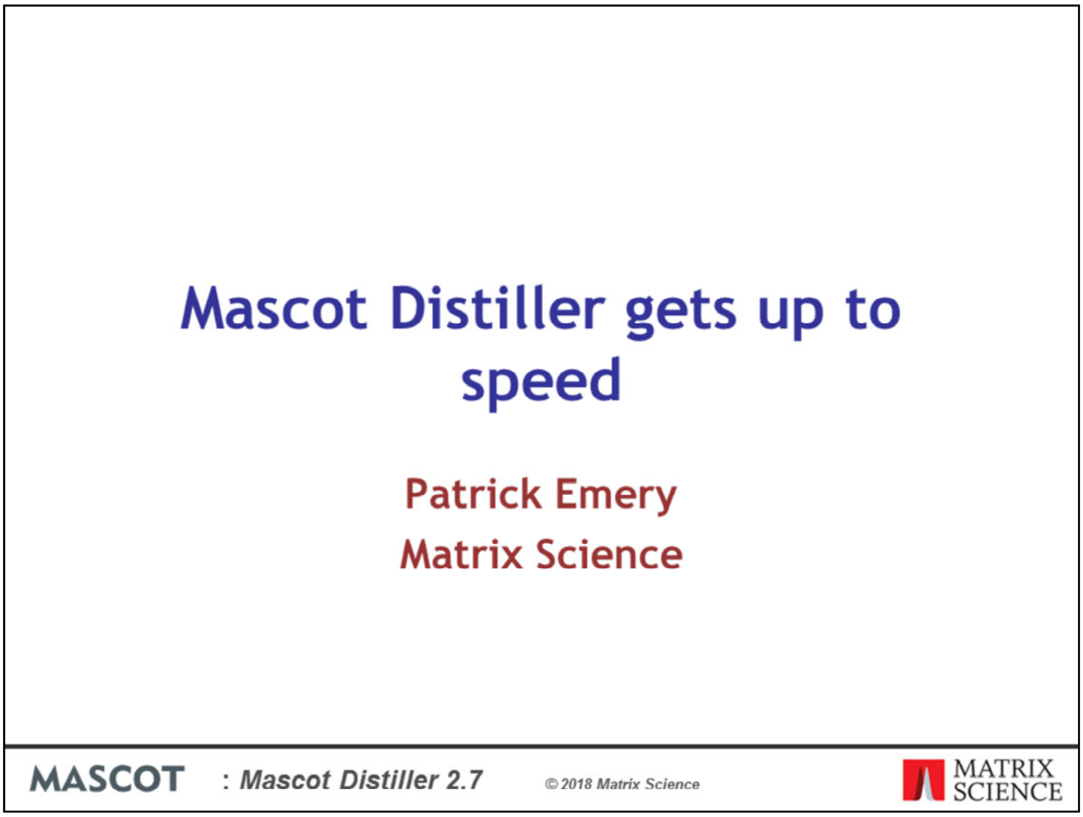
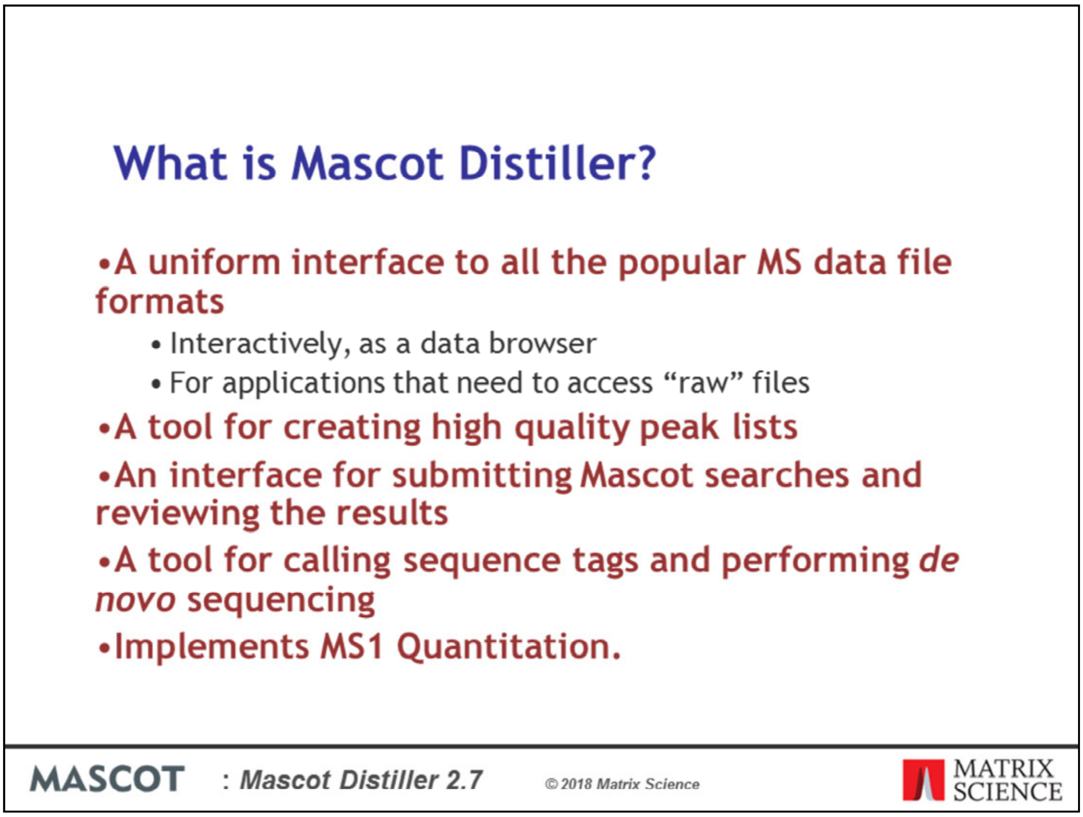


# Mascot Distiller gets up to speed

Patrick Emery

Matrix Science



## What is Mascot Distiller?

- •A uniform interface to all the popular MS data file formats
  - Interactively, as a data browser
  - For applications that need to access “raw” files
- •A tool for creating high quality peak lists
- •An interface for submitting Mascot searches and reviewing the results
- •A tool for calling sequence tags and performing *de novo* sequencing
- •Implements MS1 Quantitation.

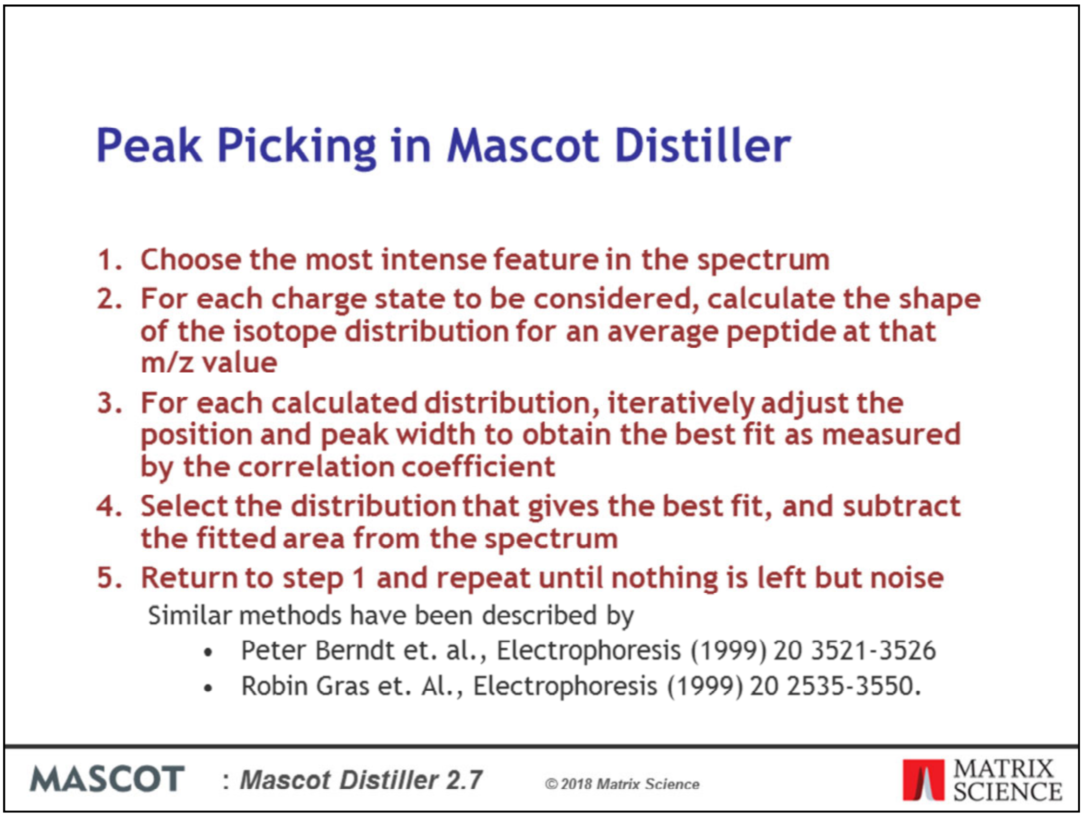
MASCOT : *Mascot Distiller 2.7* © 2018 Matrix Science MATRIX SCIENCE

Most laboratories will have instruments from more than one manufacturer. The instrument data systems are necessarily complex, so there can be a steep learning curve for someone who comes into the lab and just wants to browse their data or generate peak lists. The first benefit of Mascot Distiller is that you can access all of the popular data formats from a single user interface.

Another reason for developing Distiller was to produce high quality peak lists without having to constantly tweak peak detection parameters. Poor quality peak lists translate into poor quality Mascot scores.

Distiller is also a powerful way to review Mascot search results. And, if Mascot fails to get a match, you can perform de novo sequencing and interpret sequence tags for tag searches

Finally, Distiller is used for quantitation methods that require information from the raw data file, either because it is necessary to integrate the elution profile of each precursor peptide or because information is required for precursor peptides that were not used to trigger MS/MS scans, so are missing from the peak list.



## Peak Picking in Mascot Distiller

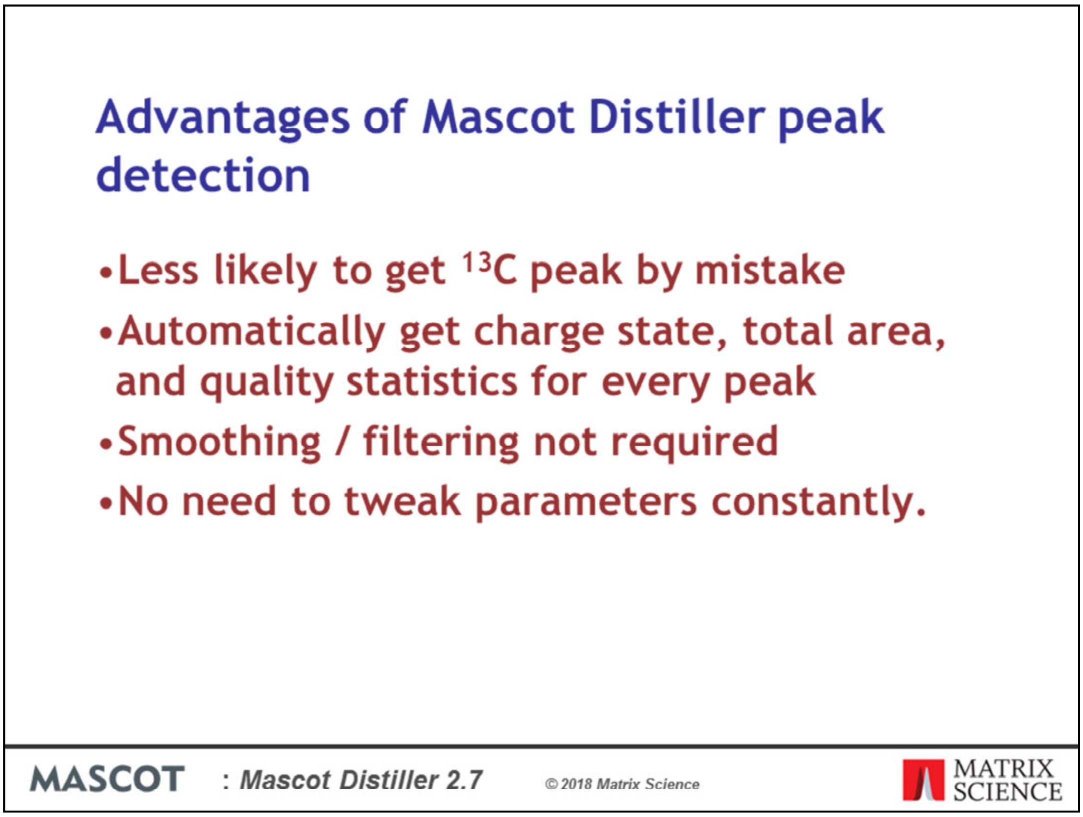
1. Choose the most intense feature in the spectrum
2. For each charge state to be considered, calculate the shape of the isotope distribution for an average peptide at that m/z value
3. For each calculated distribution, iteratively adjust the position and peak width to obtain the best fit as measured by the correlation coefficient
4. Select the distribution that gives the best fit, and subtract the fitted area from the spectrum
5. Return to step 1 and repeat until nothing is left but noise

Similar methods have been described by

- Peter Berndt et. al., Electrophoresis (1999) 20 3521-3526
- Robin Gras et. Al., Electrophoresis (1999) 20 2535-3550.

