Re-engineering key components of our data processing pipeline

Aaron Robinson

Pacific Northwest
NATIONAL LABORATORY

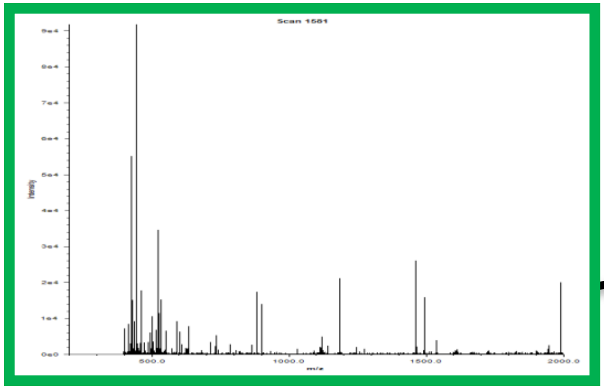
Background

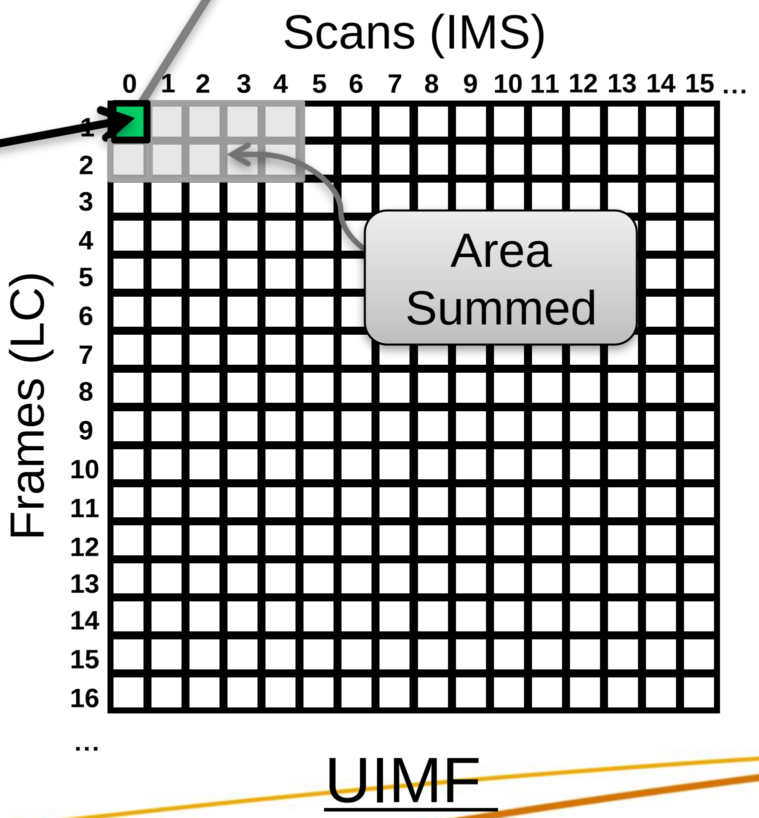
- Ion Mobility Spectrum (IMS)
  - Software in development
  - Large data volumes
- Unified Ion Mobility Format (UIMF)
  - Data management solution
  - SQLite database
  - Structured SQL schema
- Deisotoping
  - Decon2LS used for deconvolution
  - Challenges due to fast frame/scan rate

Decon2LS Deisotoping

Raw Data:

Bin	Intensity
0	0
269328	6
269328	6
269328	6
0	0
298781	20
…	…



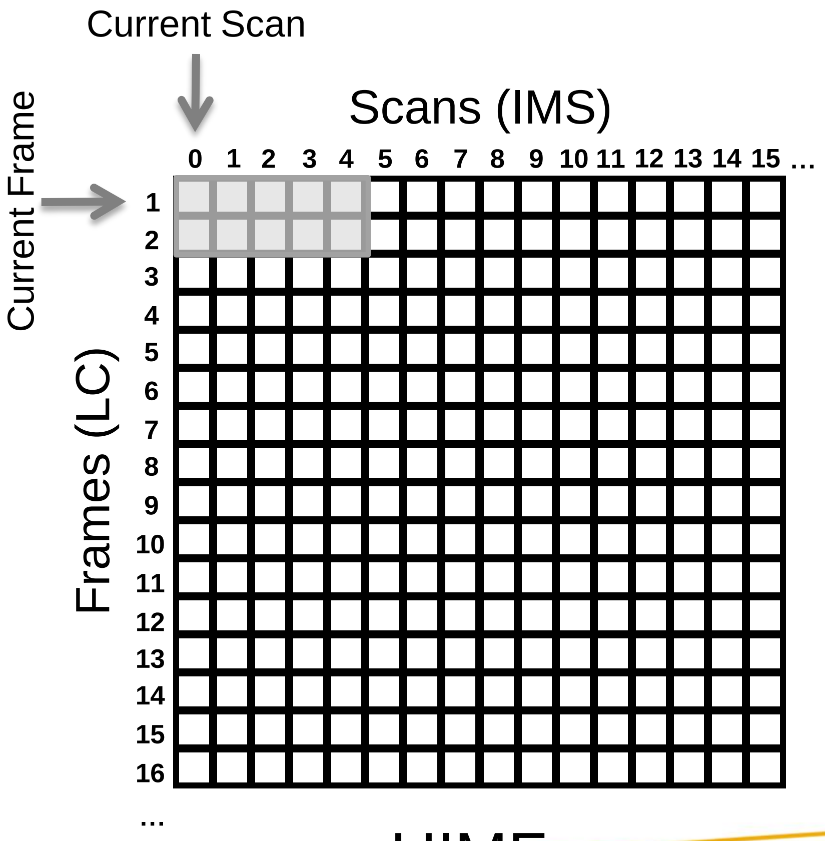


- Bins are converted into m/z values
- Summing enhances deisotoping
  - Provides greater intensities
- Sliding window: 3 Frames by 9 Scans
- Width = Current Scan ± 4
- Height = Current Frame ± 1

