Environmental Chemistry through Intelligent Atmospheric Data Analysis (EnChIlADA): A Platform for Mining ATOFMS and Other Atmospheric Data

Katie Barton, John Choiniere, Melanie Yuen, and Deborah Gross
Department of Chemistry, Carleton College

Anna Ritz, Thomas Smith, Leah Steinberg, and David Musicant
Department of Mathematics and Computer Science, Carleton College

Jamie Schauer
Environmental Chemistry and Technology, University of Wisconsin – Madison

Lei Chen, Greg Cipriano, and Raghu Ramakrishnan
Department of Computer Sciences, University of Wisconsin – Madison

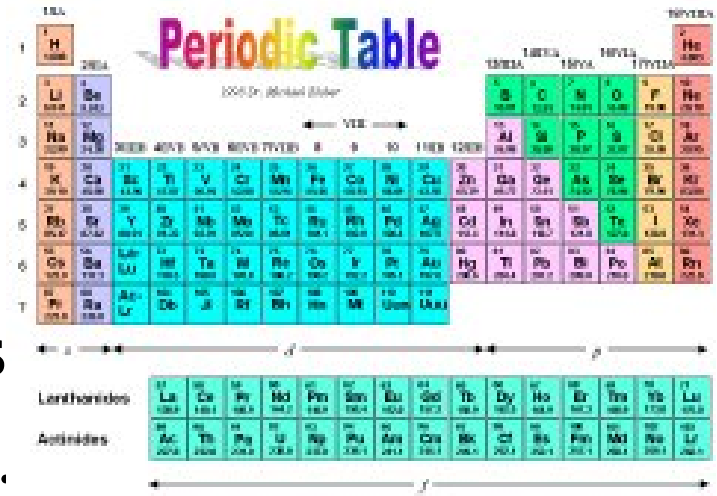
Carleton College

Today's Menu

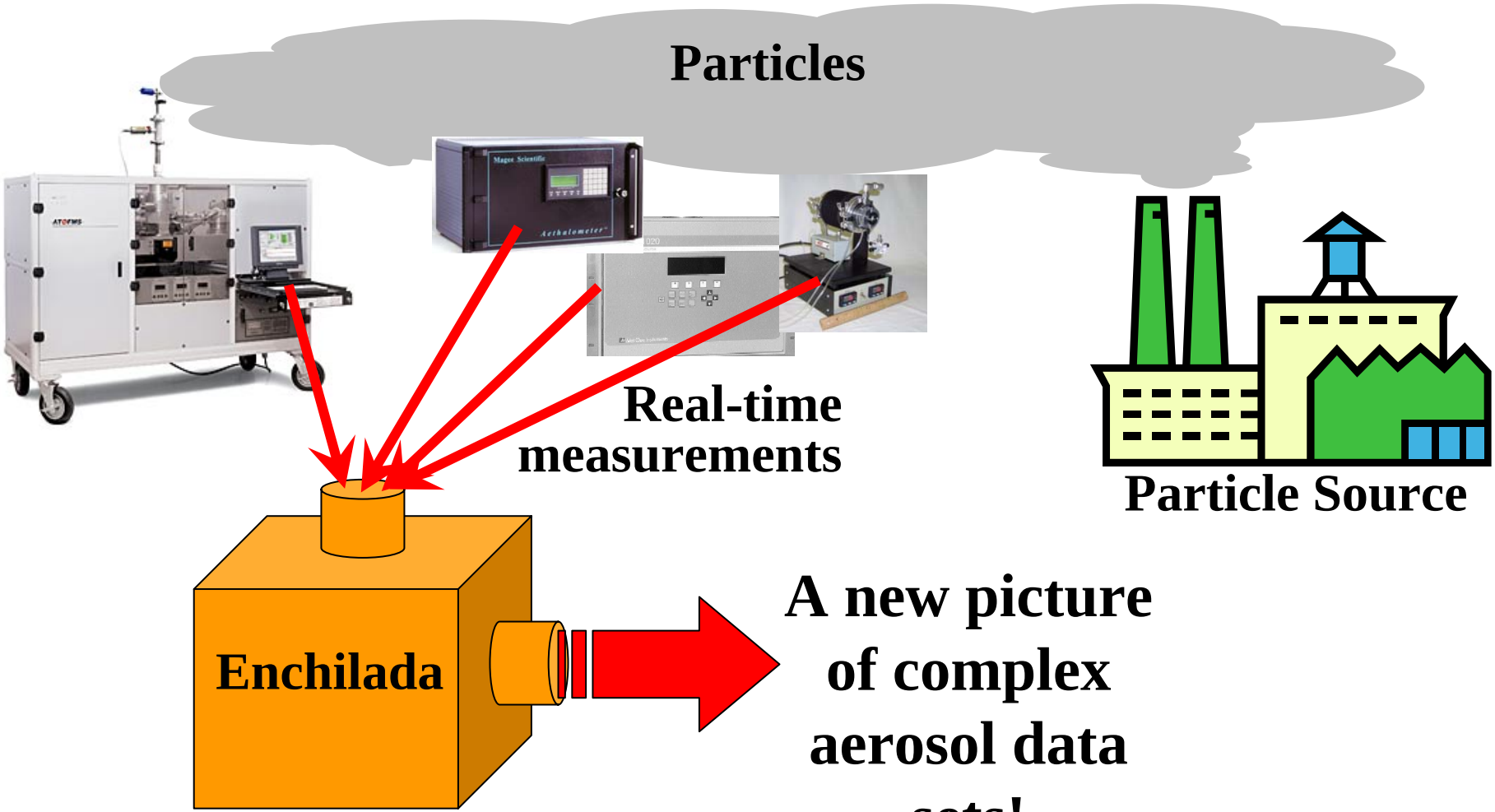
- Goals of Enchilada project
- Discussion of capabilities (current and future)
  - Data organization
  - User Interface and Analysis Tools
  - Clustering Algorithms
  - Time-series Analysis
  - Labeling ATOFMS Spectra Automatically
- Summary and Future Outlook

Project Description

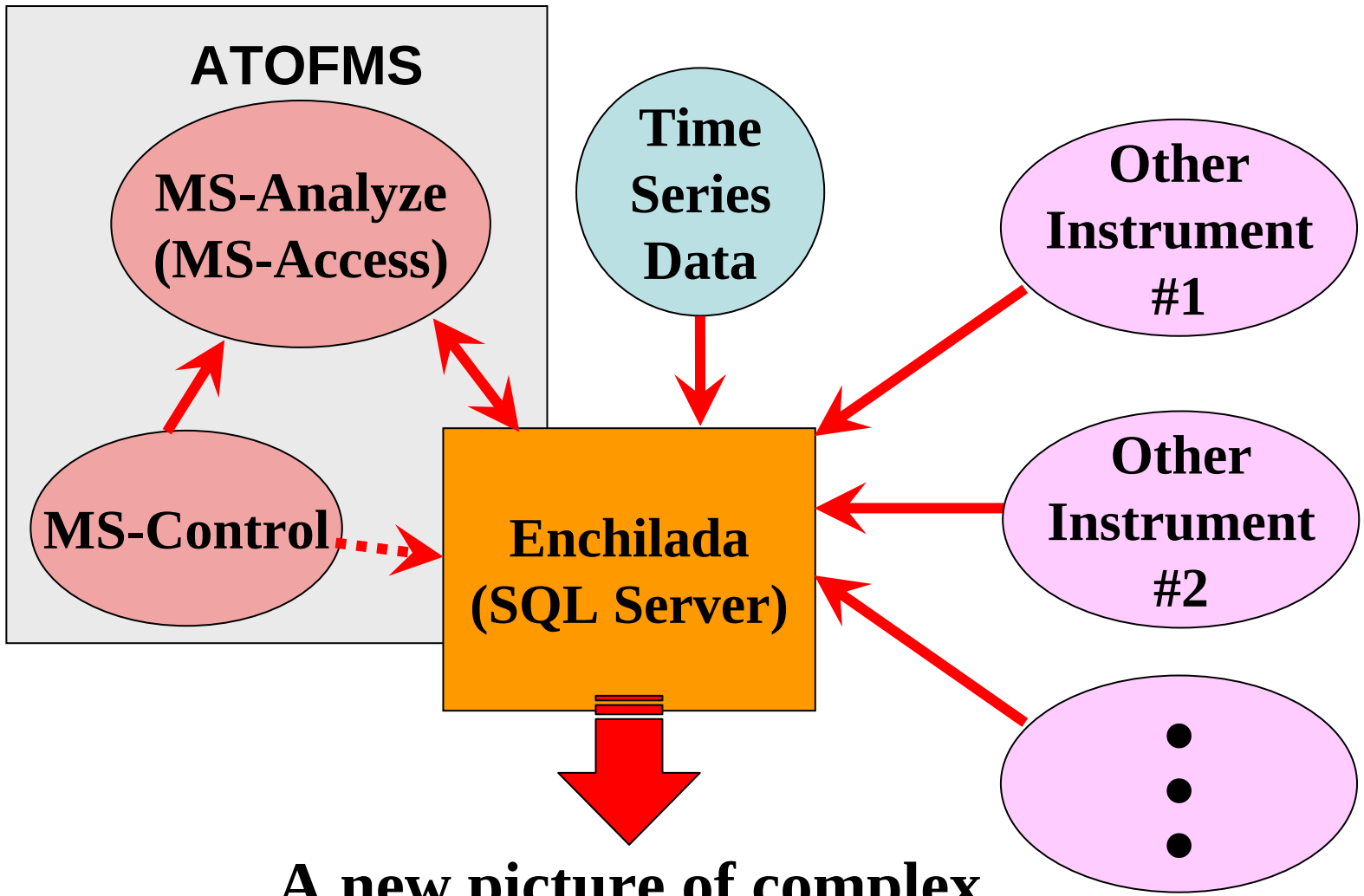
- NSF-ITR funding to develop data mining techniques for complex aerosol data sets.
- Collaboration between aerosol scientists and computer scientists:
  - Address the right questions.
  - Integrate “domain knowledge” into the data mining techniques
  - Use real data to test techniques.
  - Answer interesting questions along the way!