MASCOT : *Mascot Distiller 2.7* © 2018 Matrix Science MATRIX SCIENCE

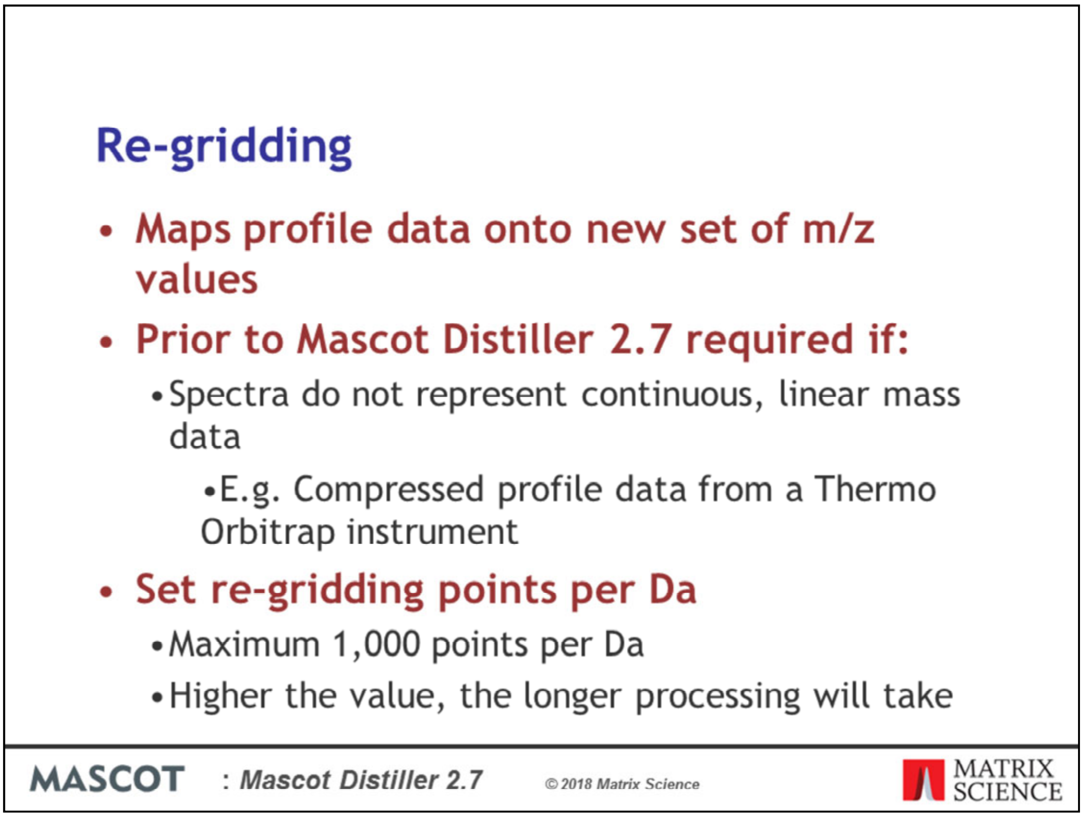
Mascot Distiller detects peaks by attempting to fit an ideal isotopic distribution to the experimental data. These are the steps in this process.



## Advantages of Mascot Distiller peak detection

- Less likely to get $^{13}C$ peak by mistake
- Automatically get charge state, total area, and quality statistics for every peak
- Smoothing / filtering not required
- No need to tweak parameters constantly.

There are a number of advantages to this approach.



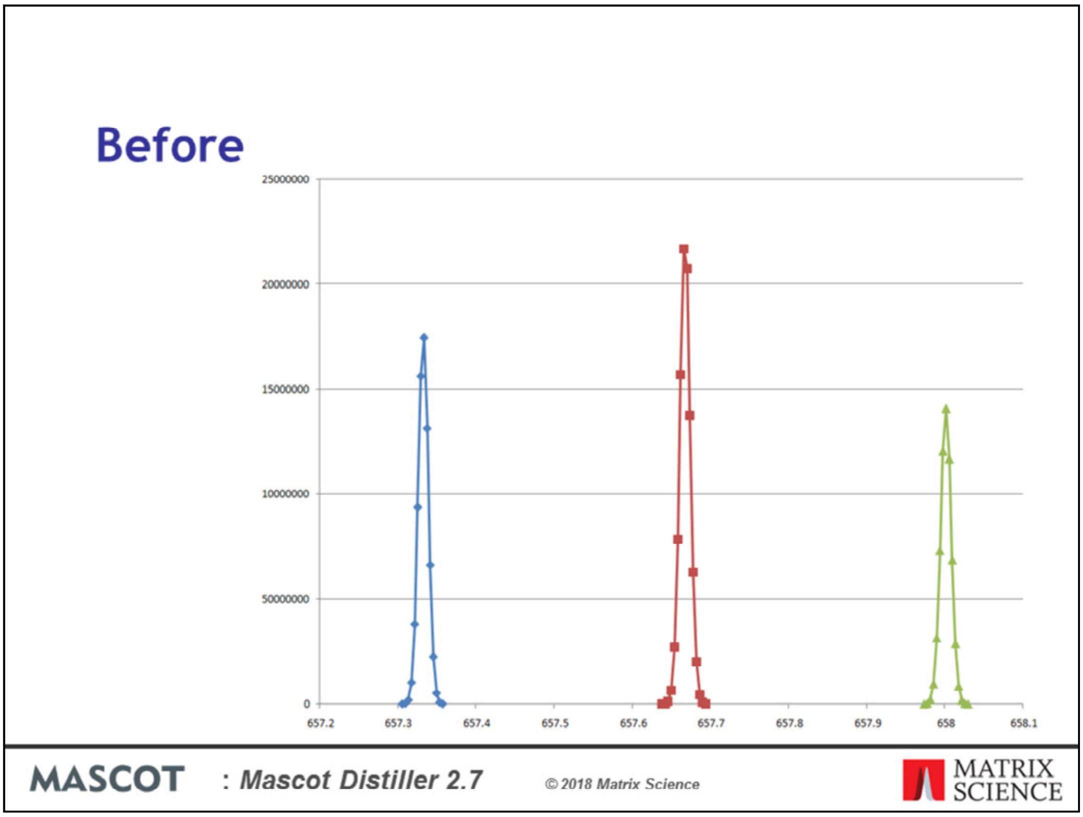
## Re-gridding

- Maps profile data onto new set of m/z values
- Prior to Mascot Distiller 2.7 required if:
  - Spectra do not represent continuous, linear mass data
    - E.g. Compressed profile data from a Thermo Orbitrap instrument
- Set re-gridding points per Da
  - Maximum 1,000 points per Da
  - Higher the value, the longer processing will take

MASCOT : *Mascot Distiller 2.7* © 2018 Matrix Science MATRIX SCIENCE

Of course, nothing is perfect…

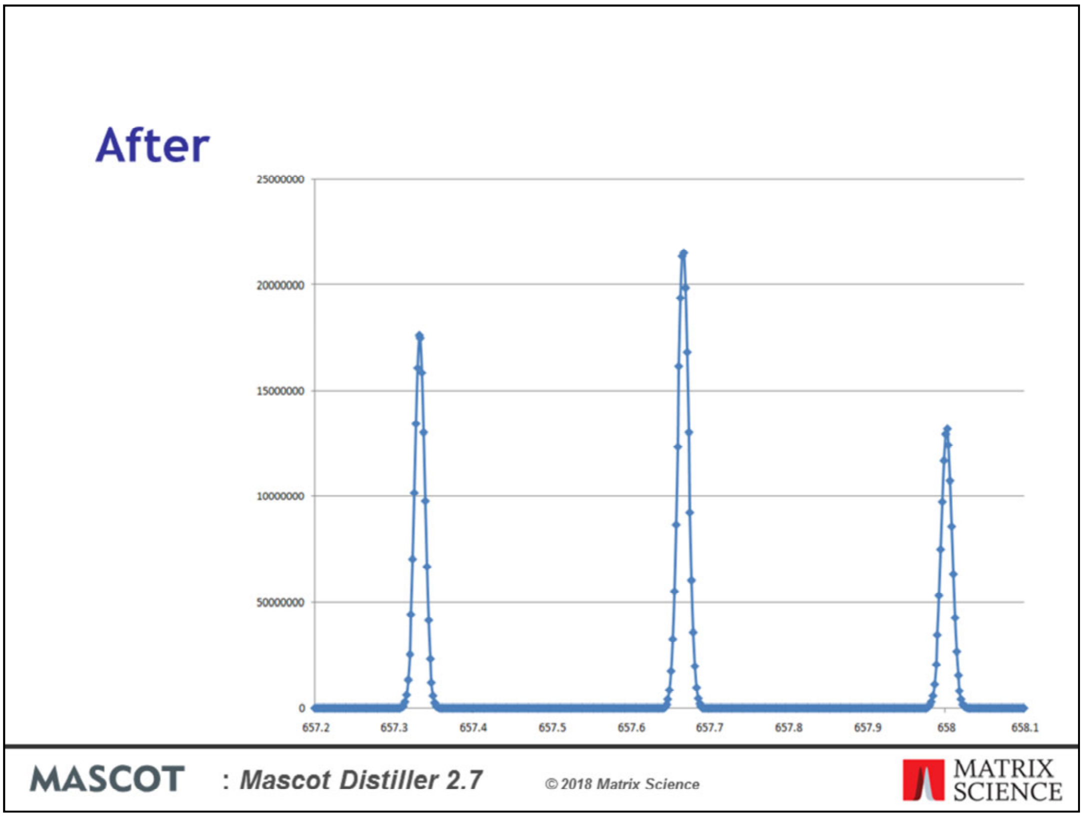
Prior to Mascot Distiller 2.7, peak picking required that the profile data was on a linear mass scale, with evenly distributed data points over the m/z axis.  However, some profile data are compressed by dropping runs of zero intensity data points. Therefore, before older versions of Mascot Distiller could process these scans, it was necessary to transform these data into evenly distributed m/z values. This is called re-gridding, and was also required before spectra with non-linear mass scales could be summed.



## Before

MASCOT : *Mascot Distiller 2.7* © 2018 Matrix Science MATRIX SCIENCE

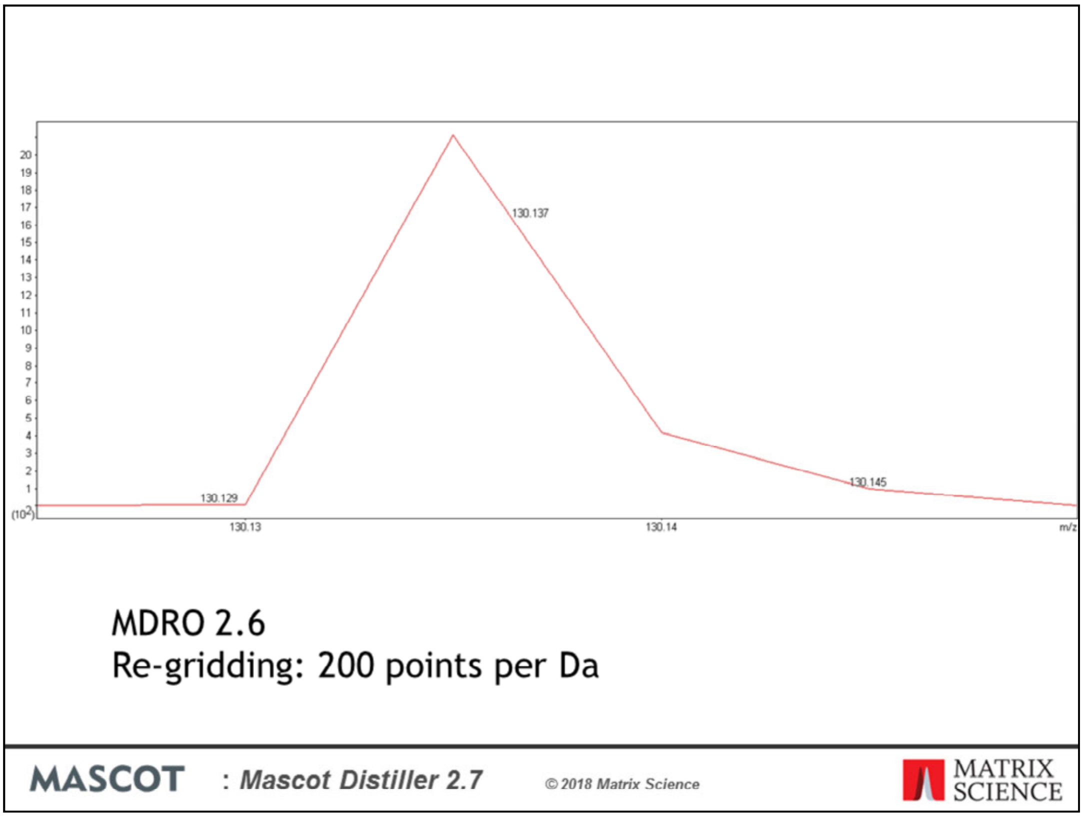
So prior to re-gridding, if you have compressed (or sparse) profile raw data, your data would look something like this - non continuous with distinct data points for the peaks and missing zero values between.



