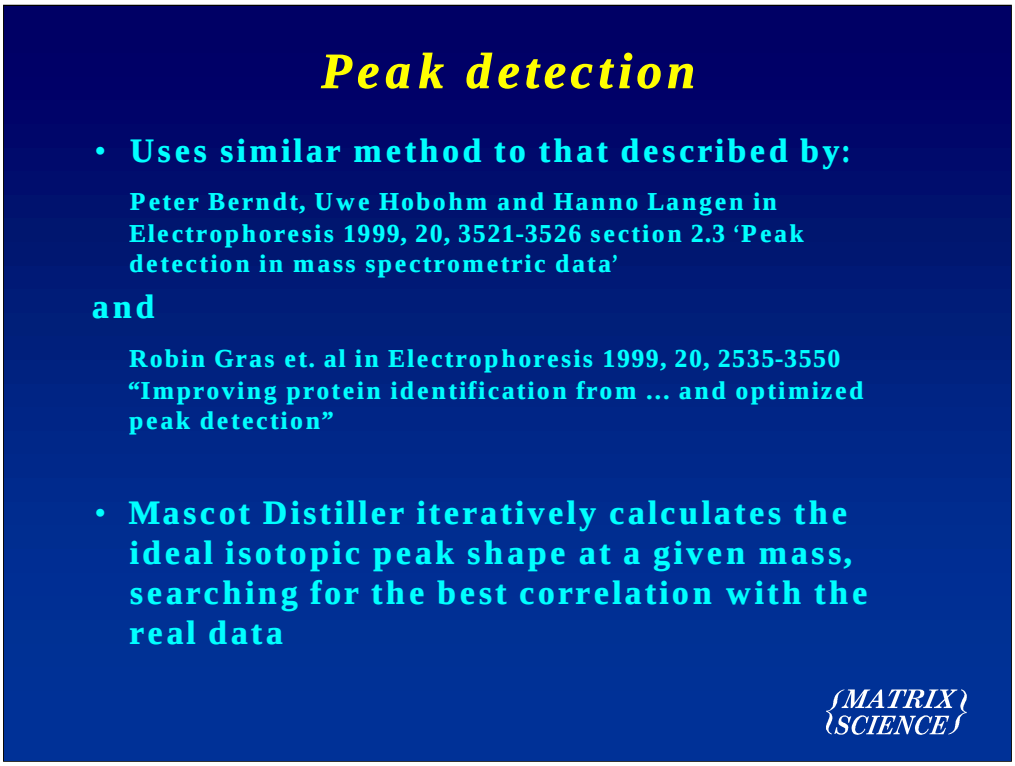
For a peptide of mass 1000 at a resolution of about 5000 we expect to see a peak cluster like this. The first peak is of course the monoisotopic peak, and at this mass and resolution will be the largest peak.

At mass 5000, the shape is very different, and it is going to be harder to pick out the first peak which is quite small.

Somewhere between these two masses...