The Aerosol Scientist's View of Enchilada:



How Enchilada Fits In



A new picture of complex aerosol data sets!

What is Data Mining?

- "The non-trivial discovery of novel, valid, comprehensible and potentially useful patterns from data" (Fayyad et al)
- "Computer! Learn something from this data, and explain it to me."
- Includes…
  - Clustering
  - Classification
  - Time series analysis

Enchilada

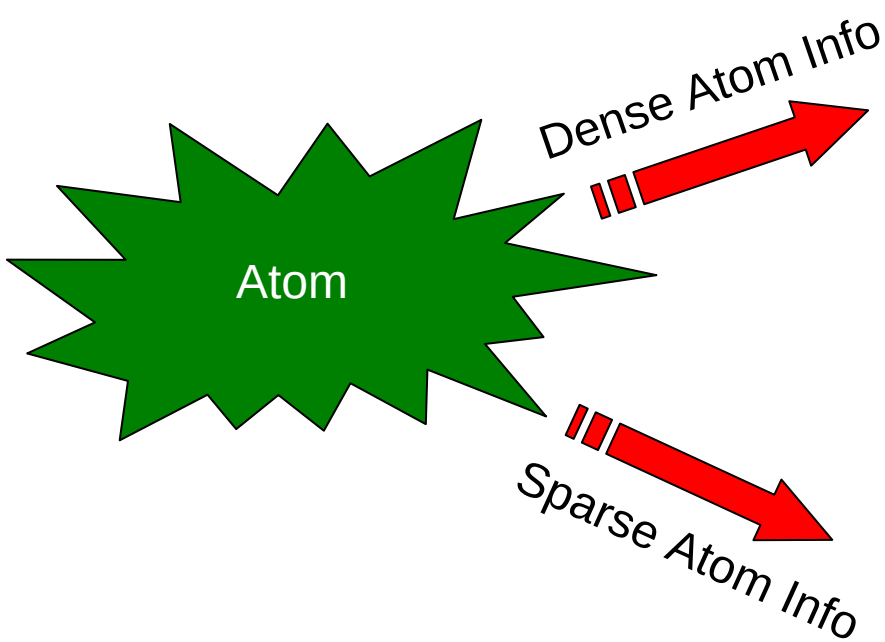
- Open Source Software
- Implements data mining tools to analyze atmospheric data (in real time)
- General
  - Will work with any relevant data, not just ATOFMS data
- Scalable
  - Will work with data that contains millions of particles
- Easy to use, intuitive

Enchilada Basics

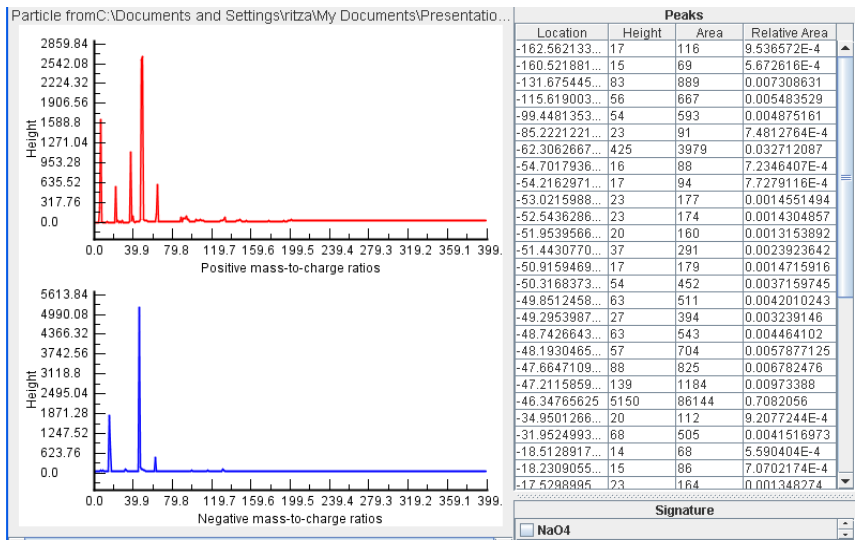
- Import ATOFMS datasets
- Organize datasets into collections
- Display particle spectra
- Query (time, size, count)
- Synchronize, Cluster, Label
- Export collections to MS-Analyze

Particle Organization in Enchilada

- Atoms
  - Not atoms in a chemistry sense!
  - Smallest units of imported data (for our purposes, particles)