## After

MASCOT : *Mascot Distiller 2.7* © 2018 Matrix Science MATRIX SCIENCE

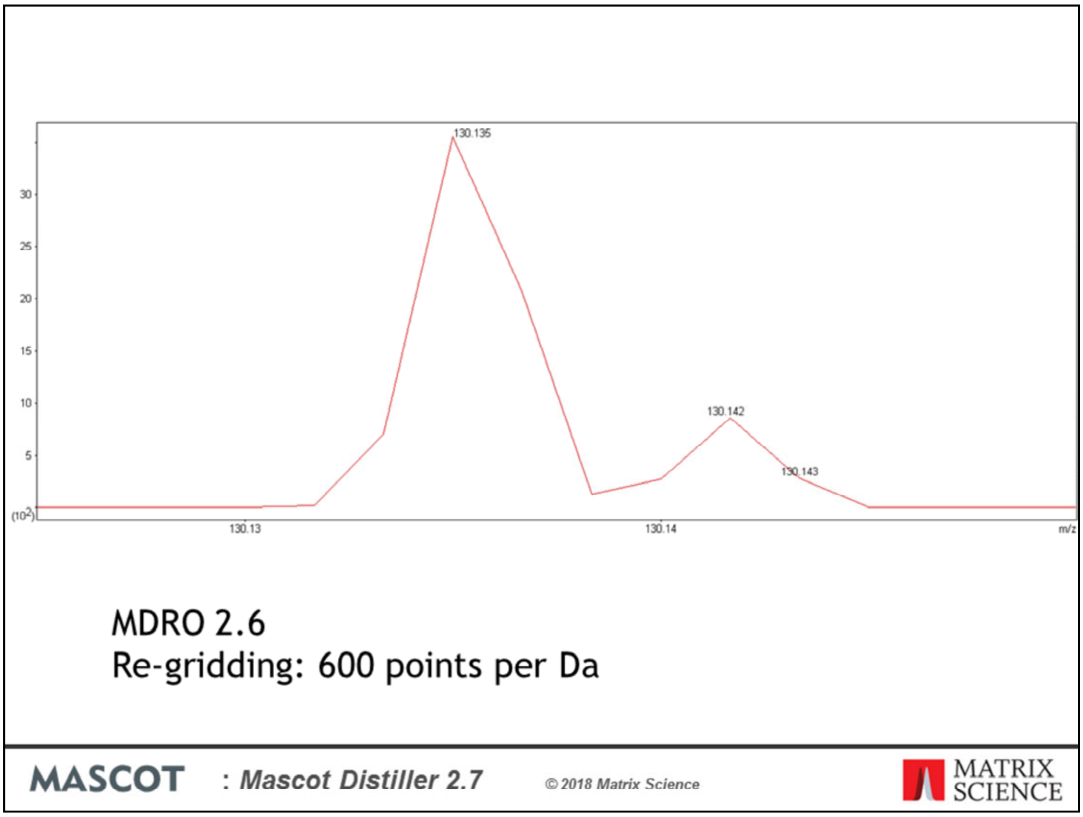
This is the same data re-gridded to 600 points per Dalton, and you now have the data points evenly spaced on the m/z axis and the zero values added back in. We now have continuous data, which is what Distiller prior to 2.7 required for peak detection.



MDRO 2.6
Re-gridding: 200 points per Da

MASCOT : *Mascot Distiller 2.7* © 2018 Matrix Science MATRIX SCIENCE

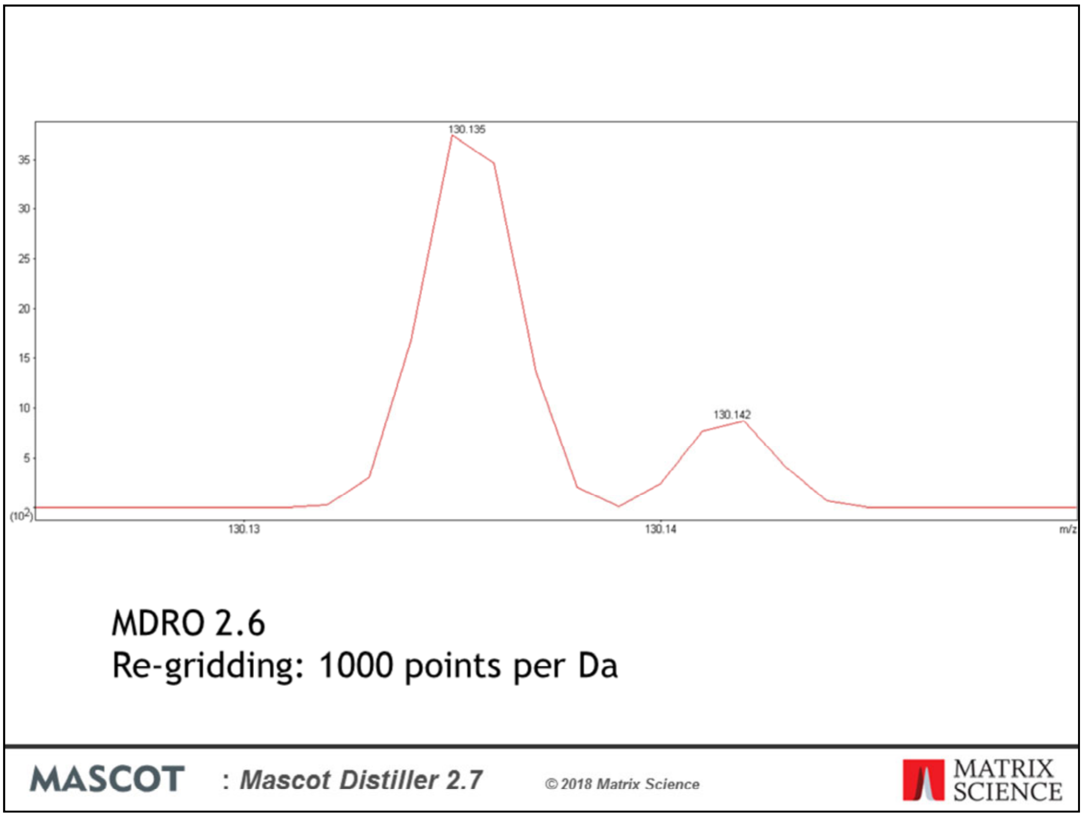
The resolution of the re-gridded data depends on the number of points per Dalton specified in the processing options. To show you the effect of this, here we are looking a some high resolution data. Initially, we have are regridding to 200 points per Dalton during peak detection. The results, are not good!



MDRO 2.6
Re-gridding: 600 points per Da

MASCOT : *Mascot Distiller 2.7* © 2018 Matrix Science MATRIX SCIENCE

This is actually TMT 10-plex data and we’re looking at a pair of reporter ion peaks. Re-gridding at 600 points per Da and we can at least now see the two reporter ion peaks in the trace, but the shapes aren’t very good. Within the reporter ion region we’ve carried out single peak peaking, rather than isotope distribution mapping, to try and capture the reporter ion peaks and the satellite peak has been split during peak detection.

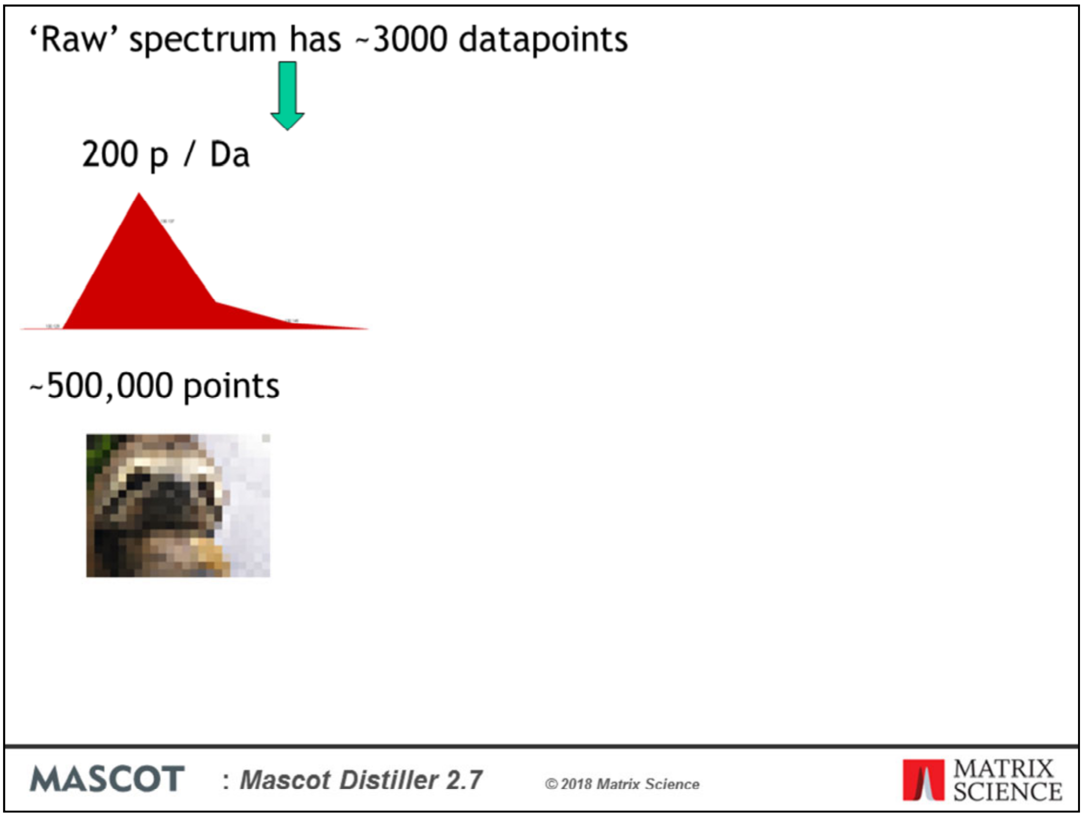


MDRO 2.6
Re-gridding: 1000 points per Da

MASCOT : Mascot Distiller 2.7

MATRIX SCIENCE

Things look a bit better using the maximum possible value of 1000 points per Da, where we have two much better defined peaks and we have successfully picked both of them. So, for high resolution data like this, you'd be looking to set the re-gridding at or close to the maximum value of 1000. However:



‘Raw’ spectrum has ~3000 datapoints

200 p / Da

~500,000 points

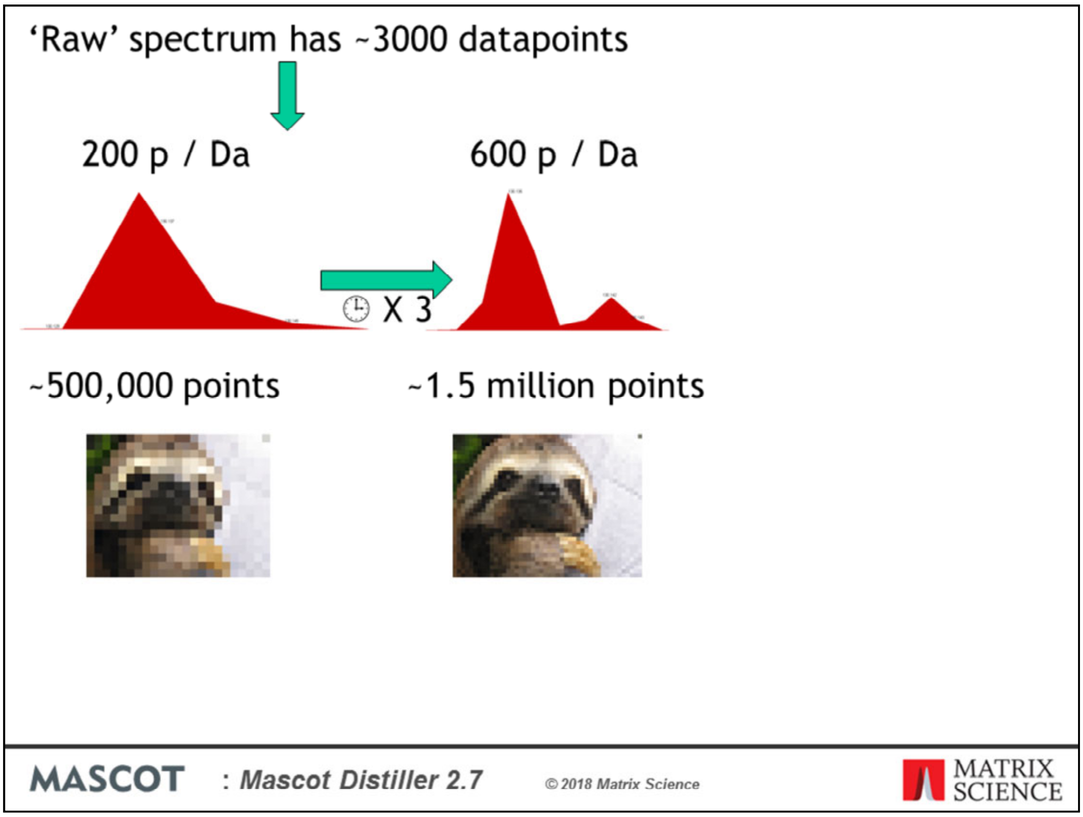
MASCOT : *Mascot Distiller 2.7*

*© 2018 Matrix Science*

MATRIX SCIENCE

Our raw spectrum contained approximately 3000 data points.  If we re-grid these to 200 points / Da, we now have increased to the number of points in the spectrum to approximately 500,000 datapoints after re-gridding, and we’re left with a fairly low resolution spectrum.