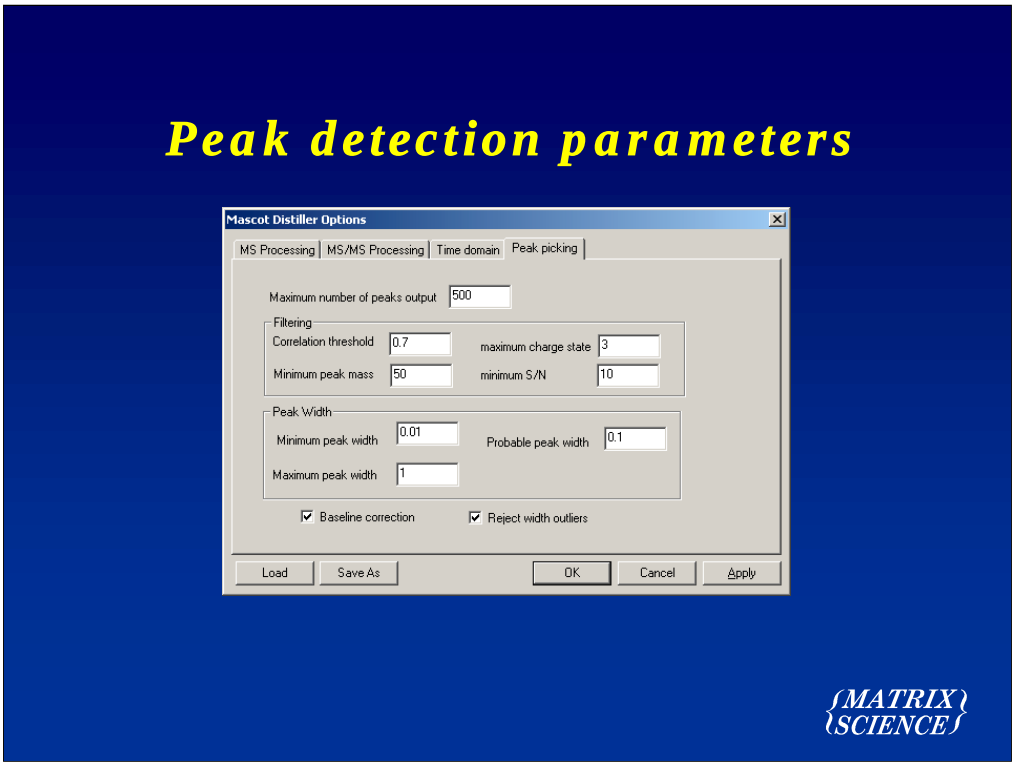
We have a cluster shaped like this, and you can see that Mascot Distiller has correctly identified the first peak in the cluster.



The method that we use to implement the peak detection is not new, but is based upon the work of some of other experts in this field. The technique has been described by Peter Berndt and colleagues at Hoffman La Roche, and also by Robin Gras, Ron Appel and others at the SIB in Geneva.

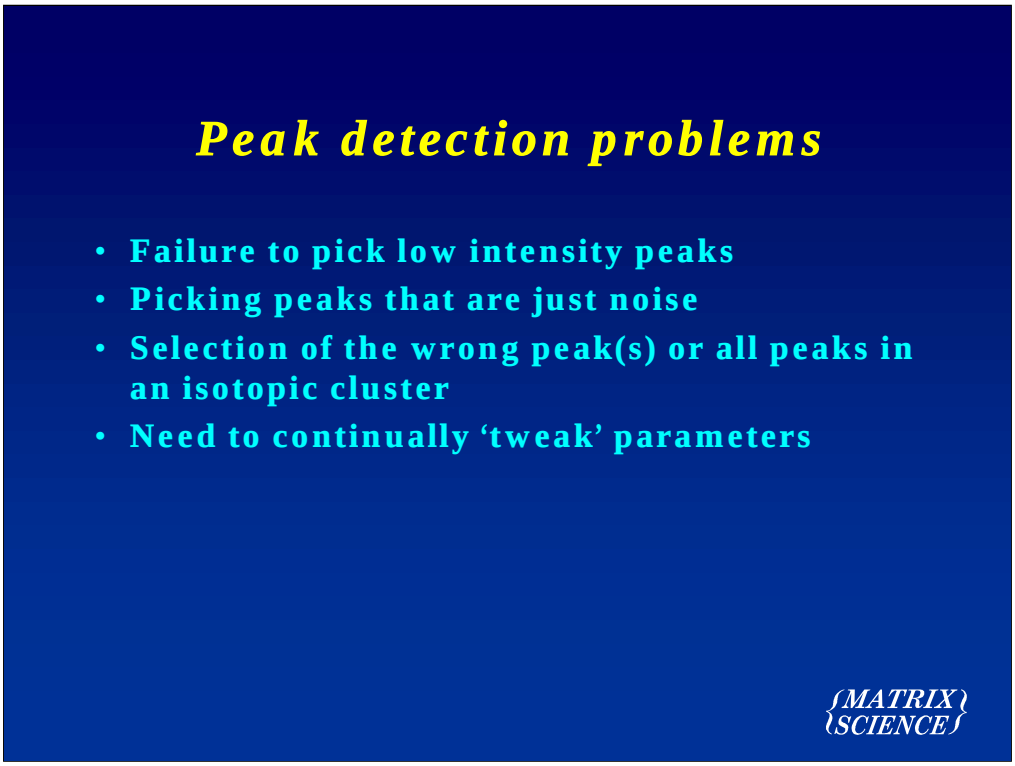
The technique is to calculate an isotopic distribution at a given mass and try and correlate this with a small section in the spectrum. If the correlation is good, then the calculated distribution is removed and the process is continued. This is an iterative process, and is computationally very intensive.