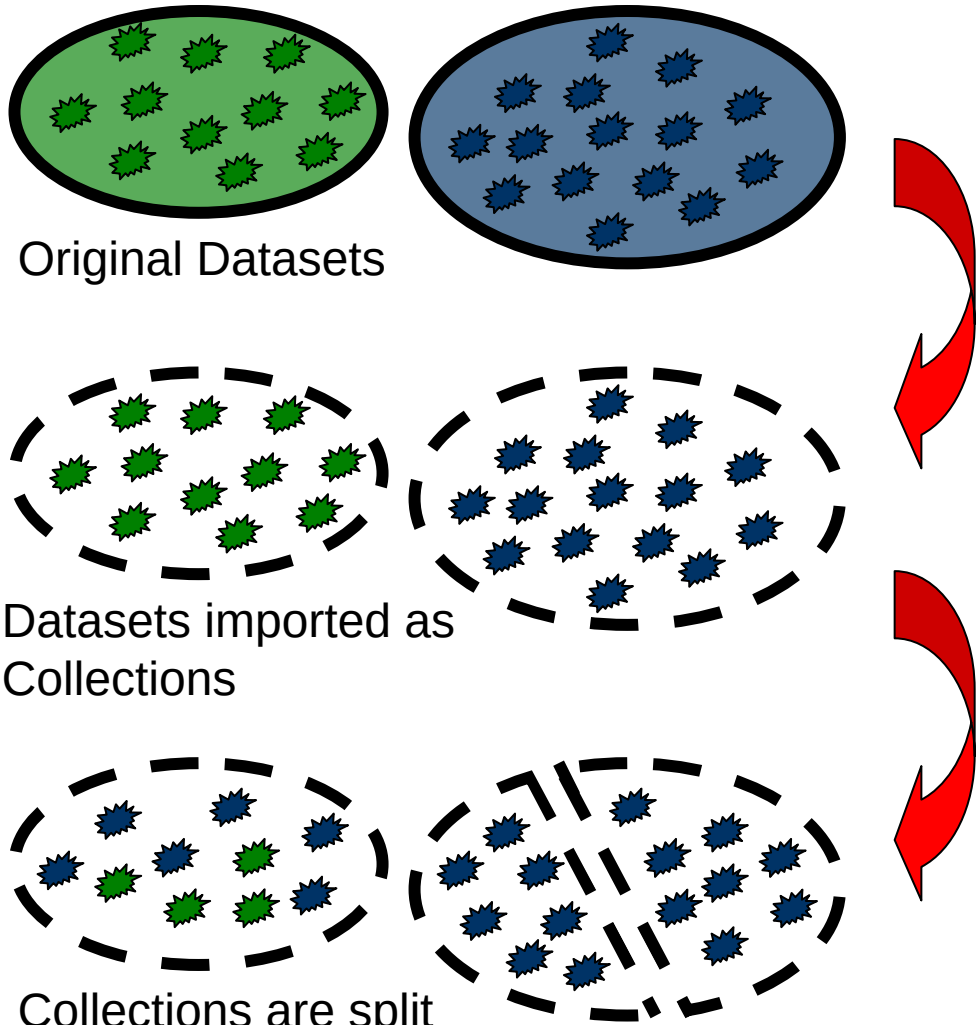


Time = 08:03:03 12/03/03
Size = 0.21 microns
Filename = particle1.amz



Dataset Organization in Enchilada

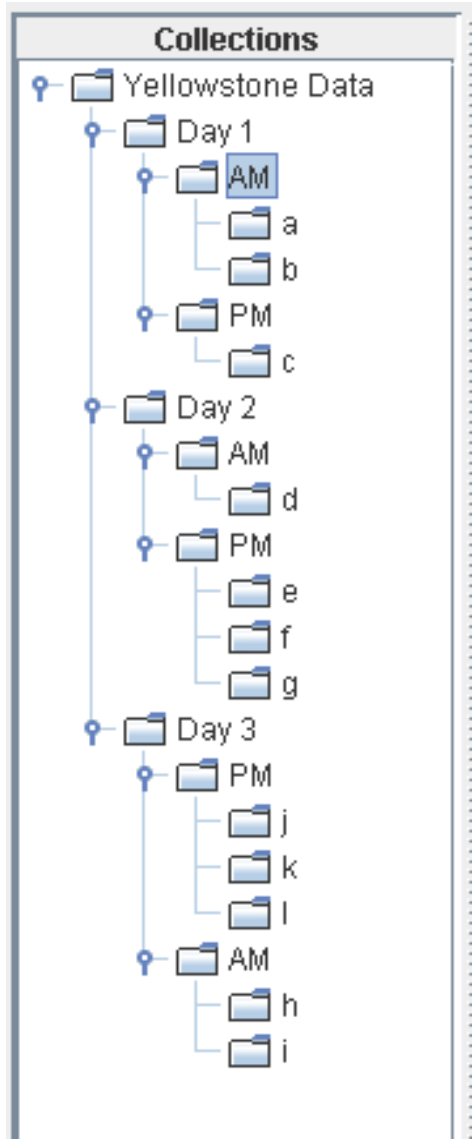
- Datasets
  - Groups of particles
  - Gathered from ATOFMS instrument in one session
  - Each atom belongs to one dataset
- Collections
  - Virtual datasets
  - Collections can be made from other collections
  - Each atom can belong to many collections



(Demo of Enchilada Tools)

The Collection Hierarchy

- Provides a way to view all imported datasets at once
  - Can be quite complex!
- Contains all the atoms from their subcollections
- Has the ability to create empty collections
- Has the ability to copy, cut, and paste collections



Datatypes in the Database

- *datatype*: the kind of data the program is working with
  - Examples: ATOFMS, TimeSeries, AMS, Mercury data
- Multiple datatypes give Enchilada power
  - Easy comparisons between instruments or over time
- Different datatypes in one database
  - Database needs to be detailed **and** generalized

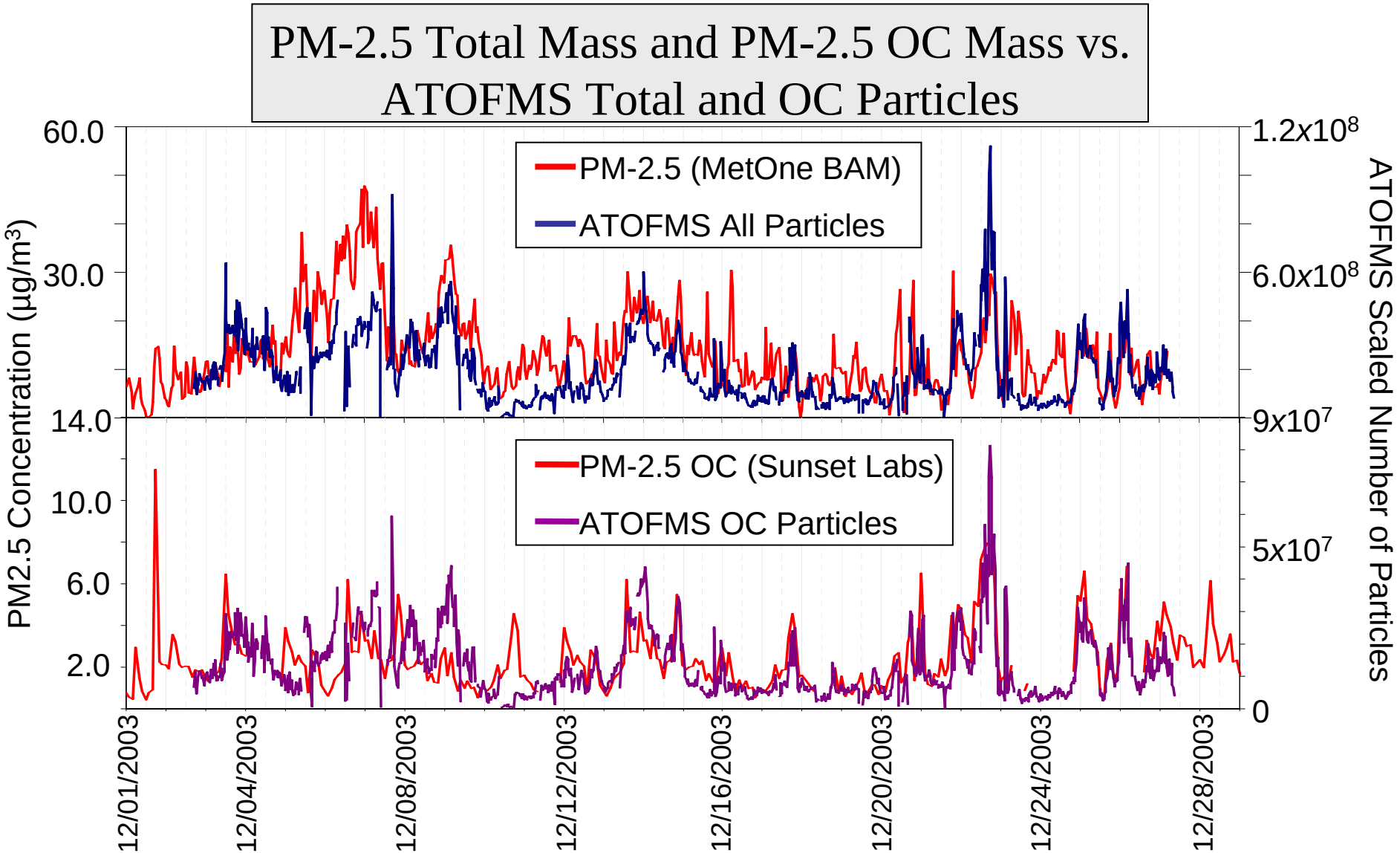
Database Representation

- Tables in the database:
  - *General tables* for information that is consistent across datatypes
  - *Datatype-specific tables* for information specific to each datatype
- Only one set of general tables in the entire database
- One set of datatype-specific tables for *each* datatype in the database