Decon2LS Deisotoping

Process:

1. Sum window & deisotope data

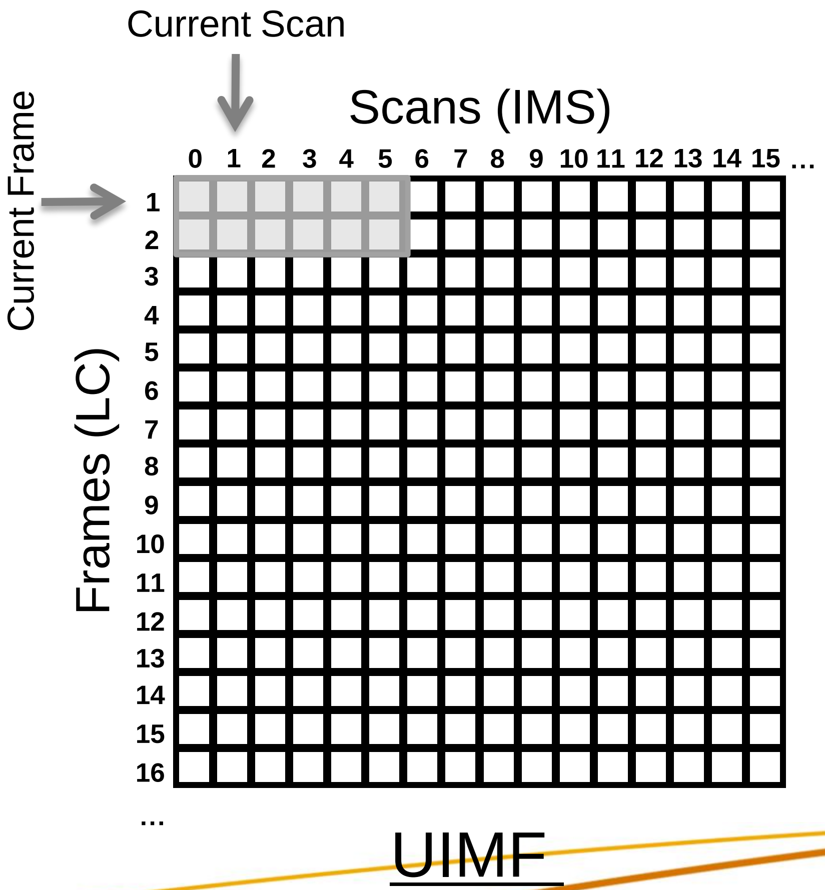


UIMF

Decon2LS Deisotoping

Process:

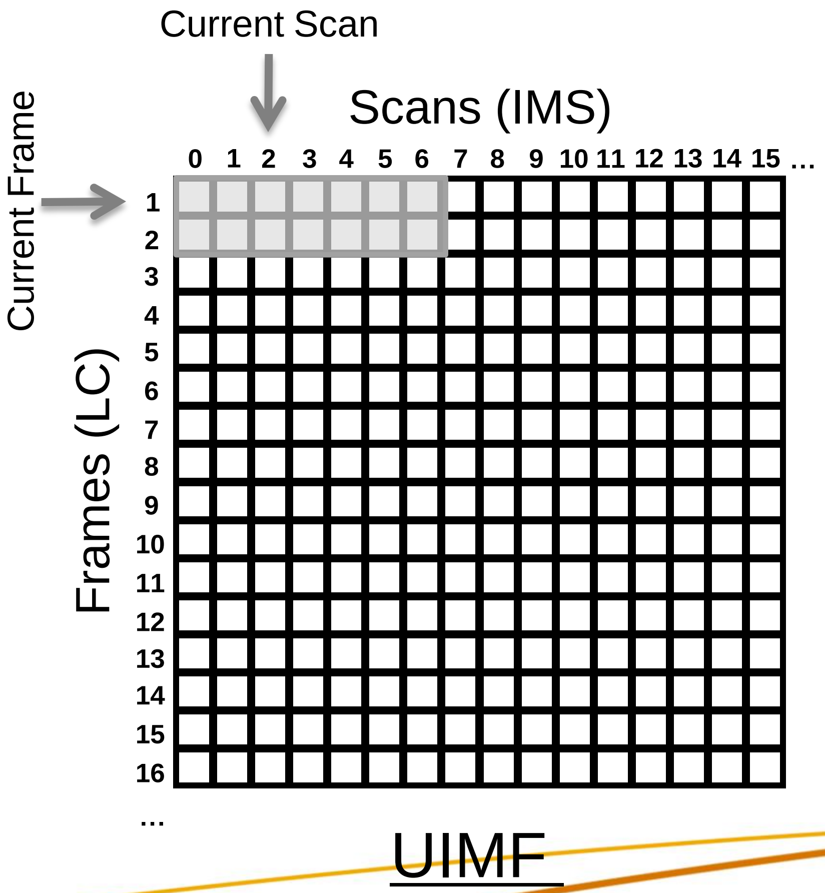
1. Sum window & deisotope data
2. Increment scan & repeat step 1



Decon2LS Deisotoping

Process:

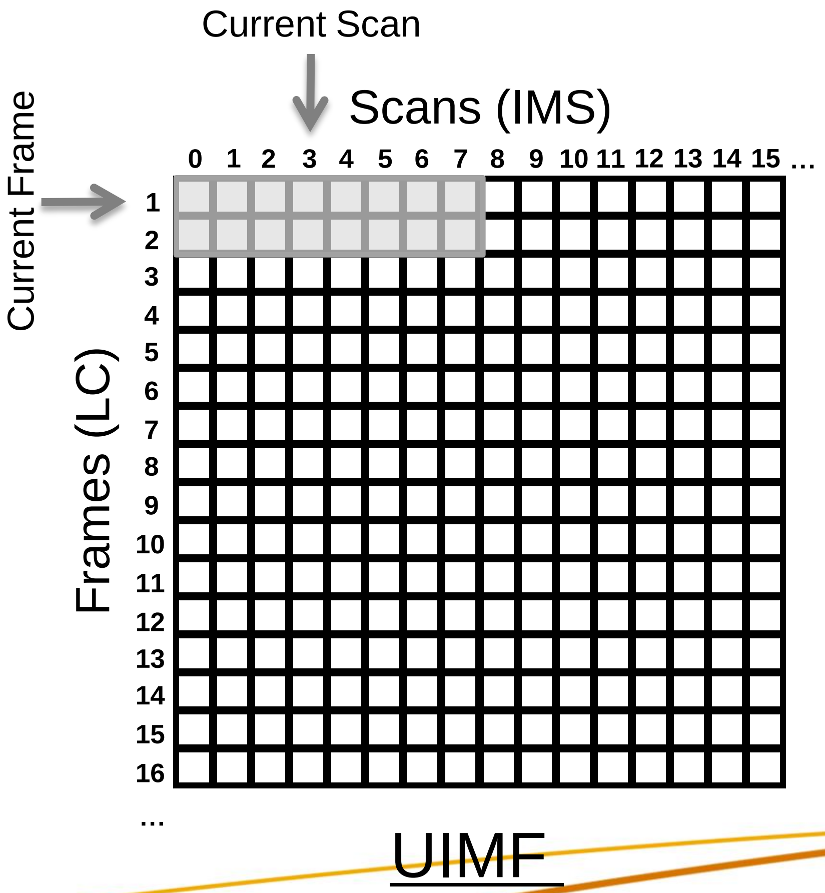
1. Sum window & deisotope data
2. Increment scan & repeat step 1
3. Iterate steps 1-2 over scan range



Decon2LS Deisotoping

Process:

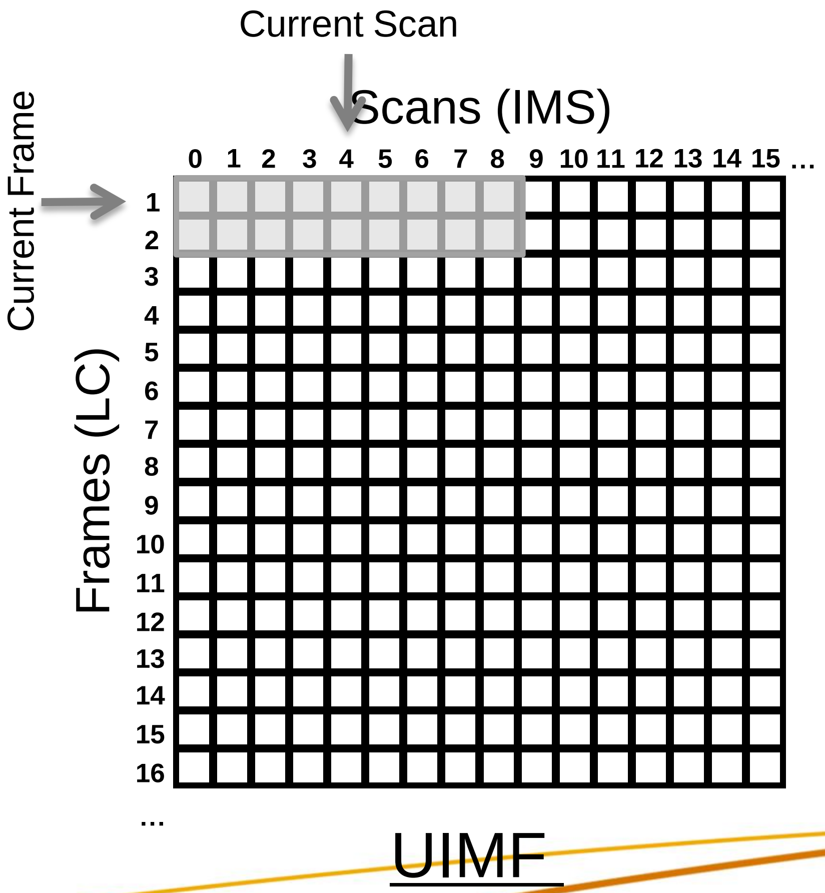
1. Sum window & deisotope data
2. Increment scan & repeat step 1
3. Iterate steps 1-2 over scan range



Decon2LS Deisotoping

Process:

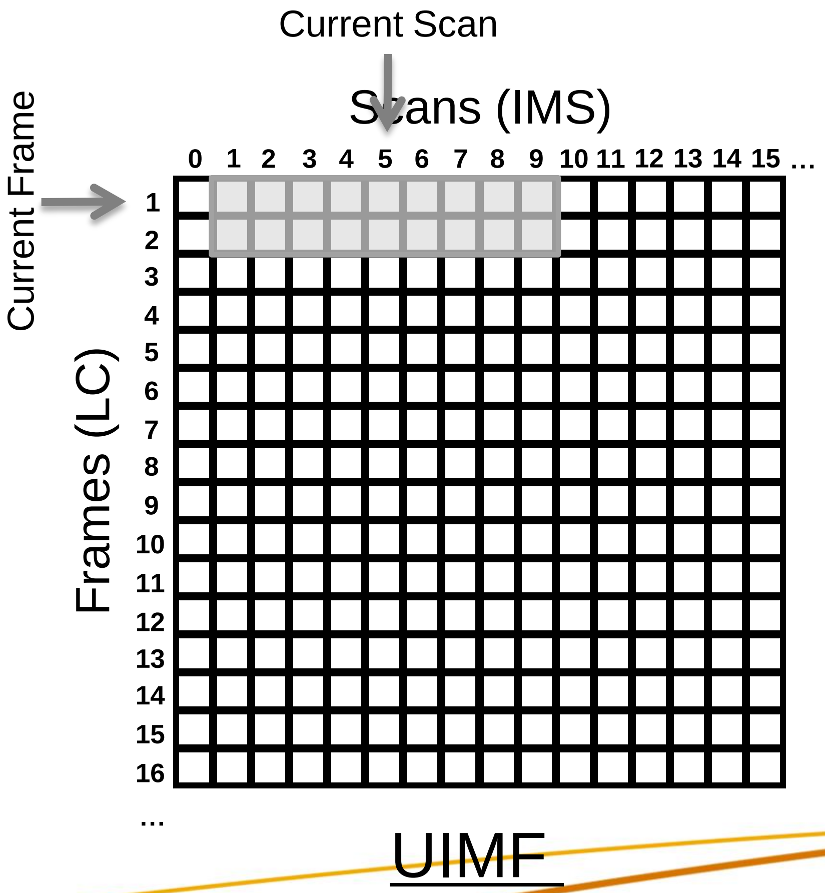
1. Sum window & deisotope data
2. Increment scan & repeat step 1
3. Iterate steps 1-2 over scan range



Decon2LS Deisotoping

Process:

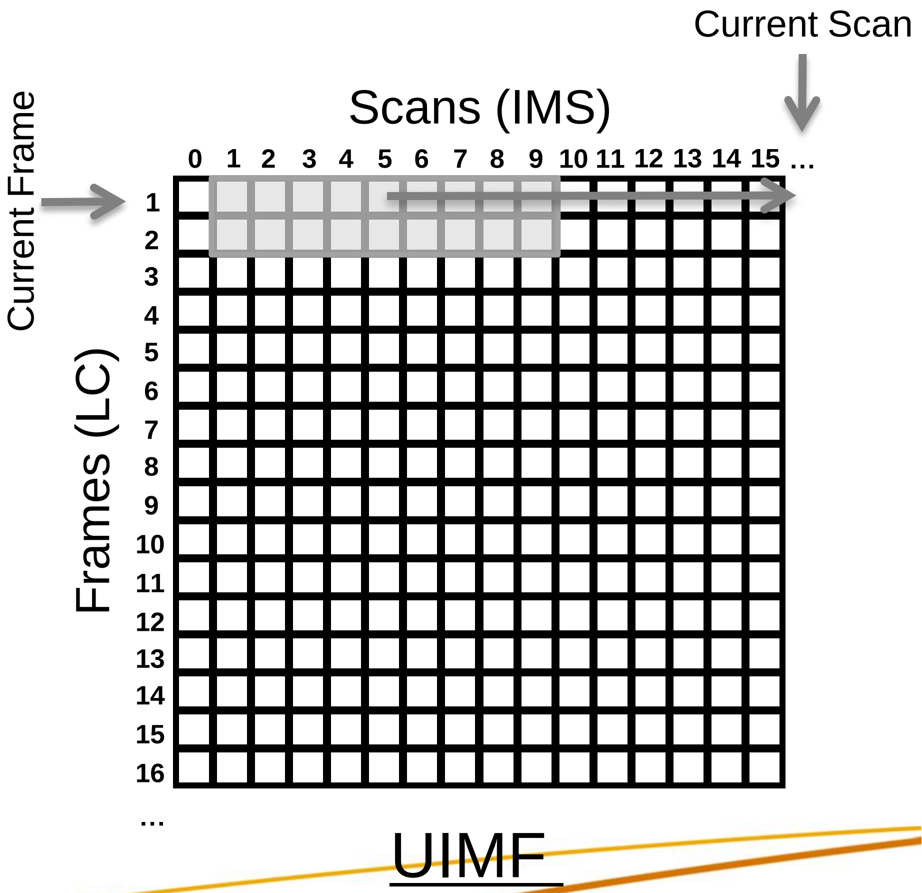
1. Sum window & deisotope data
2. Increment scan & repeat step 1
3. Iterate steps 1-2 over scan range



Decon2LS Deisotoping

Process:

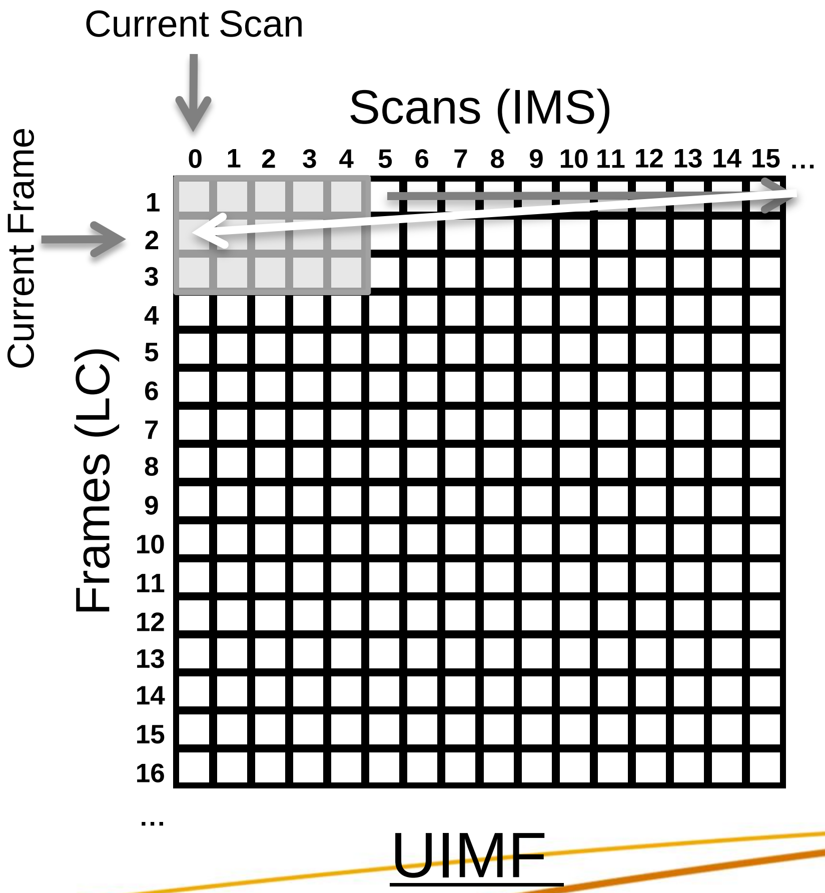
1. Sum window & deisotope data
2. Increment scan & repeat step 1
3. Iterate steps 1-2 over scan range



Decon2LS Deisotoping

Process:

1. Sum window & deisotope data
2. Increment scan & repeat step 1
3. Iterate steps 1-2 over scan range
4. Increment frame and return to 1st scan



Current Scan

Scans (IMS)

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 …

Current Frame

Frames (LC)

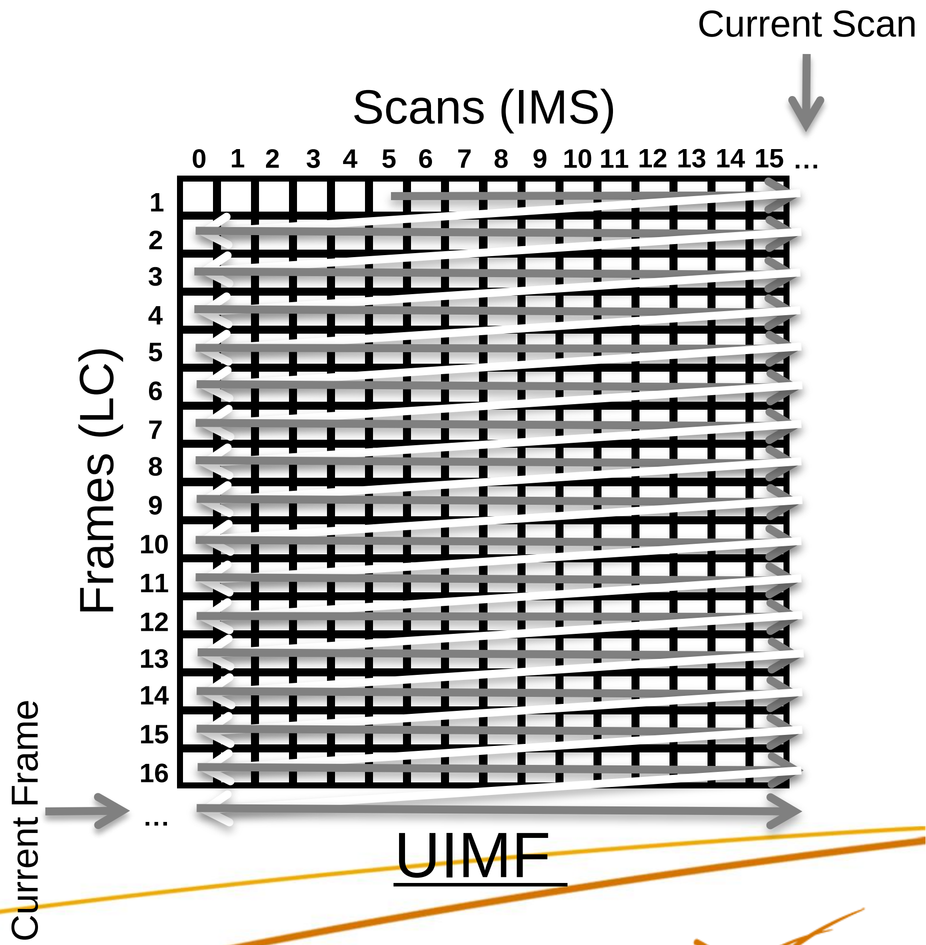
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 …

UIMF

Decon2LS Deisotoping

Process:

1. Sum window & deisotope data
2. Increment scan & repeat step 1
3. Iterate steps 1-2 over scan range
4. Increment frame and return to 1st scan
5. Repeat steps 1-4 until the whole UIMF file is traversed