So what we have done is essentially nothing new, but we have implemented it in a robust and efficient manner, making the finished product available to many people.

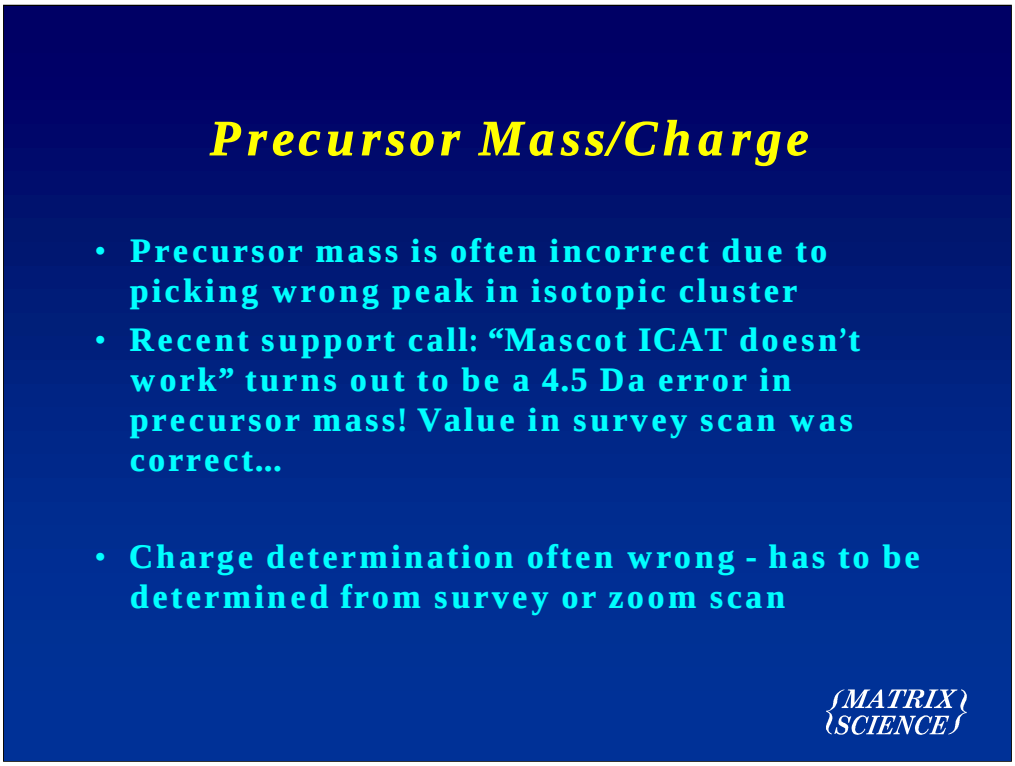


We have observed people performing peak detection using standard packages and listened to peoples feedback. The process that many people follow is to 'tweak' parameters using trial and error until they get the desired result. In most cases the users don't really understand what many or all of the parameters mean. To compound the problem, different instrument manufacturers use different techniques and different terms, making it almost impossible for a user of several different instruments to achieve optimal results.

Ideally, there would be no parameters at all, but this is simply not achievable. Our goal is that there should be a fixed set of parameters per instrument, that doesn't require any tweaking.