If we increase the points per Dalton from 200 to 600



‘Raw’ spectrum has ~3000 datapoints

200 p / Da

600 p / Da

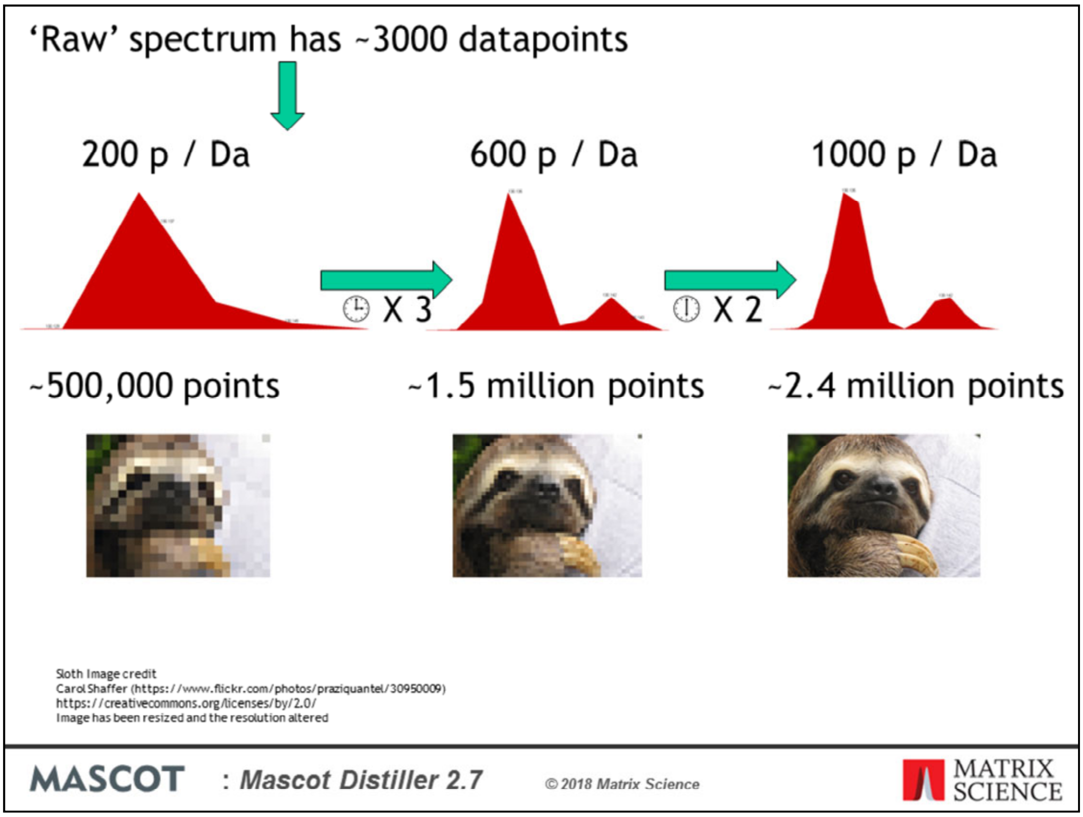
🕒 X 3

~500,000 points

~1.5 million points

MASCOT : *Mascot Distiller 2.7* © 2018 Matrix Science MATRIX SCIENCE

Our resolution improves, but now our spectrum contains approximately 1.5 million points and our processing time for the datafile increases by approximately a factor of three. When we increased to 1000 points per Dalton



‘Raw’ spectrum has ~3000 datapoints

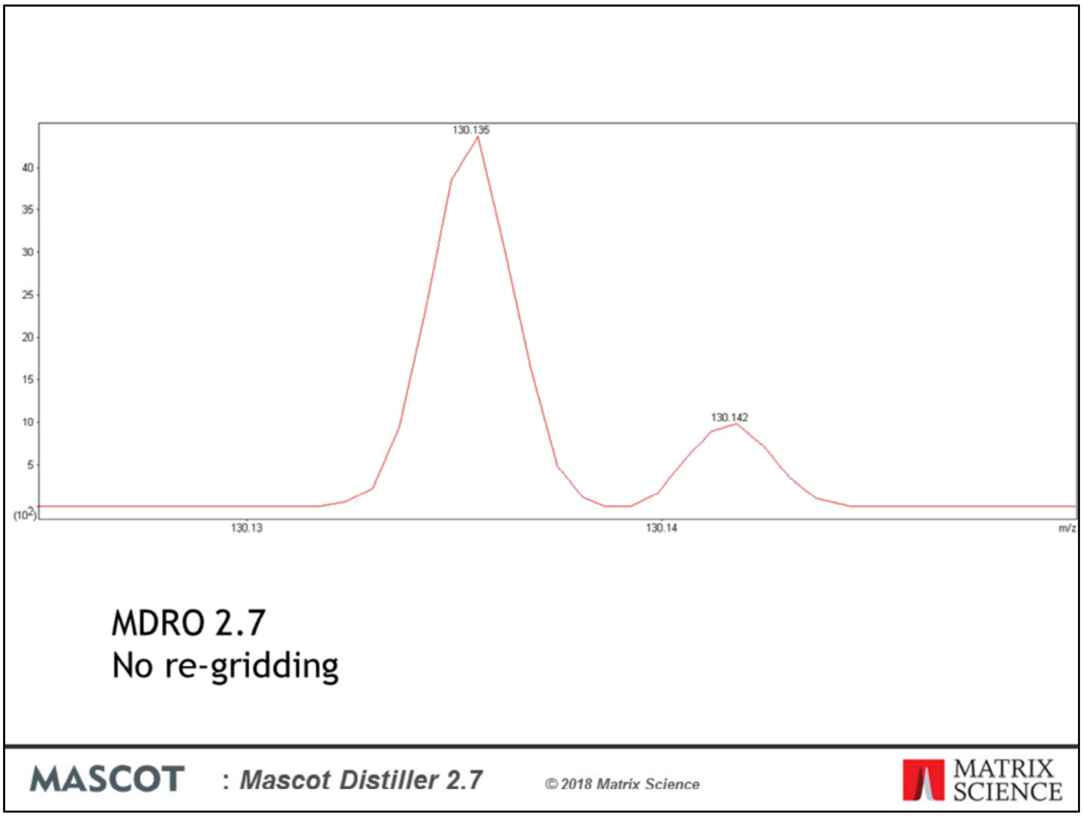
~500,000 points

~1.5 million points

~2.4 million points

Sloth Image credit
Carol Shaffer (https://www.flickr.com/photos/praziquantel/30950009)
https://creativecommons.org/licenses/by/2.0/
Image has been resized and the resolution altered

We now have pretty good resolution, but the spectrum now contains approximately 2.4 million points and processing time roughly doubles again. So, re-gridding is a major overhead, especially for high resolution data where you need to set a high number of points per Dalton; and that is assuming that 1000 points per Dalton is even sufficient to handle the data – for some very high resolution data it isn’t enough and Distiller 2.6 or earlier simply can’t handle it properly.



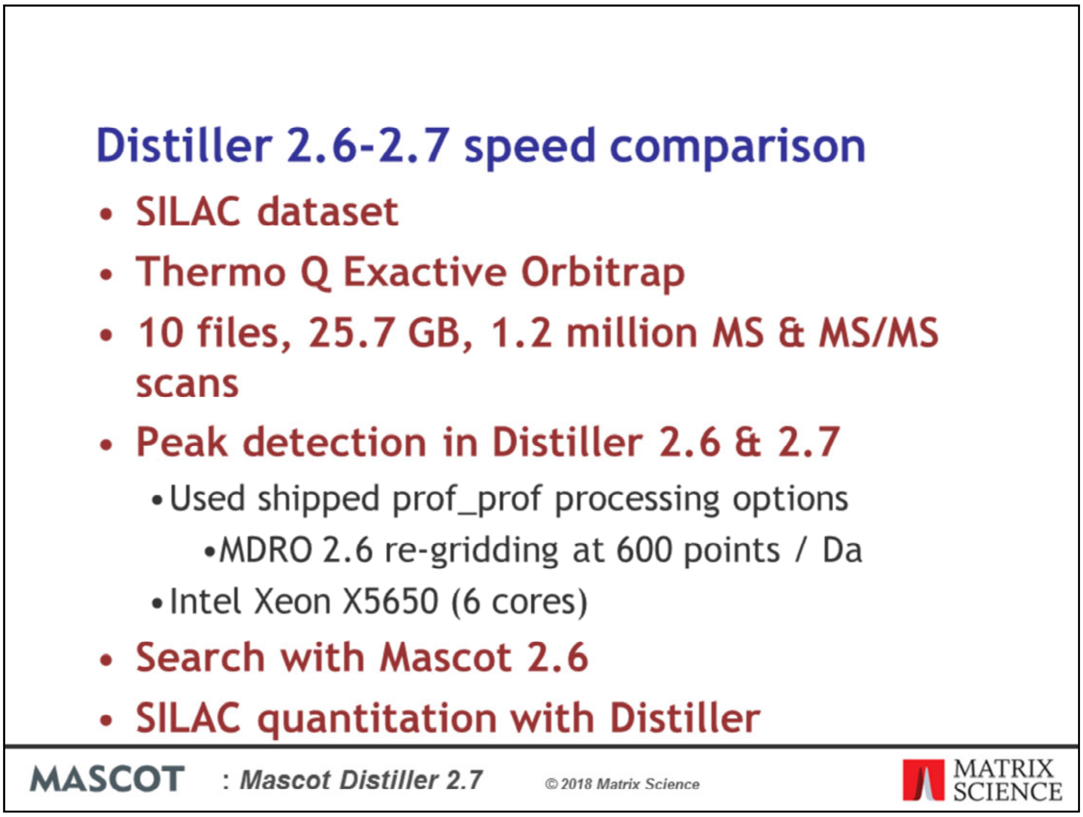
MDRO 2.7
No re-gridding

MASCOT : *Mascot Distiller 2.7*

MATRIX SCIENCE

So to improve the situation, in Mascot Distiller 2.7 we've altered the peak detection algorithm to remove the requirement for re-gridding this type of sparse data, so that it can work directly from the raw profile data.

This is what that same pair of reporter ion peaks look like in Mascot Distiller 2.7. No re-gridding has been carried out, this is plotted from the ~3K points in the source data, but the peaks are higher resolution than they were re-gridded at 1000 points per Dalton, and Distiller has had no trouble picking the reporter ion peaks correctly.



## Distiller 2.6-2.7 speed comparison

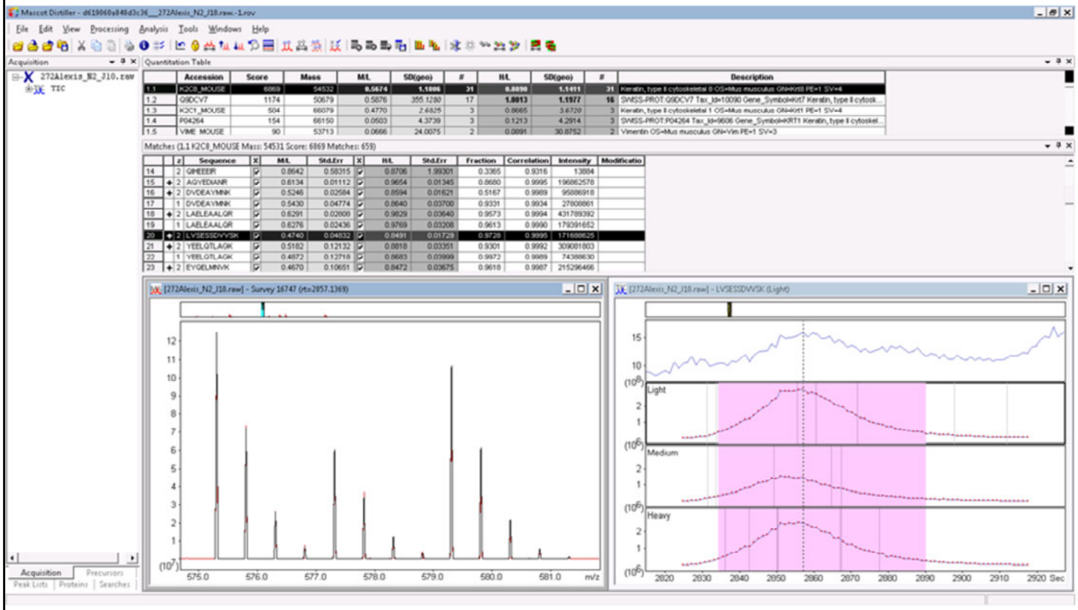
- SILAC dataset
- Thermo Q Exactive Orbitrap
- 10 files, 25.7 GB, 1.2 million MS & MS/MS scans
- Peak detection in Distiller 2.6 & 2.7
  - Used shipped prof_prof processing options
    - MDRO 2.6 re-gridding at 600 points / Da
  - Intel Xeon X5650 (6 cores)
- Search with Mascot 2.6
- SILAC quantitation with Distiller

MASCOT : *Mascot Distiller 2.7* © 2018 Matrix Science MATRIX SCIENCE

So, we seem to be getting equivalent or better results from Distiller 2.7. Now lets see if we've speeded up peak picking. In theory, removing the requirement for re-gridding on datasets where it was previously required should result in a significant speed improvement during processing.