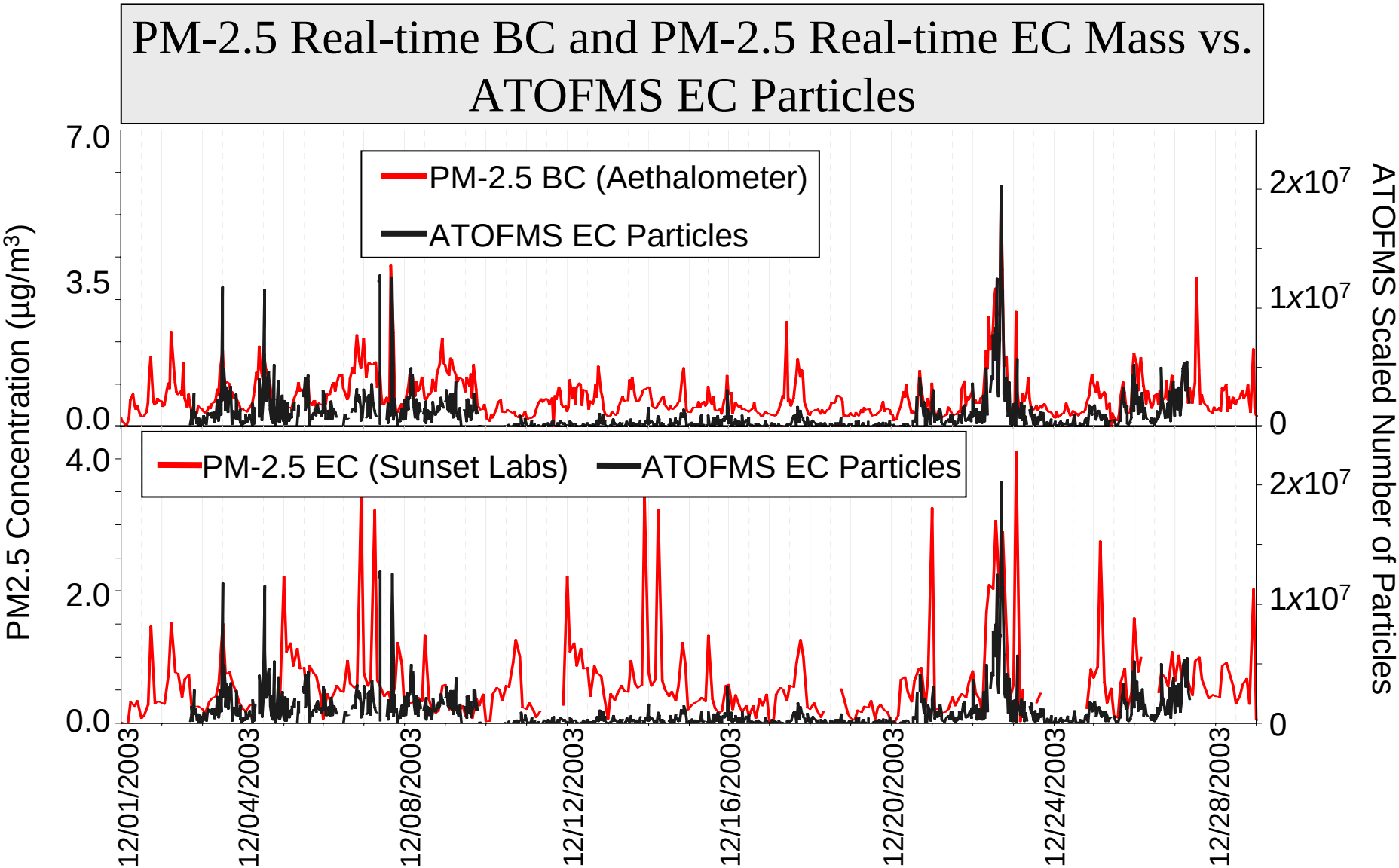
Time-Series Analysis

- Observe trends over many datasets.
  - Provide a broader picture.
  - Elucidate dynamics in data series.
- Compare different types of time series data.
  - Most other data of interest are time series.
  - Time steps need to be synchronized.

East St. Louis Time-Series



East St. Louis Time-Series



EC vs. BC by Chemical Composition?

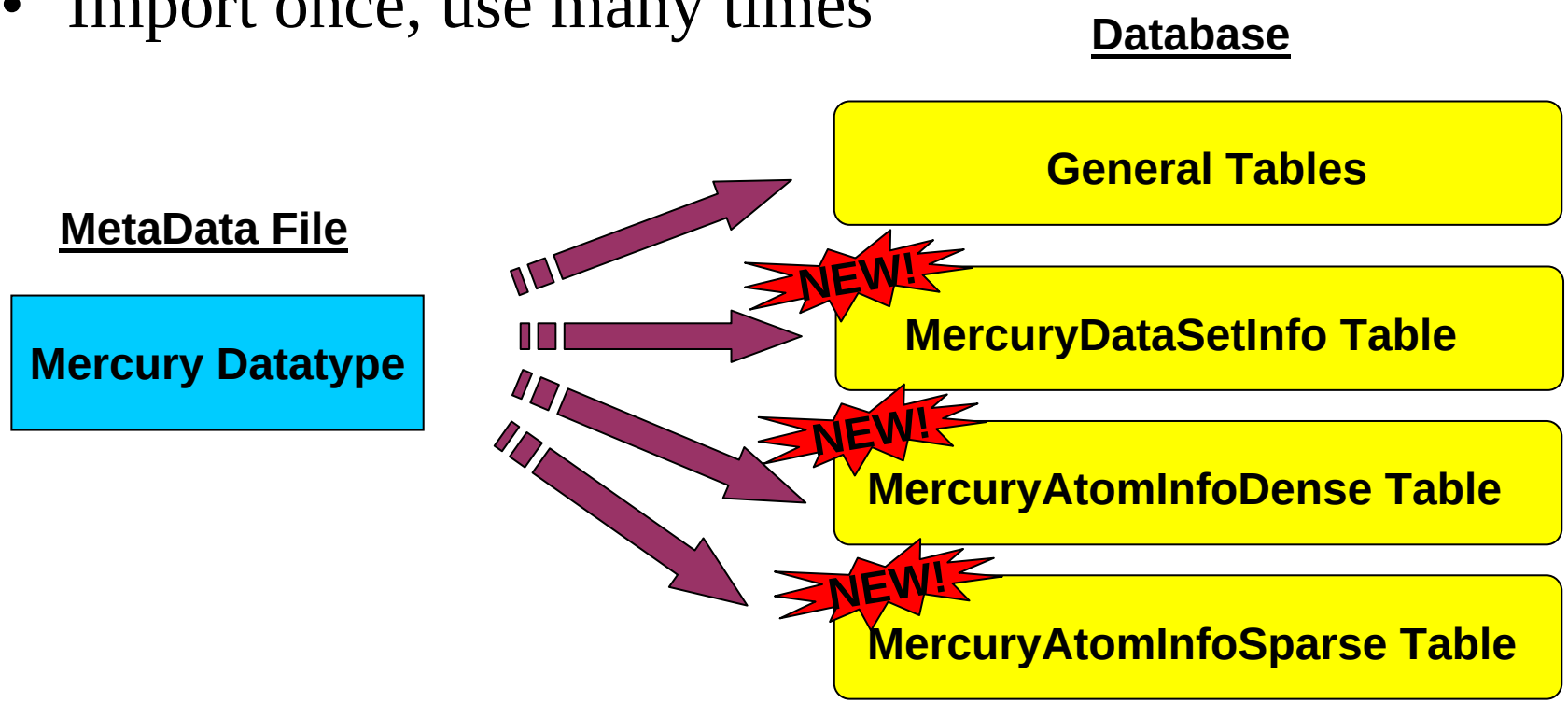
- Goals of this analysis are to use data mining to:
  - Determine the relationship(s) between EC and BC, using composition information from ATOFMS.
  - Understand which chemical components contribute to light-absorbing properties of particles.
- We want to understand these questions without using prior knowledge. Enchilada will search for the relationship(s).