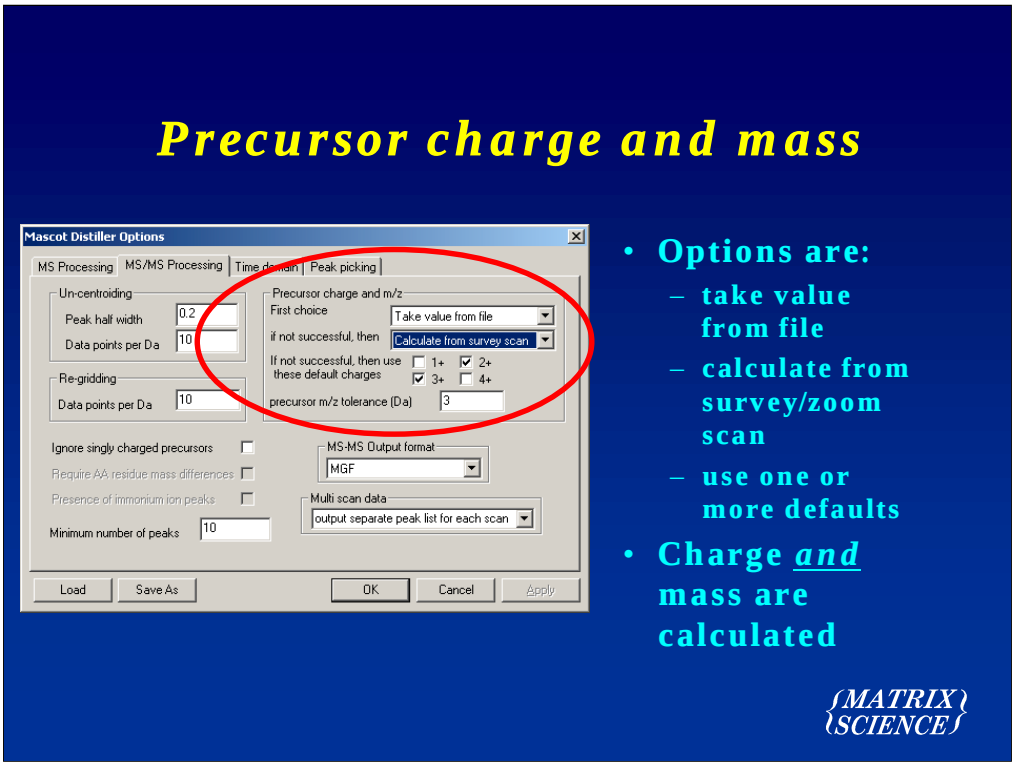
To summarise the peak detection issues, the problems with many packages are



The next big issue where things can go wrong with data reduction is failure to get the correct precursor mass and or charge for ms-ms dta.

A very recent example of this was when I had a support call from a customer who said that Mascot was not working properly with ICAT data. On further investigation it turned out that the precursor mass had an error of 4.5 Da. The person concerned checked the survey scan and it soon became clear that there was a problem in the instrument data system. The problem with these errors is that it is impossible to get a match at all if there is an error, and the error is often hard to track down. More typically we frequently see errors of 1.0 or 2.0 Daltons. Of course, it is easy in Mascot to cope with this - simply set the precursor tolerance to be 2.0 or 2.5 Daltons.

Incorrect charge determination is more of a problem, and generally results in a total failure to obtain a match.



There are three ways to determine the precursor mass and charge. Naturally, Mascot Distiller gives you all three choices, and most importantly allows you to decide on one method and if that fails try the next.