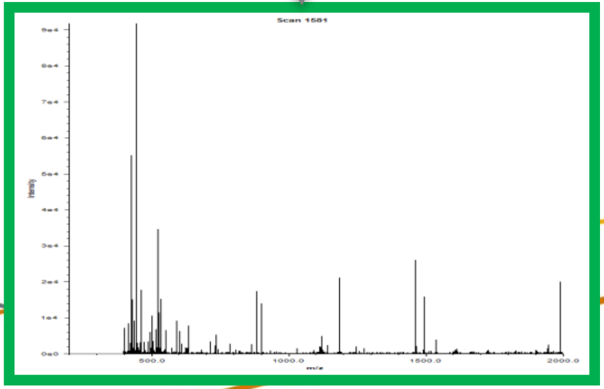
Shortcomings

- Re-reading of spectra data
  - Up to 6x required reads
- Reconverting bins to *m/z* values
  - Multiple calls to Math.Pow()
  - No dynamic programming

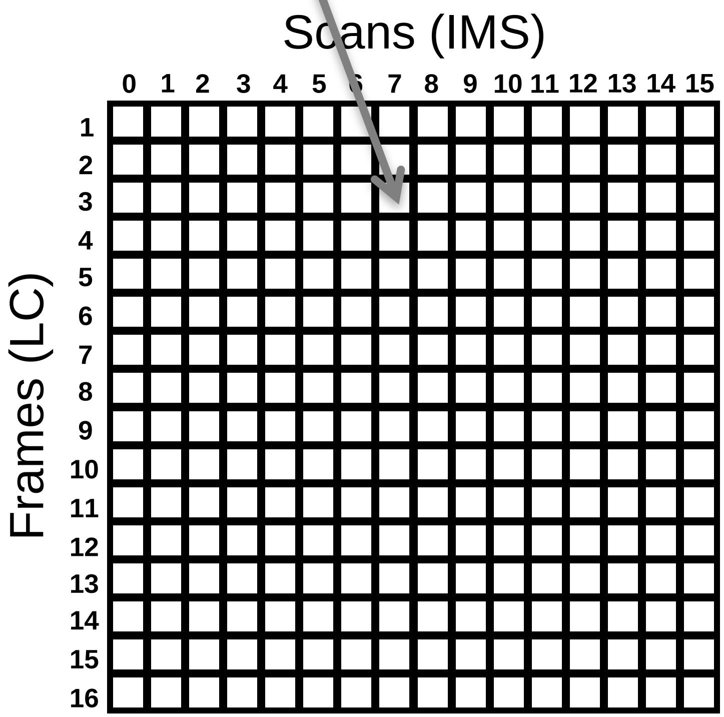
Raw Data:

Bin	Intensity
0	0
269328	6
269328	6
269328	6
0	0
298781	20
…	…

T = Bin * BinWidth / 1000



This data was re-read 27 times!!!



UIMF

UIMF Library Improvements

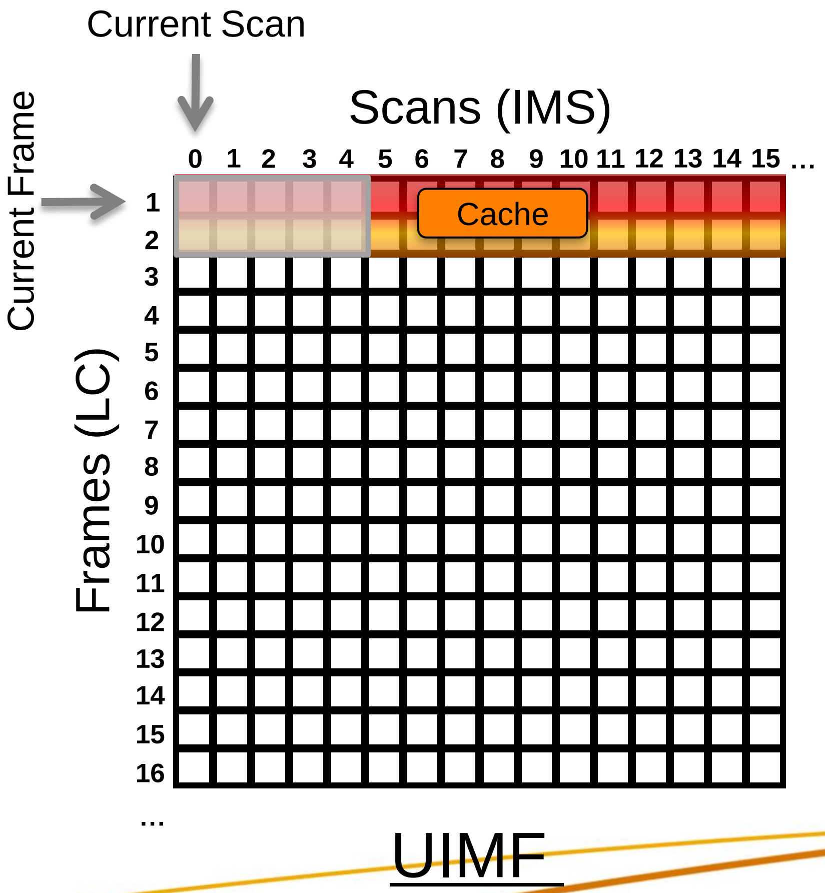
- Spectra caching
  - Data is only read once
  - All scans are cached per frame range
  - 2x List<List<int[]>>
    - Bins & Intensities

- Bins to *m/z* values caching
  - T values and powers calculated once
  - Dynamic programming implemented