To test this, we took a publically available SILAC dataset from the PRIDE repository. The dataset comprises 10 raw files, saved as compressed profile data. The data were captured using a Thermo Q-Exactive instrument. The raw files were processed using Mascot Distiller 2.6 and 2.7 using the shipped prof_prof processing options. With Distiller 2.6, this therefore requires re-gridding. The system used for processing was an old Intel Xeon X5650 based system – these are quite old processors now and you'd get better performance from a more recent i7 based desktop system for example.

The generated peaklists were then searched using Mascot 2.6 against SwissProt with a mouse taxonomy filter, and a contaminants database. Finally, the results were quantified in Distiller.


Mascot Distiller - d619060a848d3c36__272Alexis_N2_J18.raw.-1.rov
File Edit View Processing Analysis Tools Windows Help
Acquisition
272Alexis_N2_J10.raw
TIC
Acquisition
Precursors
Peak Lists
Proteins
Searches

Quantitation Table

| | Accession | Score | Mass | M/L | SD(geo) | # | H/L | SD(geo) | # | Description |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.1 | K2C8_MOUSE | 6869 | 54532 | 0.5674 | 1.1806 | 31 | 0.8890 | 1.1411 | 31 | Keratin, type II cytoskeletal 8 OS=Mus musculus GN=Krt8 PE=1 SV=4 |
| 1.2 | Q9DCV7 | 1174 | 50679 | 0.5876 | 355.1280 | 17 | 1.8013 | 1.1977 | 16 | SWISS-PROT:Q9DCV7 Tax_Id=10090 Gene_Symbol=Krt7 Keratin, type II cytosk... |
| 1.3 | K2C1_MOUSE | 504 | 66079 | 0.4770 | 2.6825 | 3 | 0.8665 | 3.6720 | 3 | Keratin, type II cytoskeletal 1 OS=Mus musculus GN=Krt1 PE=1 SV=4 |
| 1.4 | P04264 | 154 | 66150 | 0.0503 | 4.3739 | 3 | 0.1213 | 4.2914 | 3 | SWISS-PROT:P04264 Tax_Id=9606 Gene_Symbol=KRT1 Keratin, type II cytoskel... |
| 1.5 | VIME_MOUSE | 90 | 53713 | 0.0666 | 24.0075 | 2 | 0.0891 | 30.8752 | 2 | Vimentin OS=Mus musculus GN=Vim PE=1 SV=3 |

Matches (1.1 K2C8_MOUSE Mass: 54531 Score: 6869 Matches: 659)

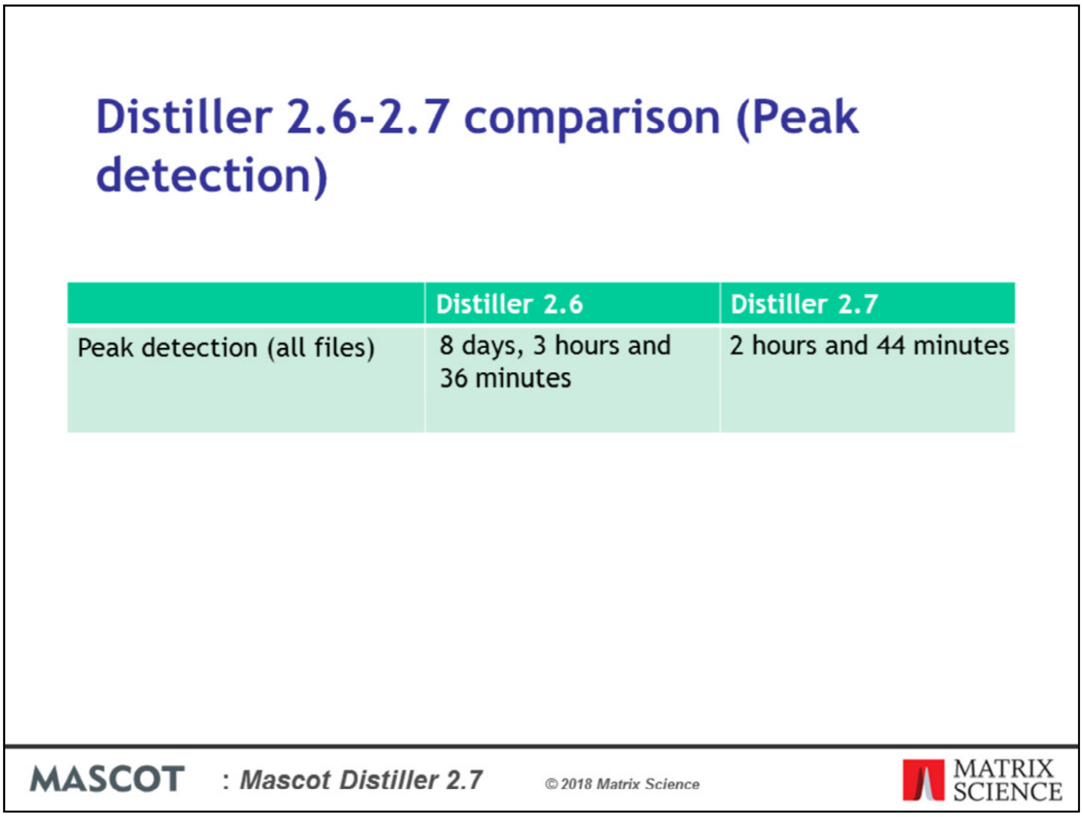
| | | z | Sequence | X | M/L | Std.Err | X | H/L | Std.Err | Fraction | Correlation | Intensity | Modificatio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 14 | | 2 | QIHEEIR | ☑ | 0.8642 | 0.58315 | ☑ | 0.8706 | 1.99301 | 0.3365 | 0.9316 | 13884 | |
| 15 | ✦ | 2 | AQYEDIANR | ☑ | 0.6134 | 0.01112 | ☑ | 0.9654 | 0.01345 | 0.8680 | 0.9995 | 196862578 | |
| 16 | ✦ | 2 | DVDEAYMNK | ☑ | 0.5246 | 0.02584 | ☑ | 0.8594 | 0.01621 | 0.5167 | 0.9989 | 95886918 | |
| 17 | | 1 | DVDEAYMNK | ☑ | 0.5430 | 0.04774 | ☑ | 0.8640 | 0.03700 | 0.9331 | 0.9934 | 27808861 | |
| 18 | ✦ | 2 | LAELEAALQR | ☑ | 0.6291 | 0.02808 | ☑ | 0.9829 | 0.03640 | 0.9573 | 0.9994 | 431789392 | |
| 19 | | 1 | LAELEAALQR | ☑ | 0.6276 | 0.02436 | ☑ | 0.9769 | 0.03208 | 0.9613 | 0.9990 | 179391652 | |
| 20 | ✦ | 2 | LVSESSDVVSK | ☑ | 0.4740 | 0.04832 | ☑ | 0.8491 | 0.01729 | 0.9728 | 0.9995 | 171688625 | |
| 21 | ✦ | 2 | YEELQTLAGK | ☑ | 0.5182 | 0.12132 | ☑ | 0.8818 | 0.03351 | 0.9301 | 0.9992 | 309081803 | |
| 22 | | 1 | YEELQTLAGK | ☑ | 0.4872 | 0.12718 | ☑ | 0.8683 | 0.03999 | 0.9972 | 0.9989 | 74388630 | |
| 23 | ✦ | 2 | EYQELMNVK | ☑ | 0.4670 | 0.10651 | ☑ | 0.8472 | 0.03675 | 0.9618 | 0.9987 | 215296466 | |

MASCOT : *Mascot Distiller 2.7*

*© 2018 Matrix Science*

MATRIX SCIENCE

The entire process was automated using Mascot Daemon, which saved separate Distiller project files for each of the raw files. The project files contain the picked peaklists, the Mascot search and Distiller quantitation results – here we're looking at a single SILAC peptide quantitation match with the XICs for the peptide shown in the window on the lower right.



## Distiller 2.6-2.7 comparison (Peak detection)

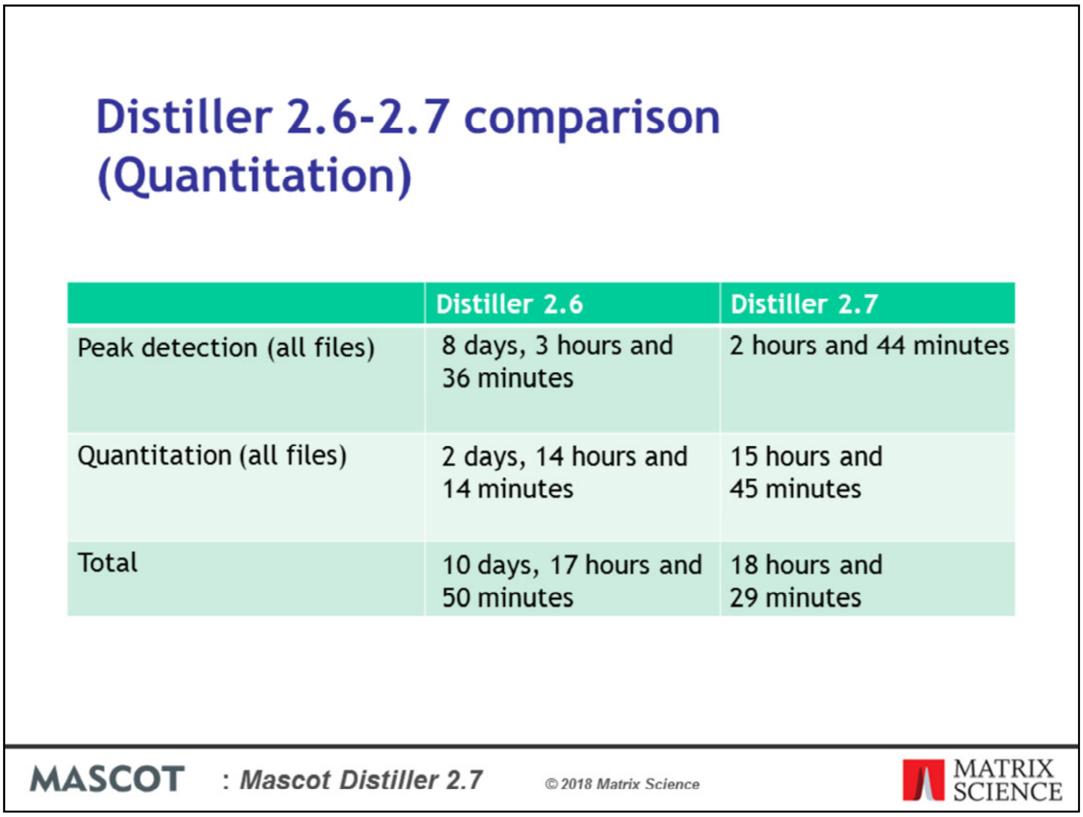
| | Distiller 2.6 | Distiller 2.7 |
|---|---|---|
| Peak detection (all files) | 8 days, 3 hours and 36 minutes | 2 hours and 44 minutes |

MASCOT : *Mascot Distiller 2.7* *© 2018 Matrix Science* MATRIX SCIENCE

So, how long did all this take? Using Distiller 2.6 and re-gridding with 600 p/Da, simply processing all the raw data to peaklists took …. Over 8 days!

Re-gridding is a significant processing overhead regardless of the resolution - even if we'd dropped the re-gridding resolution to 200 points per Da, which would have been far too low for these data, it would take nearly 3 days to process all the files.