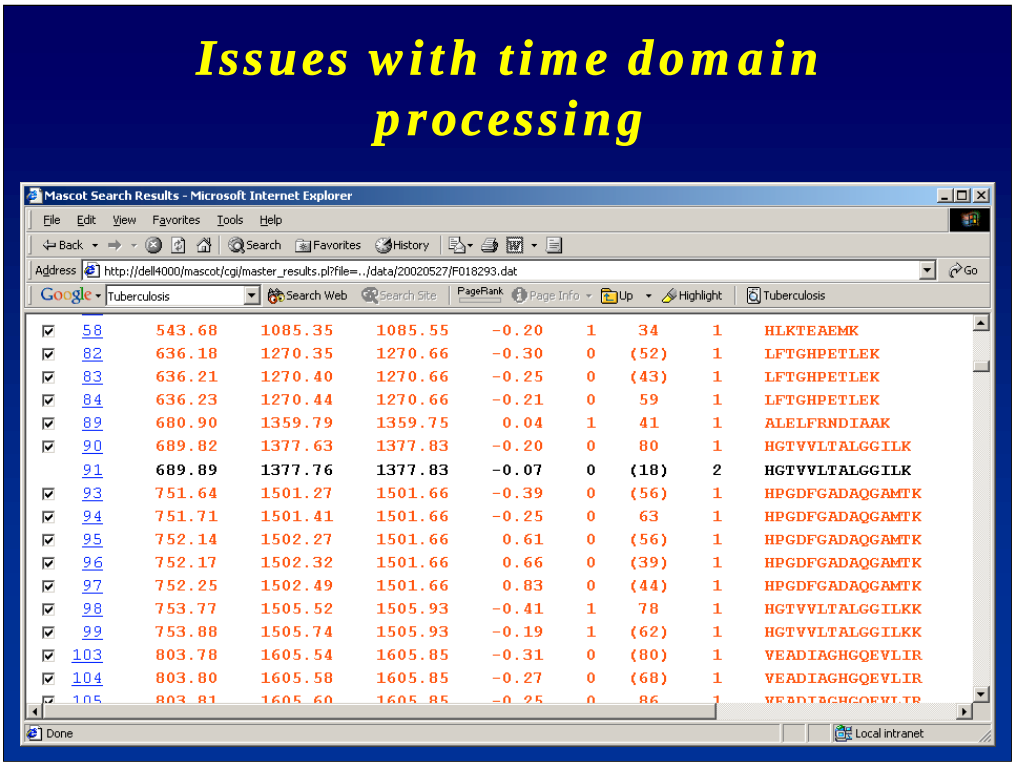
The choices are:

- take the value from the file. The mass value will normally be correct, although, as we have seen, there can be severe problems, and some values may be 1 or 2 Daltons out due to problems choosing the incorrect peak from an isotopic envelope. Furthermore, at least one system only saves the value as an integer.

- the second choice is to calculate the values from the survey or zoom scan. Mascot distiller uses it's peak detection routines to find a peak nearest to the one in the file, and also determines the charge. If the survey or zoom scan is poor quality, then this will obviously fail. Obviously if there is no survey or zoom scan in the file, then this will not succeed.

- the final fail safe is to try a number of different charge states.