Flexible Importation

- How do we get all these different datatypes into Enchilada?
- Customized XML file formats for importing datatypes and data
  - XML (Extensible Markup Language) is an industry standard for exchanging data.
- Any datatype that can be represented in these file formats could be imported.

MetaData Files

- Contain information about a datatype
- Used to create new datatype-specific tables in the database
- Import once, use many times



EnchiladaData Files

- Contain the actual data to import into the program
- XML example:

```
<atominfodense>
    <field>ExampleParticle</field>
    <field>2003-12-03 08:03:03</field>
    <atominfosparse table="Peaks">
      <field>12</field>
      <field>.02568247</field>
    </atominfosparse>
</atominfodense>
```

(Demo of Enchilada Importer & Time Series)

Clustering: A Data Mining Technique

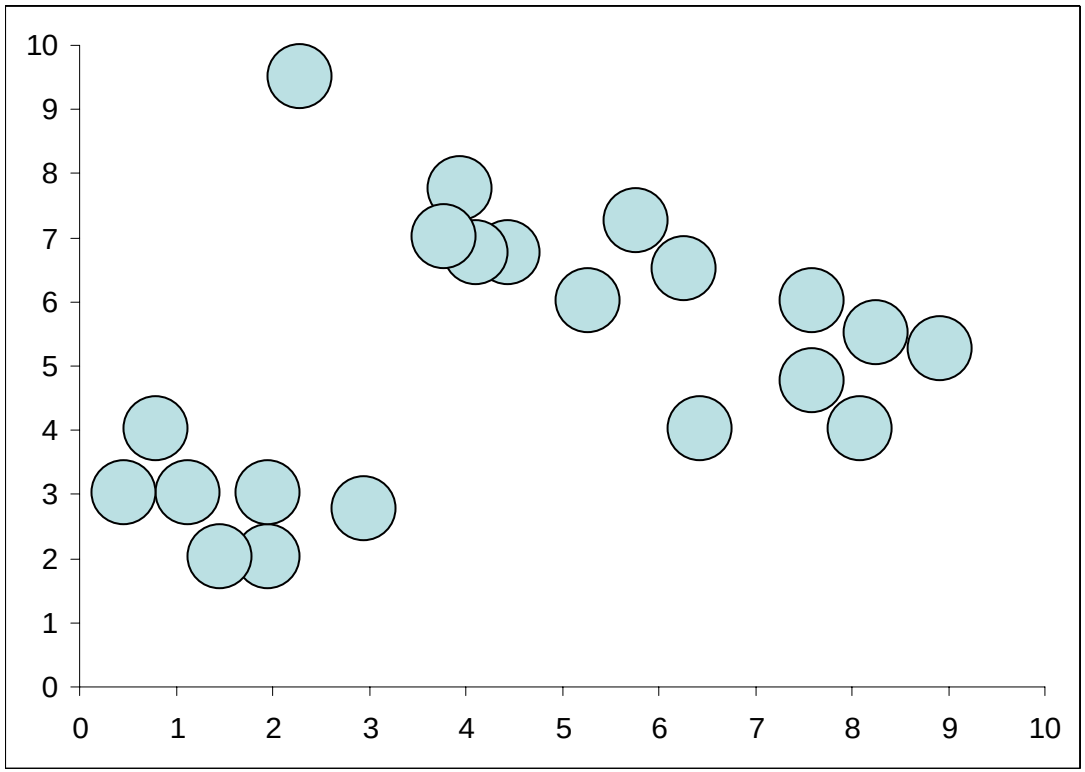
- Answers the question *"What are the main groups of data points in this database?"*
- For chemists: a fast way to find out about the major types of particles present in a sample.
- With time series aggregation, we can explore simultaneously clustering different datatypes.

Clustering

The purpose of clustering is to find groups of similar objects.

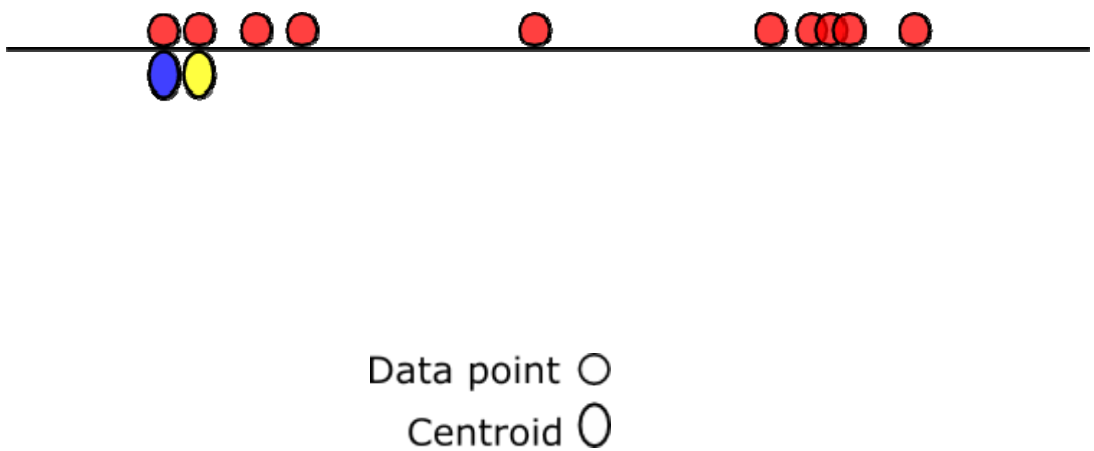
How many groups are in this graph?

- 4 clusters
- 3 clusters
- 2 clusters
- 1 cluster
- Or lots of clusters



k-Means Demo in One Dimension

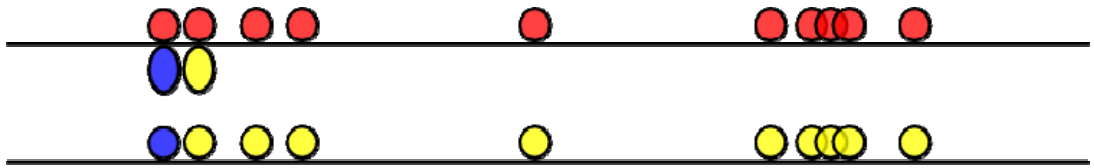
Step 1.
Choose initial cluster centers somehow (here, the first two points are chosen arbitrarily).



There are better methods of choosing initial cluster centers in the Data Mining literature.

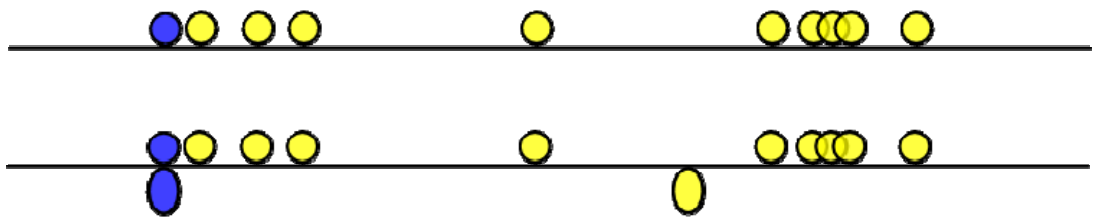
k-Means Demo in One Dimension

Step 2.
Assign each point to the centroid to which it is closest. Each group of points is now a cluster.