Spectra Caching

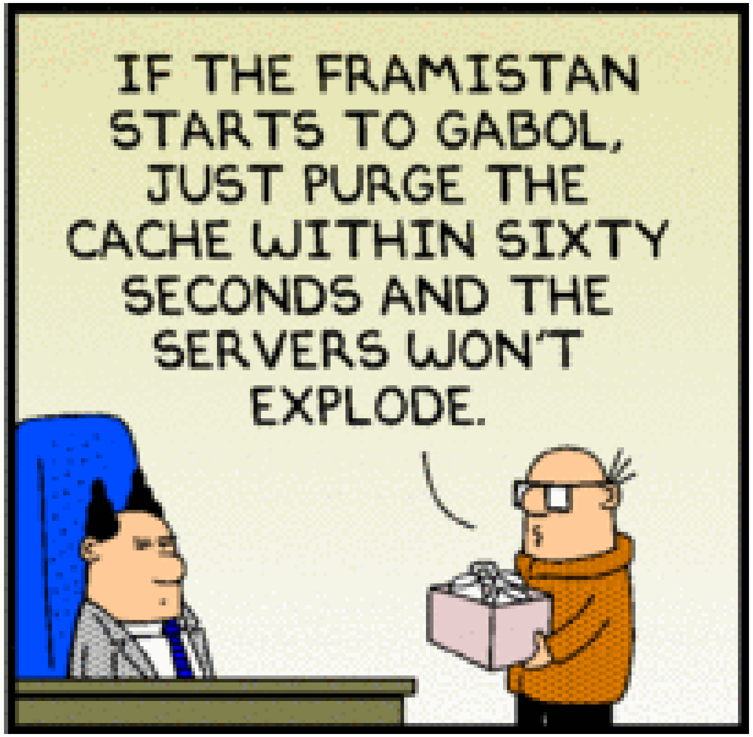
- First call creates cache
  - Frames queried one at a time
  - Spectra arrays append to list
  - Frames are added to a 2nd list

- As sliding window moves
  - Trailing frames are removed
  - New frames are added



Bins to m/z Values Caching

- Two-dimensional Array
  - Rows = Number of bins
  - Columns = t, t^3, t^5, t^7, t^9, t^{11}
- Dynamic Programming
  - Calculate t and t^2
  - $t^3 = t * t^2$
  - $t^5 = t^3 * t^2$
  - And so on…


IF THE FRAMISTAN STARTS TO GABOL, JUST PURGE THE CACHE WITHIN SIXTY SECONDS AND THE SERVERS WON'T EXPLODE.