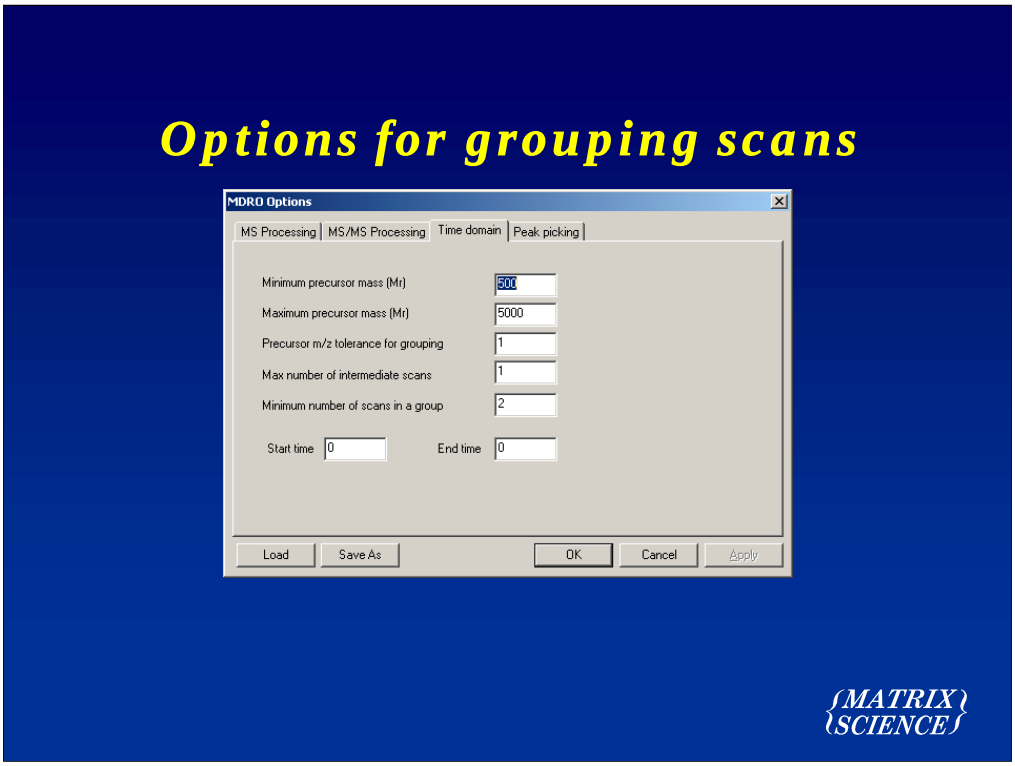
Issues with time domain processing



☑	58	543.68	1085.35	1085.55	-0.20	1	34	1	HLKTEAEMK
☑	82	636.18	1270.35	1270.66	-0.30	0	(52)	1	LFTGHPETLEK
☑	83	636.21	1270.40	1270.66	-0.25	0	(43)	1	LFTGHPETLEK
☑	84	636.23	1270.44	1270.66	-0.21	0	59	1	LFTGHPETLEK
☑	89	680.90	1359.79	1359.75	0.04	1	41	1	ALELFRNDIAAK
☑	90	689.82	1377.63	1377.83	-0.20	0	80	1	HGTVVLTALGGILK
	91	689.89	1377.76	1377.83	-0.07	0	(18)	2	HGTVVLTALGGILK
☑	93	751.64	1501.27	1501.66	-0.39	0	(56)	1	HPGDFGADAQGAMTK
☑	94	751.71	1501.41	1501.66	-0.25	0	63	1	HPGDFGADAQGAMTK
☑	95	752.14	1502.27	1501.66	0.61	0	(56)	1	HPGDFGADAQGAMTK
☑	96	752.17	1502.32	1501.66	0.66	0	(39)	1	HPGDFGADAQGAMTK
☑	97	752.25	1502.49	1501.66	0.83	0	(44)	1	HPGDFGADAQGAMTK
☑	98	753.77	1505.52	1505.93	-0.41	1	78	1	HGTVVLTALGGILKK
☑	99	753.88	1505.74	1505.93	-0.19	1	(62)	1	HGTVVLTALGGILKK
☑	103	803.78	1605.54	1605.85	-0.31	0	(80)	1	VEADIAGHGQEVLIR
☑	104	803.80	1605.58	1605.85	-0.27	0	(68)	1	VEADIAGHGQEVLIR
☑	105	803.81	1605.60	1605.85	-0.25	0	86	1	VEADIAGHGQEVLIR

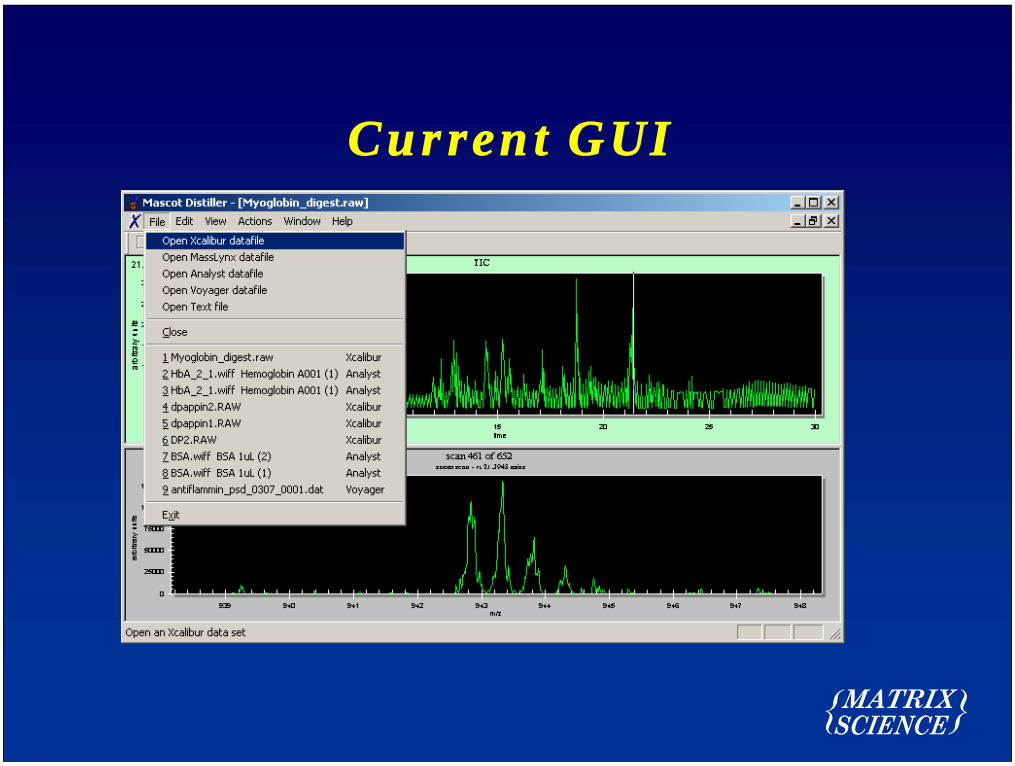
It is quite common to see Mascot results as shown here. As you can see, there are many identical peptides. Better signal to noise can be obtained by averaging these spectra together.