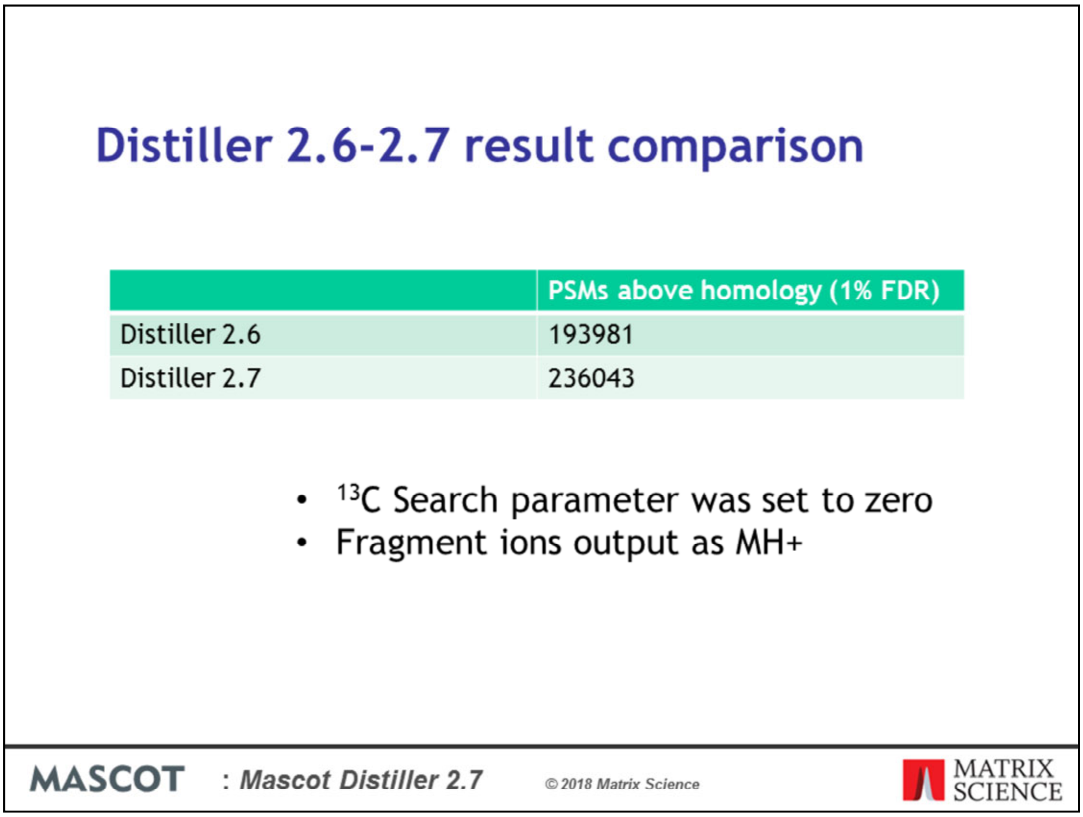
With Mascot Distiller 2.7, the improvement is almost embarrassing as processing now only took 2 and ¾ hours….I think we can agree that this is a significant improvement!



## Distiller 2.6-2.7 comparison (Quantitation)

| | Distiller 2.6 | Distiller 2.7 |
|---|---|---|
| Peak detection (all files) | 8 days, 3 hours and 36 minutes | 2 hours and 44 minutes |
| Quantitation (all files) | 2 days, 14 hours and 14 minutes | 15 hours and 45 minutes |
| Total | 10 days, 17 hours and 50 minutes | 18 hours and 29 minutes |

During the quantitation phase, Distiller has to carry out some additional peak detection on MS scans in the XIC regions. With Distiller 2.6, quantifying all 10 of the data files took a little over 2 and a half days, giving us a total processing time of over 10 days. With Distiller 2.7, we see a significant improvement again, and quantitation took 15 and 3/4 hours for all 10 data files, giving us a total processing time of 18 and a half hours.

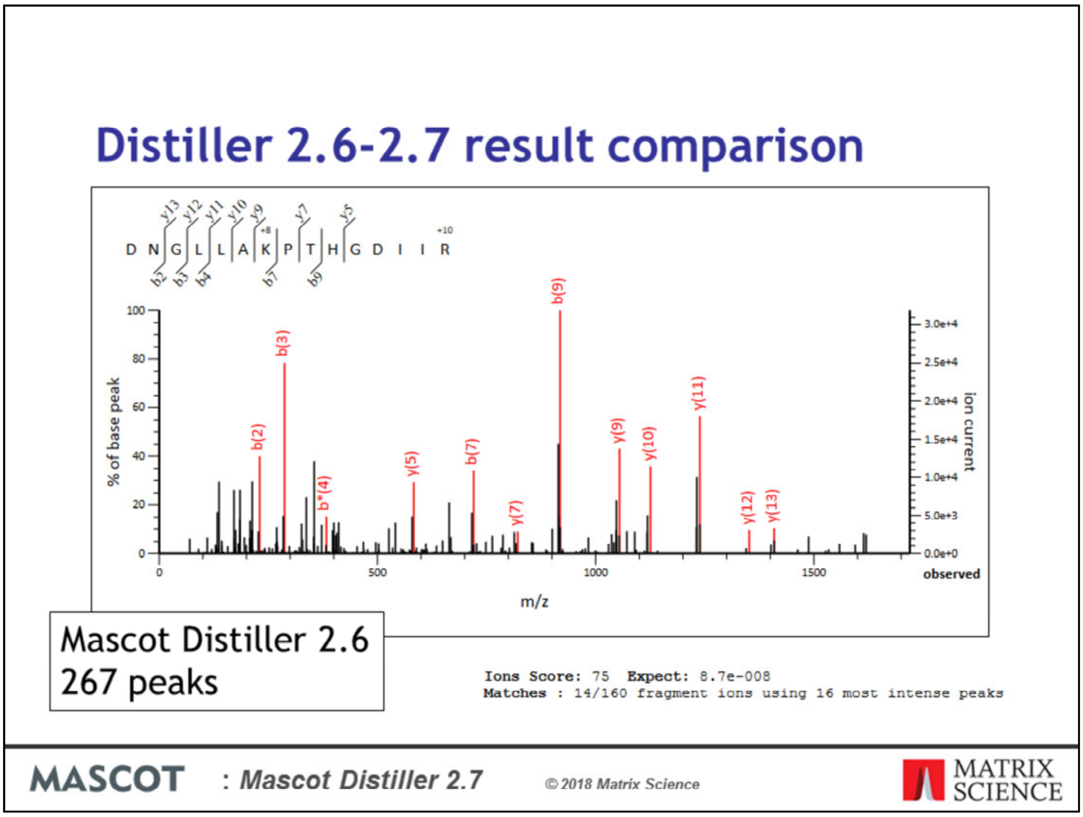


## Distiller 2.6-2.7 result comparison

| | PSMs above homology (1% FDR) |
|---|---|
| Distiller 2.6 | 193981 |
| Distiller 2.7 | 236043 |

- $^{13}C$ Search parameter was set to zero
- Fragment ions output as MH+

Taking a look at the number of significant PSMs at a 1% False Discovery Rate, we can see that we’re getting more significant matches back from Distiller 2.7 – a roughly 20% improvement.



## Distiller 2.6-2.7 result comparison

Mascot Distiller 2.6
267 peaks

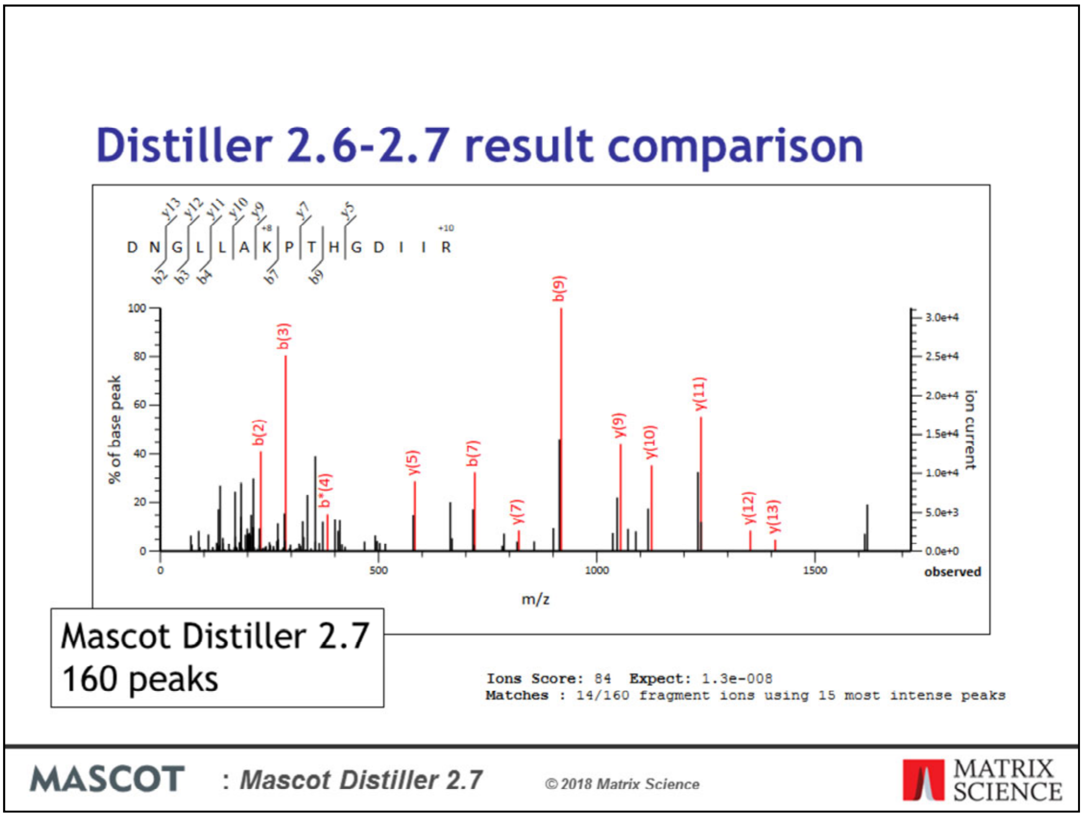
```
Ions Score: 75  Expect: 8.7e-008
Matches : 14/160 fragment ions using 16 most intense peaks
```

MASCOT : *Mascot Distiller 2.7* © 2018 Matrix Science MATRIX SCIENCE

Lets delve into some of the reasons we’re seeing better results from Distiller 2.7.

For some matches, we’re seeing an improvement in score.

This is a match from a Peak List generated by Mascot Distiller 2.6, which contains 267 selected fragment ion peaks.  It’s a reasonable match match, and gets an ions score of 75. Keep your eyes on the unmatched noise peaks…



## Distiller 2.6-2.7 result comparison

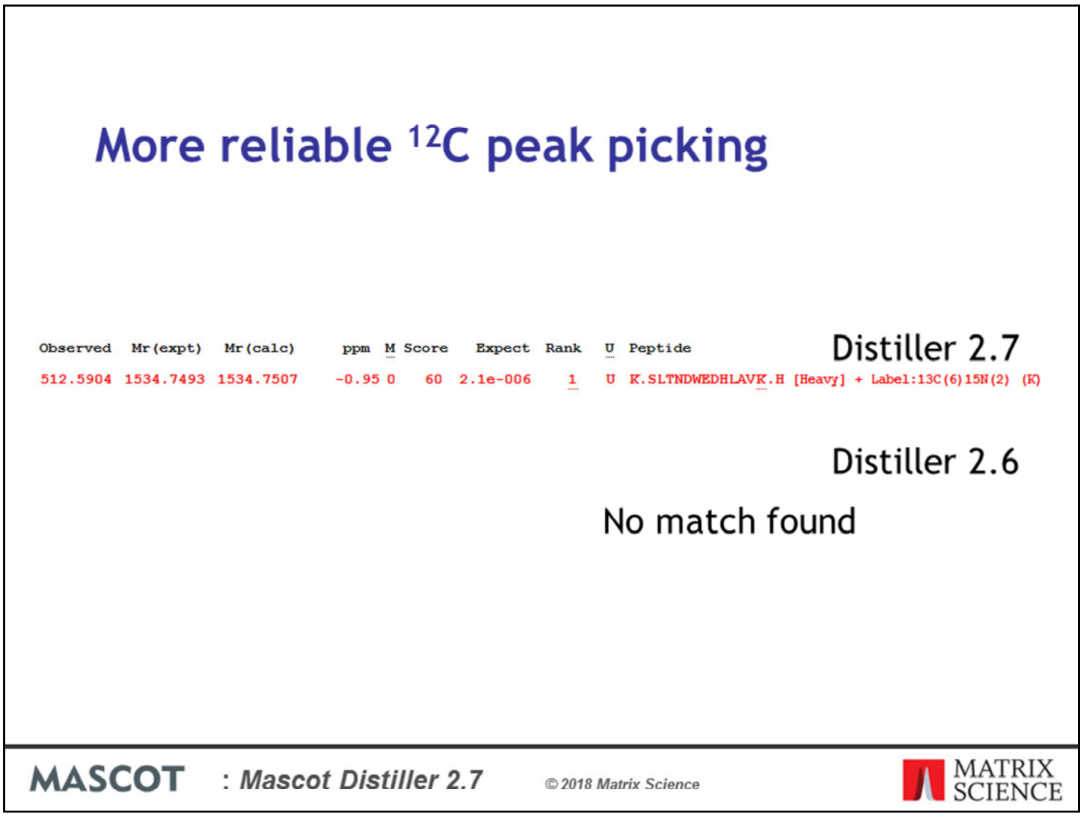
Mascot Distiller 2.7
160 peaks

```
Ions Score: 84  Expect: 1.3e-008
Matches : 14/160 fragment ions using 15 most intense peaks
```

MASCOT : *Mascot Distiller 2.7* *© 2018 Matrix Science* MATRIX SCIENCE

And this is the match to the peak list for the same MS/MS scan generated by Mascot Distiller 2.7. It's a very similar match, but hopefully you can see we have a cleaner peak list, containing just 160 peaks, and we have fewer unmatched noise peaks – giving us better signal to noise, which gives us a slightly higher ions score of 84.