Decon2LS Results

HSer_2pt0_420_100_c2_150um_fr560_Cheetah_0001

Bins: 138000

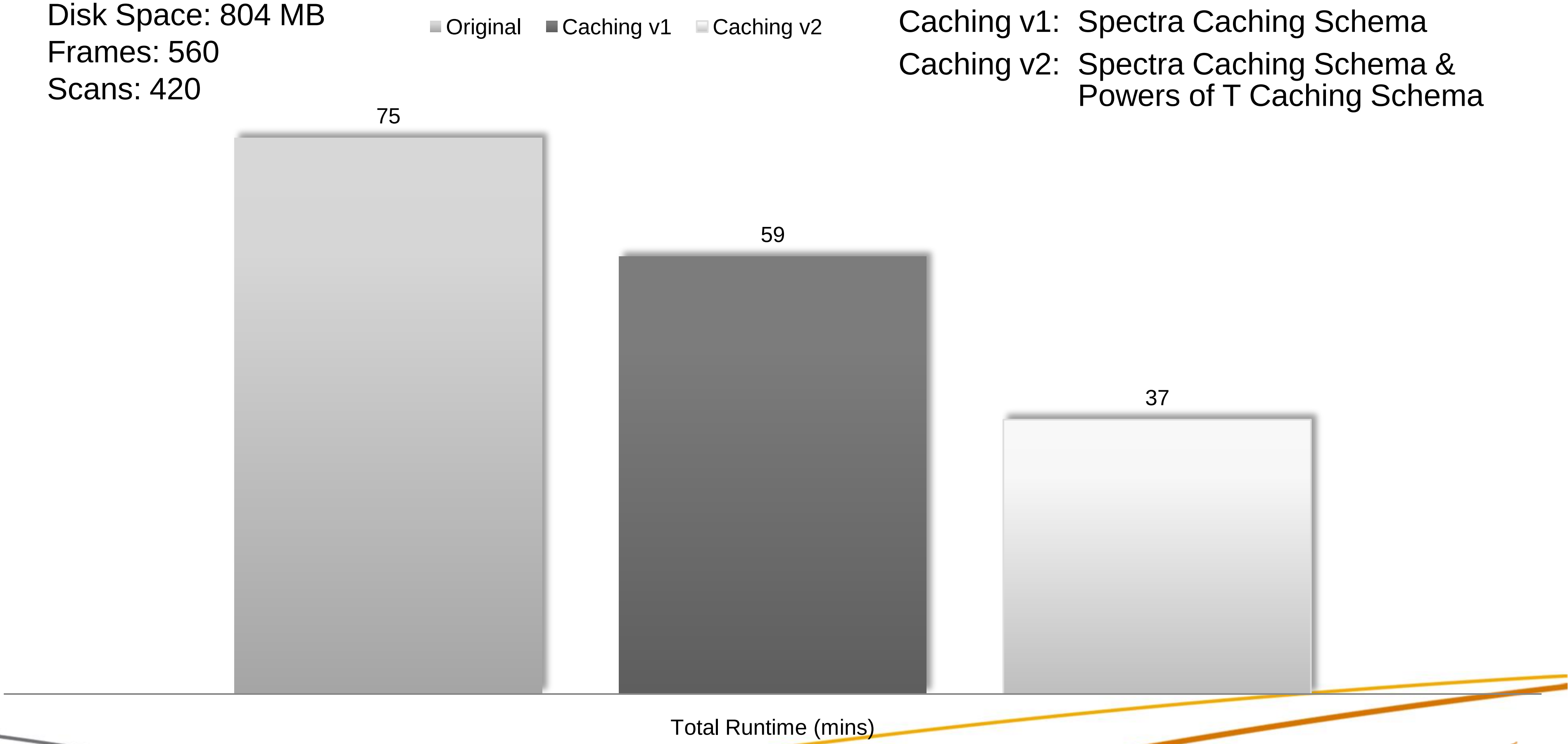
Disk Space: 804 MB

Frames: 560

Scans: 420

Caching v1: Spectra Caching Schema

Caching v2: Spectra Caching Schema & Powers of T Caching Schema

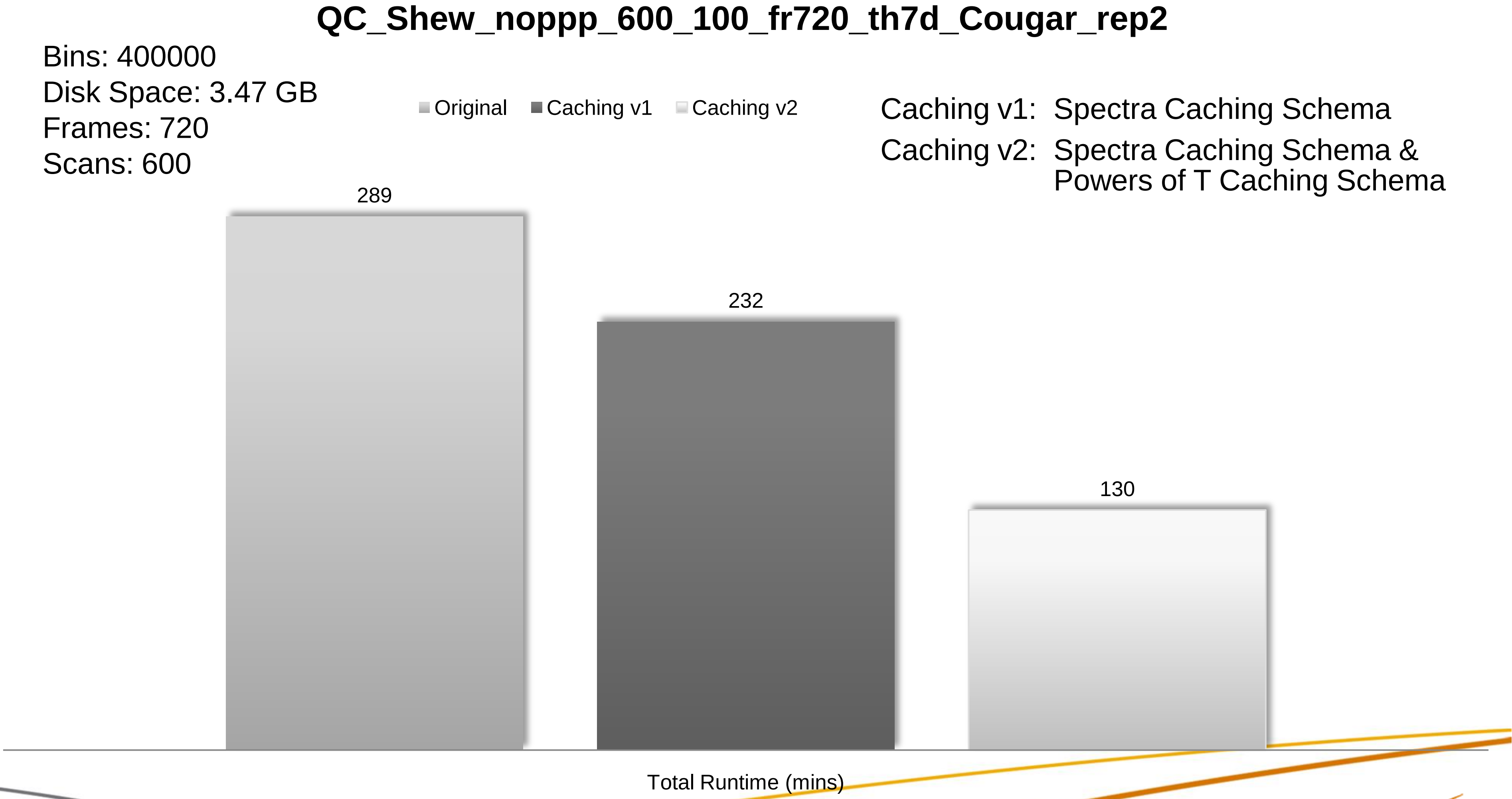


Decon2LS Results

Bins: 400000
Disk Space: 3.47 GB
Frames: 720
Scans: 600

Caching v1: Spectra Caching Schema

Caching v2: Spectra Caching Schema & Powers of T Caching Schema



Tradeoffs

- Specific data access pattern
  - Random access
    - Invalid output
- Addition memory requirement
  - Spectra cache: ~64.6 MB
  - T values cache: ~18.3 MB

Acknowledgements

Informatics:

- Anuj Shah
- Kevin Crowell
- Gordon Slysz
- Brian LaMarche
- Gordon Anderson

SULI Program:

- Karen Wieda
- Frances Skomurski

Funding:

U.S. DEPARTMENT OF ENERGY