The principles for doing this are quite simple, but somewhat surprisingly, many software packages don't provide nearly enough control over which scans will be grouped together



Time Domain parameters control how scans are grouped together

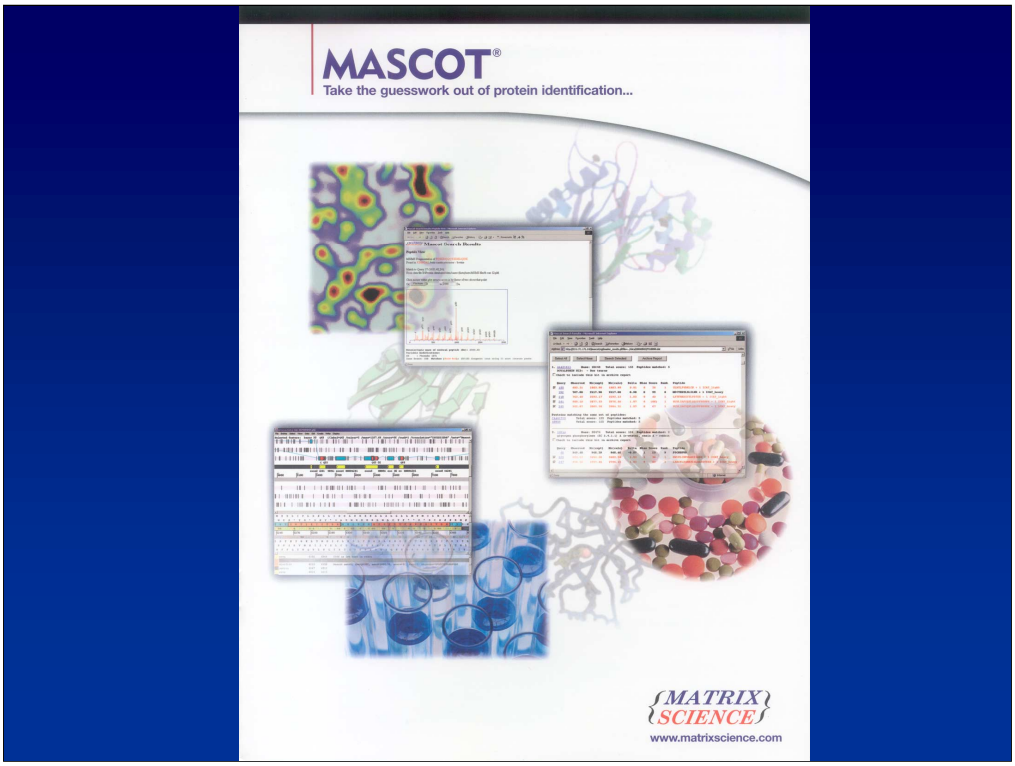
The features of Mascot Distiller are ...



The GUI is very basic, and is only intended as a debugging tool.

More importantly, Mascot Distiller is a function library that can be called from Mascot Daemon or from your own code

The development has been a three way partnership with Biovision and GSK


MASCOT®
Take the guesswork out of protein identification...
MATRIX SCIENCE
www.matrixscience.com