There are plenty of examples of this from more marginal matches, and this sometimes lifts the match above the significance threshold.



## More reliable $^{12}$C peak picking

Distiller 2.7

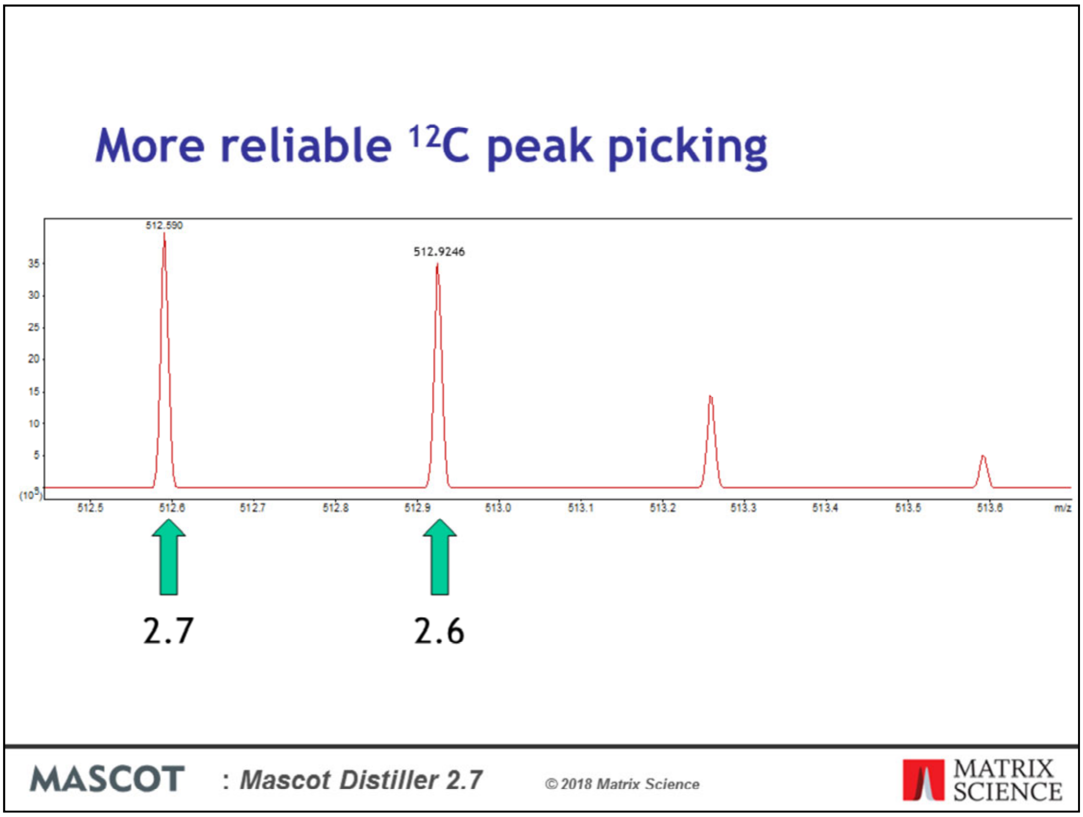
```
Observed  Mr(expt)   Mr(calc)     ppm  M Score   Expect  Rank   U  Peptide
512.5904 1534.7493 1534.7507    -0.95 0    60  2.1e-006    1    U  K.SLTNDWEDHLAVK.H [Heavy] + Label:13C(6)15N(2) (K)
```

Distiller 2.6

No match found

Although the isotope distribution mapping peak picking technique employed by Mascot Distiller is far less likely to pick the 13C peak by mistake, on occasion it does still happen, and this is something else we've improved in Distiller 2.7.

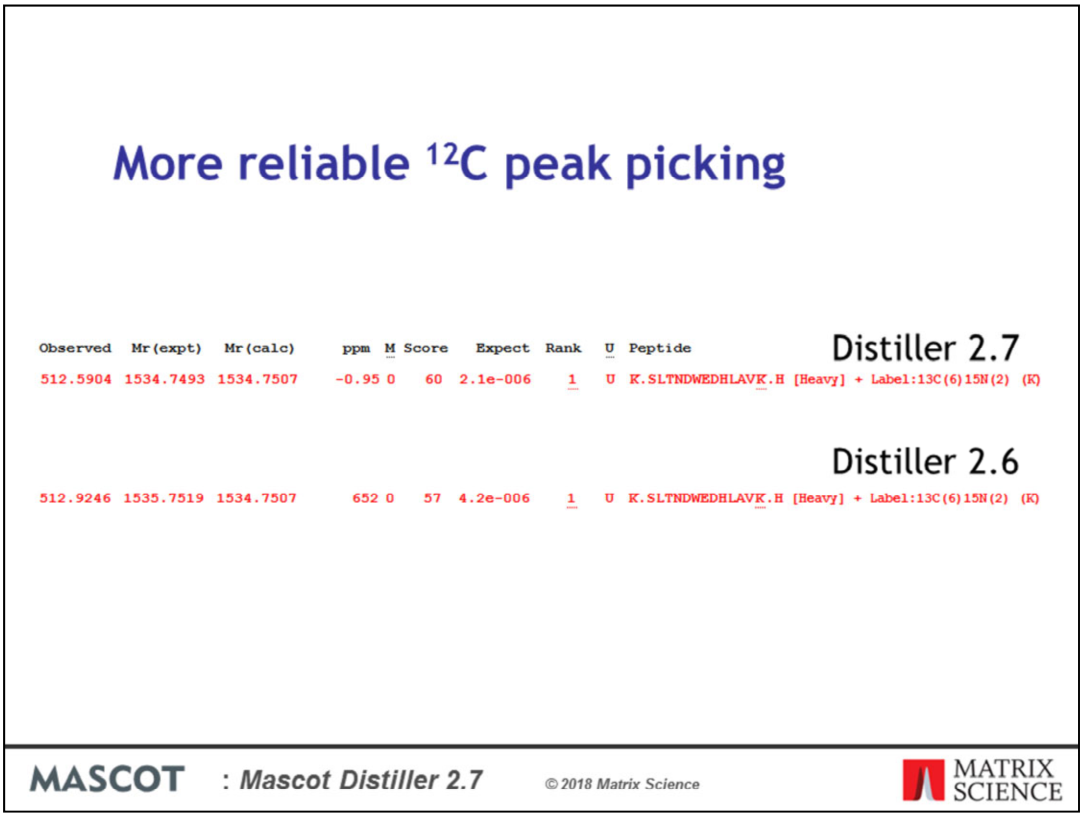
Take a look at this match from the same MS/MS scan from the Distiller 2.7 peak list – we have a reasonably high scoring match to a heavy labelled peptide. But in the search result from Distiller 2.6, we didn't find a match at all.



## More reliable $^{12}$C peak picking

MASCOT : *Mascot Distiller 2.7* © 2018 Matrix Science MATRIX SCIENCE

If we look at the parent survey scan in this region, Distiller 2.7 has correctly picked the 12C peak – and we can see that Mascot Distiller 2.6 clearly picked the 13C peak



## More reliable $^{12}$C peak picking

Distiller 2.7

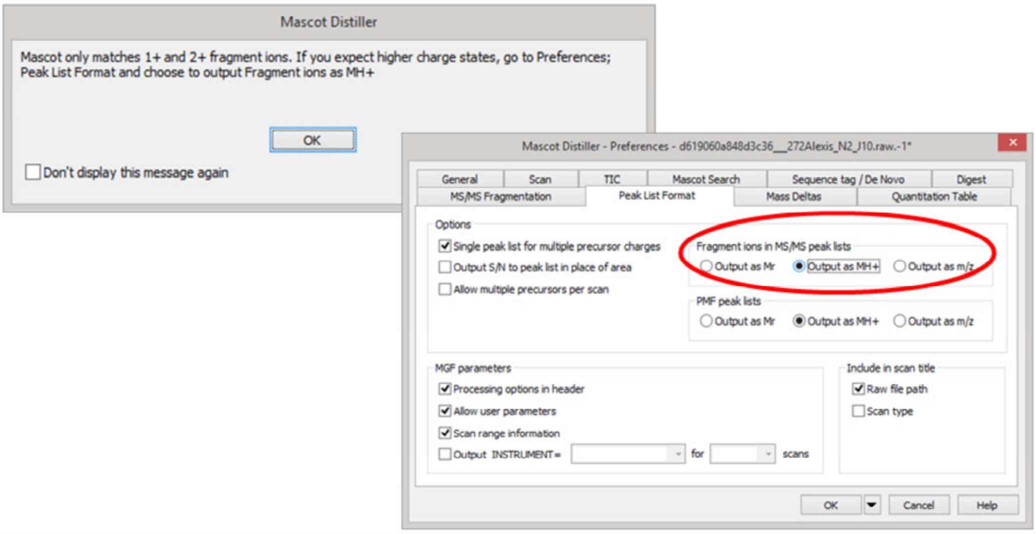
```
Observed  Mr(expt)  Mr(calc)      ppm  M Score   Expect Rank   U Peptide
512.5904 1534.7493 1534.7507    -0.95 0    60  2.1e-006    1   U K.SLTNDWEDHLAVK.H [Heavy] + Label:13C(6)15N(2) (K)
```

Distiller 2.6

```
512.9246 1535.7519 1534.7507      652 0    57  4.2e-006    1   U K.SLTNDWEDHLAVK.H [Heavy] + Label:13C(6)15N(2) (K)
```

Repeating the search from Distiller 2.6 with the 13C search parameter set to 1, and now we get the match with a very similar score to that from Distiller 2.7, but with a precursor mass delta of just over 1Da

# More accurate charge state calling



Mascot Distiller

Mascot only matches 1+ and 2+ fragment ions. If you expect higher charge states, go to Preferences; Peak List Format and choose to output Fragment ions as MH+

□ Don't display this message again

Mascot Distiller - Preferences - d619060a848d3c36__272Alexis_N2_J10.raw.-1*

Options

☑ Single peak list for multiple precursor charges
□ Output S/N to peak list in place of area
□ Allow multiple precursors per scan

Fragment ions in MS/MS peak lists

○ Output as Mr ● Output as MH+ ○ Output as m/z

PMF peak lists

○ Output as Mr ● Output as MH+ ○ Output as m/z

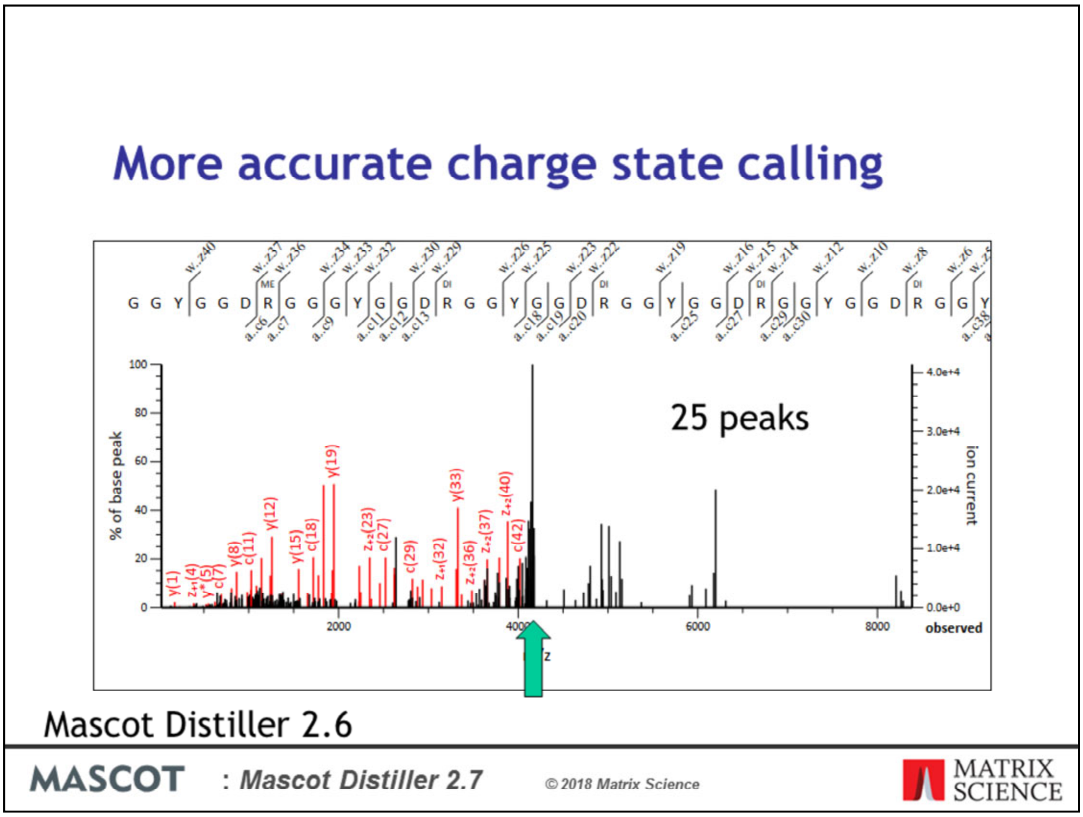
MGF parameters

☑ Processing options in header
☑ Allow user parameters
☑ Scan range information
□ Output INSTRUMENT= for scans

Include in scan title

☑ Raw file path
□ Scan type

Something you may have seen if you've submitted a Mascot search from Distiller is this warning dialog, reminding you that Mascot only searches for singly and doubly charge fragment ions. If you have higher charge state fragment ions present in your data, you should de-charge the fragment ion masses to MH+ in the peak lists using the Peak List format options dialog. Of course, the accuracy of this relies on the fragment ion charge states having been accurately determined in the first place.