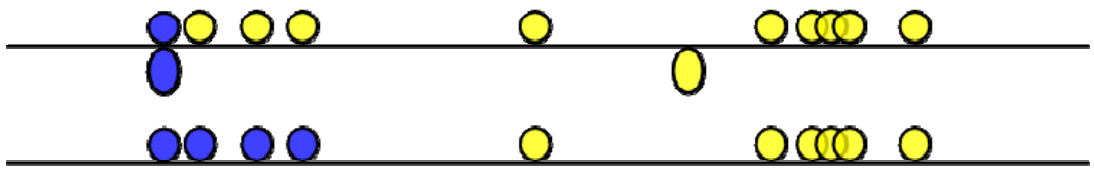
k-Means Demo in One Dimension

Step 3.
Average all the points in each cluster to find the centroid.



k-Means Demo in One Dimension

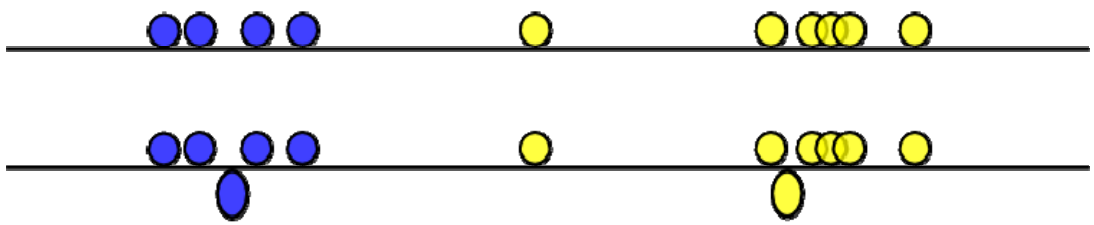
Step 2 again.
Assign each point to its nearest centroid.



Note that three points change which cluster they belong to.

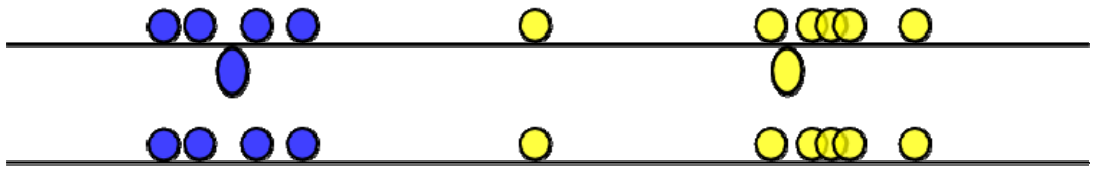
k-Means Demo in One Dimension

Step 3 again.
Average all the points in each cluster to find the new centroid.



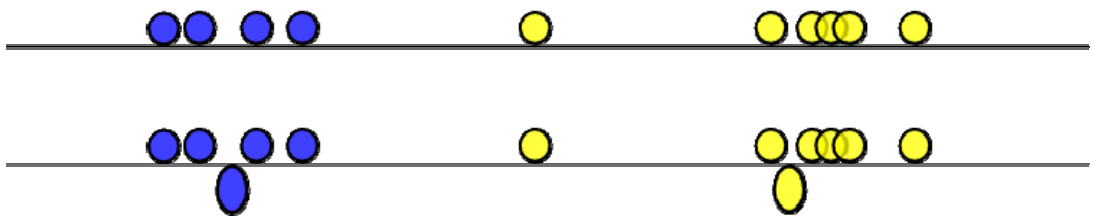
k-Means Demo in One Dimension

Step 2 again.
Assign each point to its nearest centroid.



k-Means Demo in One Dimension

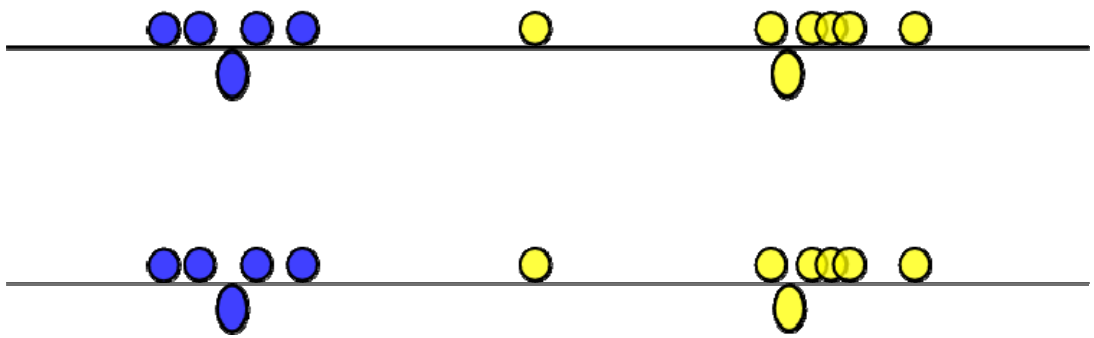
Step 3 again.
Average all the points in each cluster to find the new centroid.



k-Means Demo in One Dimension

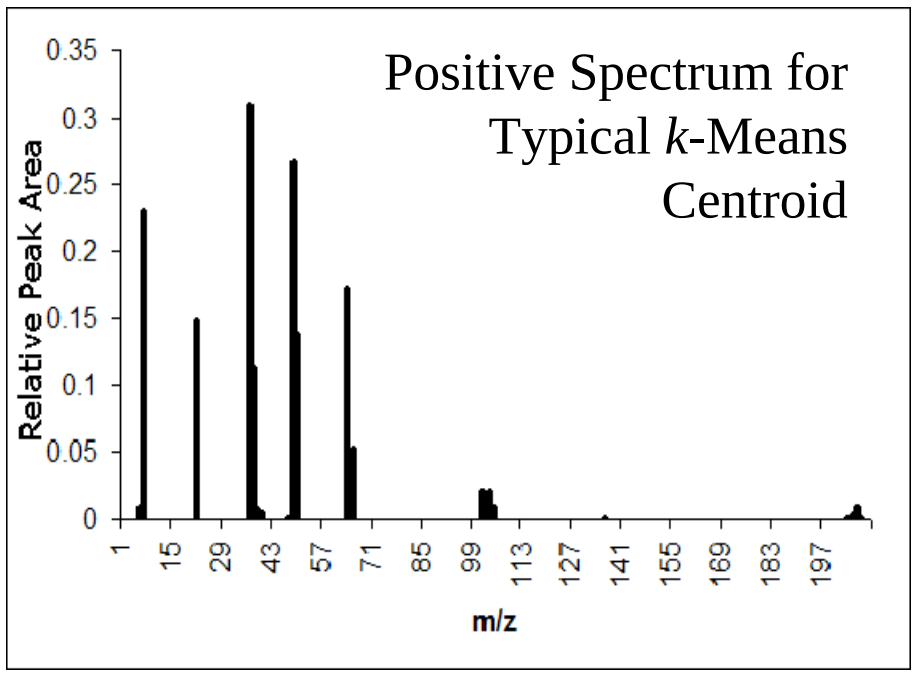
Step 4.
Notice that the centroids in the last two averaging steps were the same.

This means that the algorithm is done.



Clustering Spectra

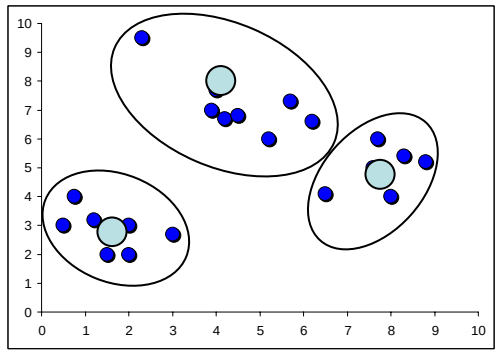
- Treat each *m*/*z* value as a dimension (so that we store information on the relative concentrations of each ion)
- Over 600 dimensions
- Apply the same algorithms we used for 2D data
- Cluster centers still summarize the particles



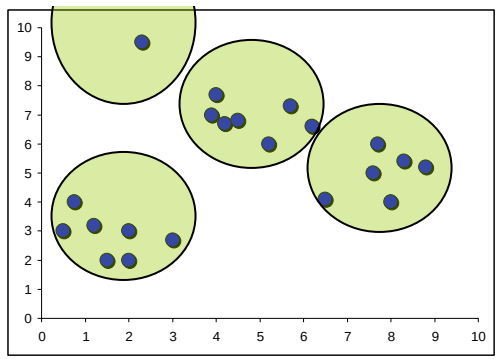
(Demo of Clustering in Enchilada)

Different Clustering Approaches

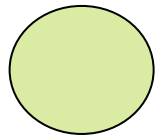
- *k*-Means/*k*-Medians
  - Cluster # parameter, use mean or median to find centers
- ART-2a
  - "Vigilance" parameter: cluster radius
  - Considers particles one at a time, rather than in a group.
  - Has been used with ATOFMS before



k = 3

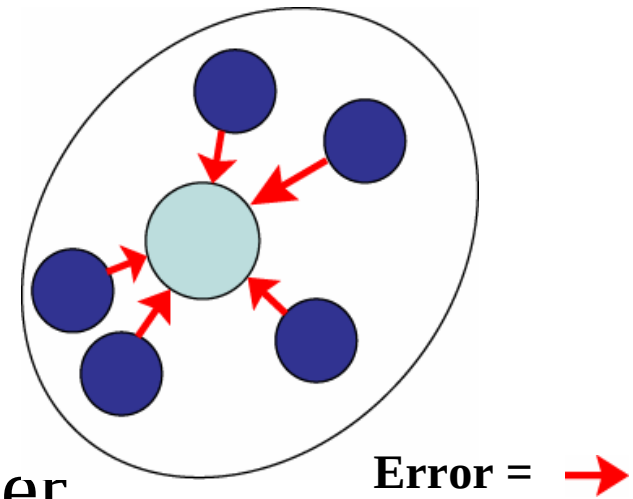


vigilance =



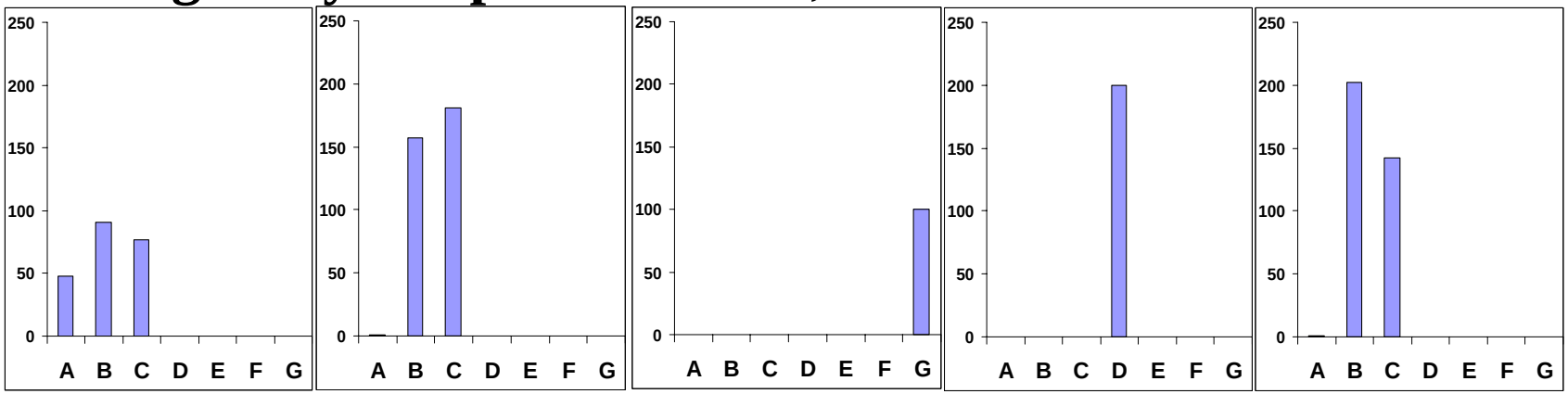
Clustering Experiment #1

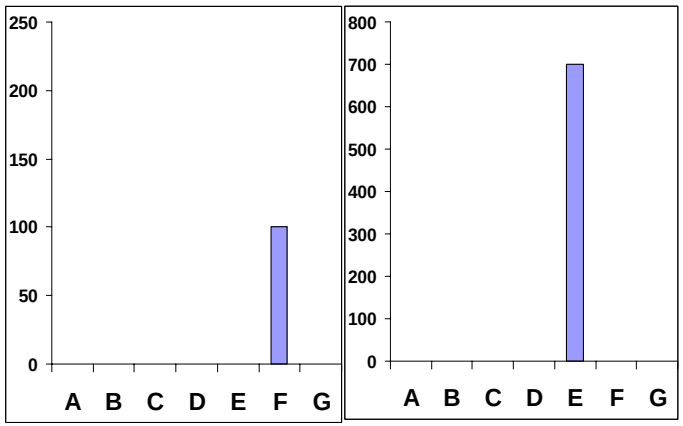
- How good is a clustering algorithm?
  - Synthetic Dataset
    - 2000-particle dataset with 7 real particles
    - We know what the clusters should look like
  - Error
    - The average distance from every point in a given cluster to its center