## More accurate charge state calling

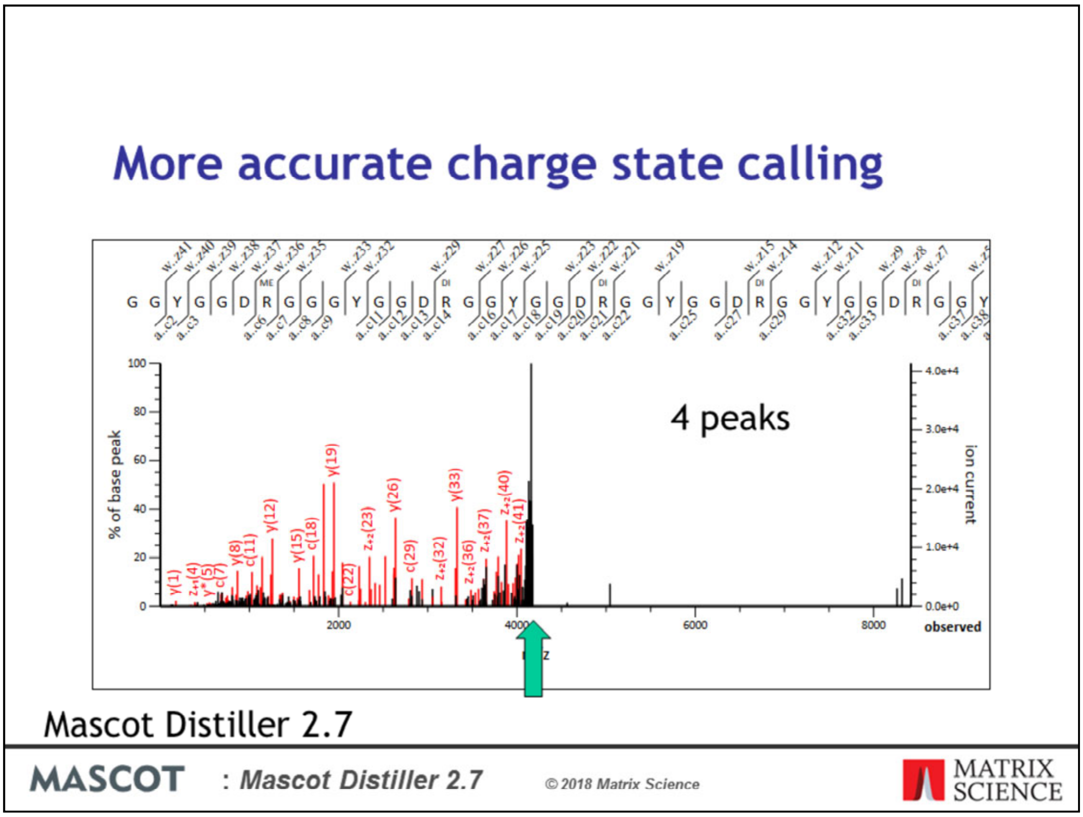
25 peaks

Mascot Distiller 2.6

MASCOT : *Mascot Distiller 2.7*

MATRIX SCIENCE

This is a match from a different dataset to a decharged MS/MS peaklist from a 6+ precursor. The mass of the precursor is 4174.83, so we shouldn't have any fragment peaks above that value in the peaklist – in fact we have 25 peaks above the precursor mass.



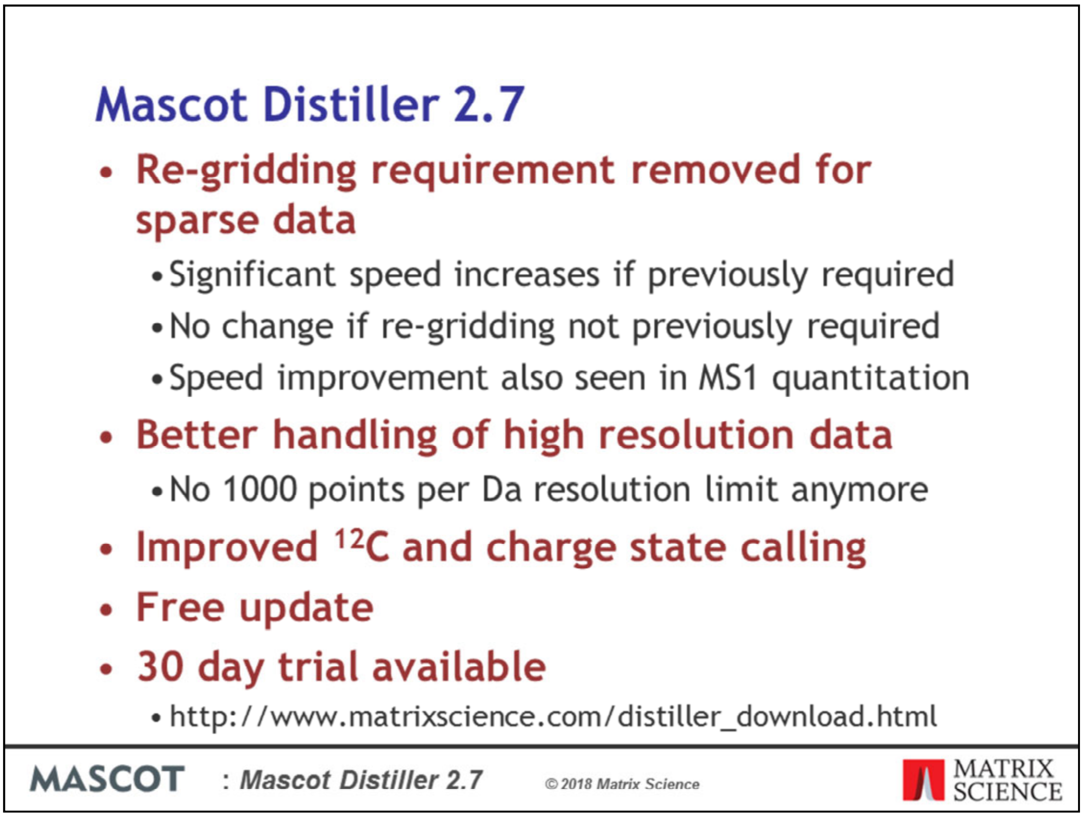
## More accurate charge state calling

4 peaks

Mascot Distiller 2.7

MASCOT : *Mascot Distiller 2.7* © 2018 Matrix Science MATRIX SCIENCE

Here is the peak list from the same MS/MS spectrum generated by Mascot Distiller 2.7. We now only have 4 peaks with masses greater than the precursor – in other words, our charge state calling on the fragment ions was more accurate.



## Mascot Distiller 2.7

- Re-gridding requirement removed for sparse data
  - Significant speed increases if previously required
  - No change if re-gridding not previously required
  - Speed improvement also seen in MS1 quantitation
- Better handling of high resolution data
  - No 1000 points per Da resolution limit anymore
- Improved $^{12}C$ and charge state calling
- Free update
- 30 day trial available
  - http://www.matrixscience.com/distiller_download.html

MASCOT : *Mascot Distiller 2.7* *© 2018 Matrix Science* MATRIX SCIENCE

So, with Mascot Distiller 2.7, we've removed the requirement for re-gridding of sparse profile data, and this results in significant speed increases when processing datasets which previously required that step. However, if your data didn't previously require re-gridding, then you're unlikely to see any significant speed increases.

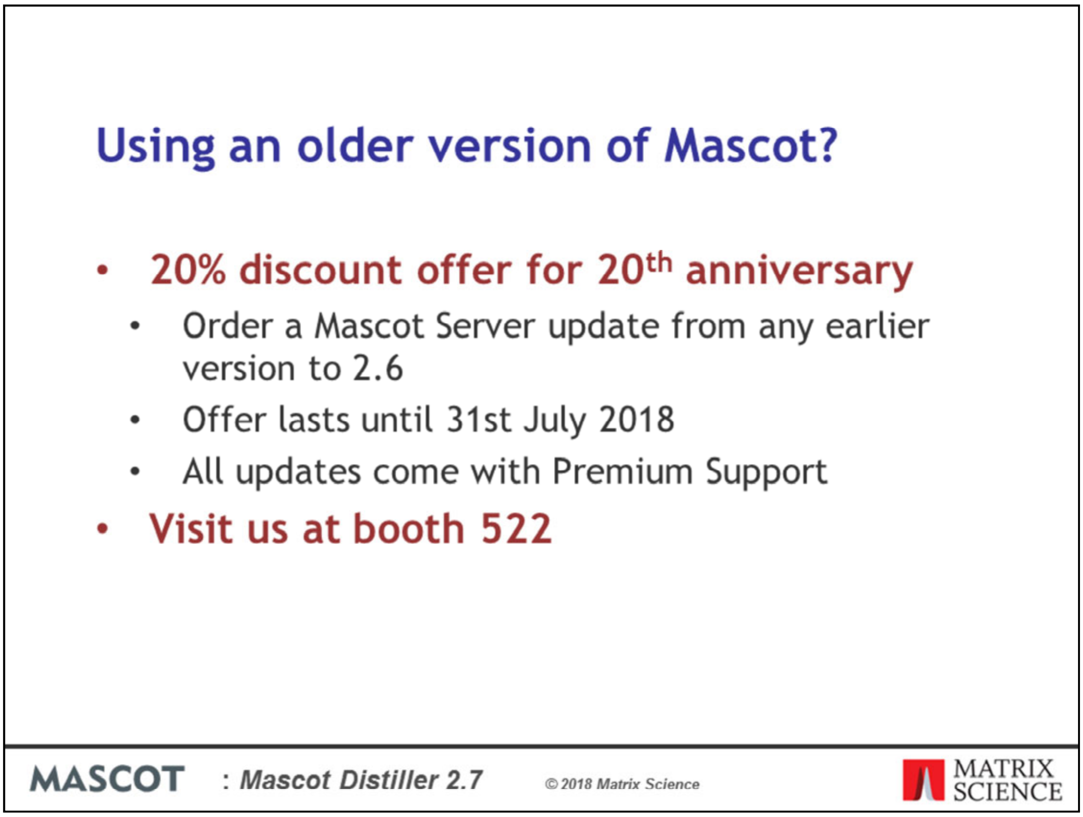
You also see speed improvements in MS1 based quantitation for these types of data, because precursor based quantitation requires additional peak picking during XIC calculation.

Because Distiller doesn't need to re-grid data anymore, it can handle high resolution data much better than it could before – it won't bump into the 1000 points per Da limit it used to have.

We've also improved the 12C and ion charge state calling.

If you already have Mascot Distiller in-house, and you haven't already updated, then this is a free update.

If you want to test Mascot Distiller on your data, then you can get a 30-day evaluation licence from us – details on distiller_download webpage on our public website.



## Using an older version of Mascot?

- **20% discount offer for 20th anniversary**
  - Order a Mascot Server update from any earlier version to 2.6
  - Offer lasts until 31st July 2018
  - All updates come with Premium Support
- **Visit us at booth 522**

MASCOT : *Mascot Distiller 2.7* *© 2018 Matrix Science* MATRIX SCIENCE

And finally, if you’re on an earlier version of Mascot now is a great time to update! 2018 is the 20th anniversary year of Matrix Science Ltd, and to celebrate, we’re offering 20% off on all Mascot Server updates. Come talk to us at booth 522 after the presentation and we’re happy to send you a quote immediately.