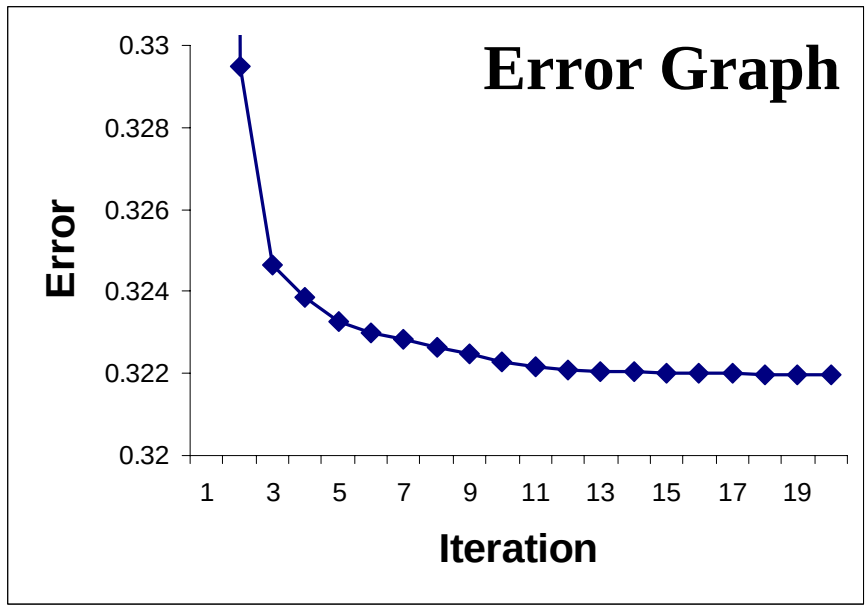


Quantitative Analysis of Clustering

Homogeneity Graphs: *k*-Means, *k* = 7







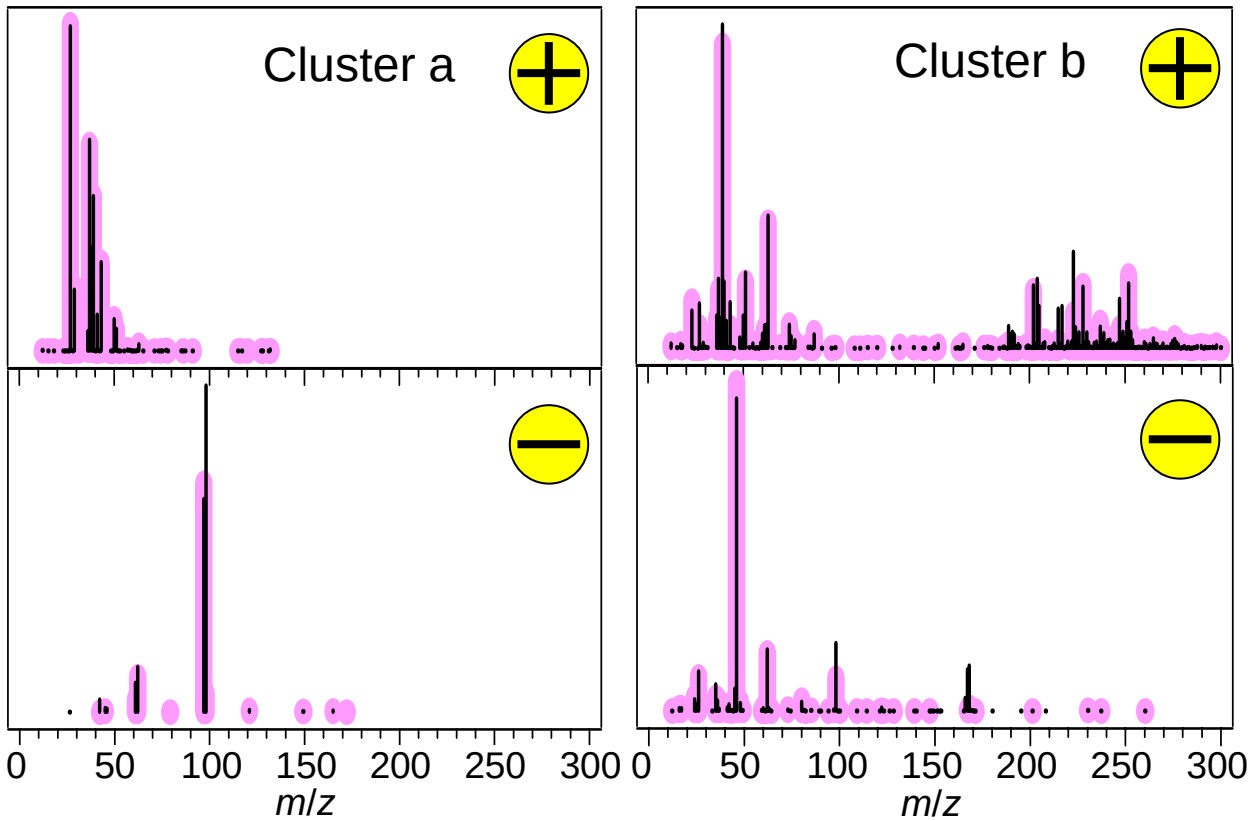
Clustering Experiment #2

- ART-2a versus *k*-Means
  - ART-2a has been used before for ATOFMS data
  - *k*-Means is a more standard clustering algorithm
    - More user-friendly than ART-2a
- This experiment:
  - ~2000 particles sampled in East St. Louis
  - Same number of clusters from each algorithm
  - Compare cluster centroids to evaluate algorithm performance

Chemical Comparison of Clustering

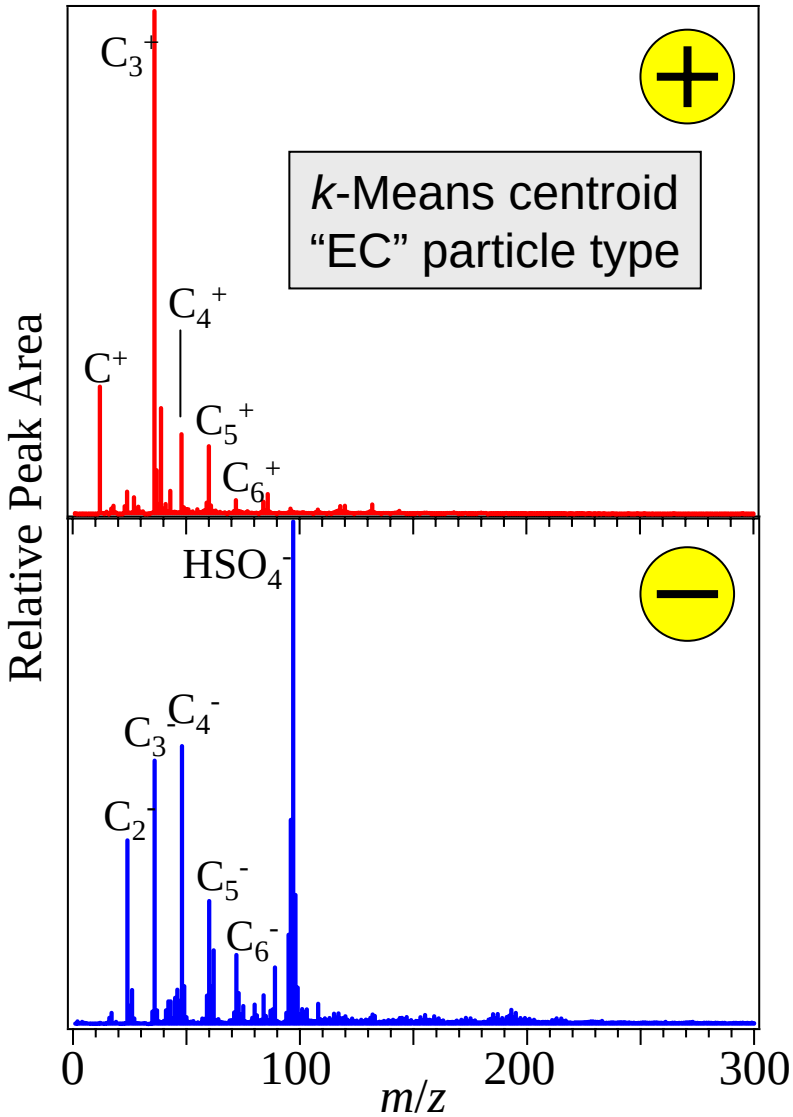
- ART-2a and *k*-Means are comparable
- Clusters make chemical sense
- *k*-Means calculations are more efficient
- Other algorithms will also be compared

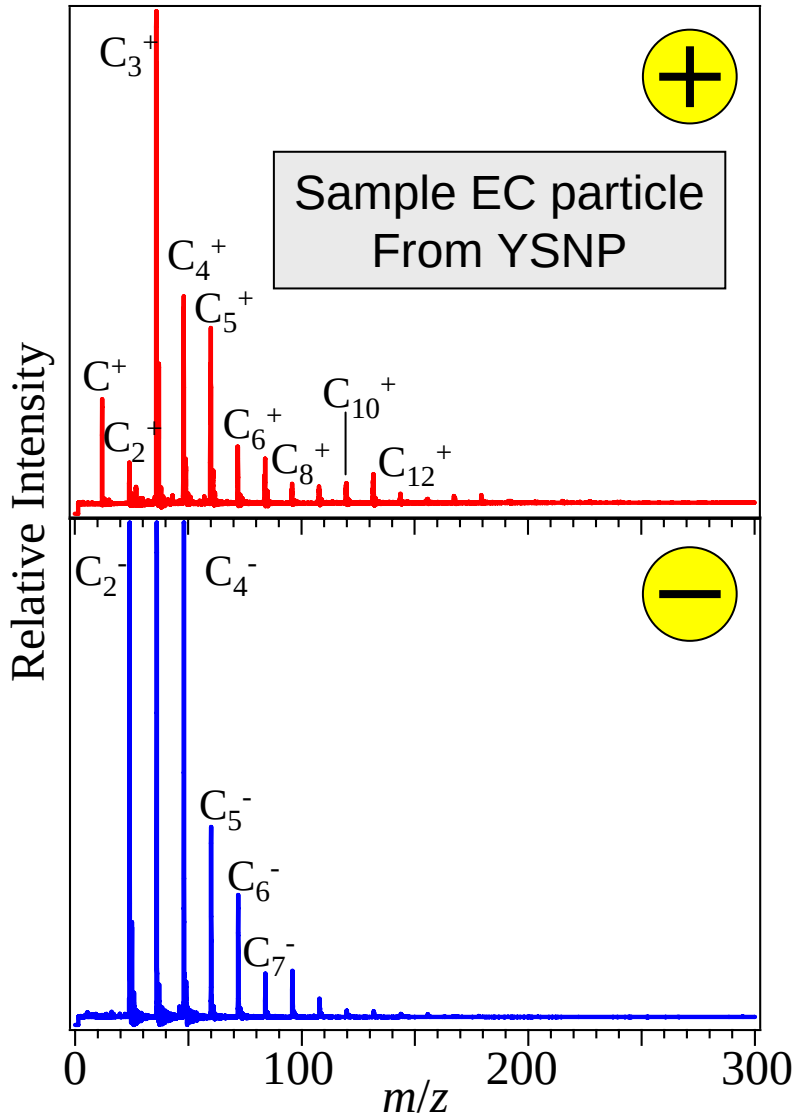
2 of the 9 Cluster Centroids from 2000 East St. Louis ATOFMS Particles



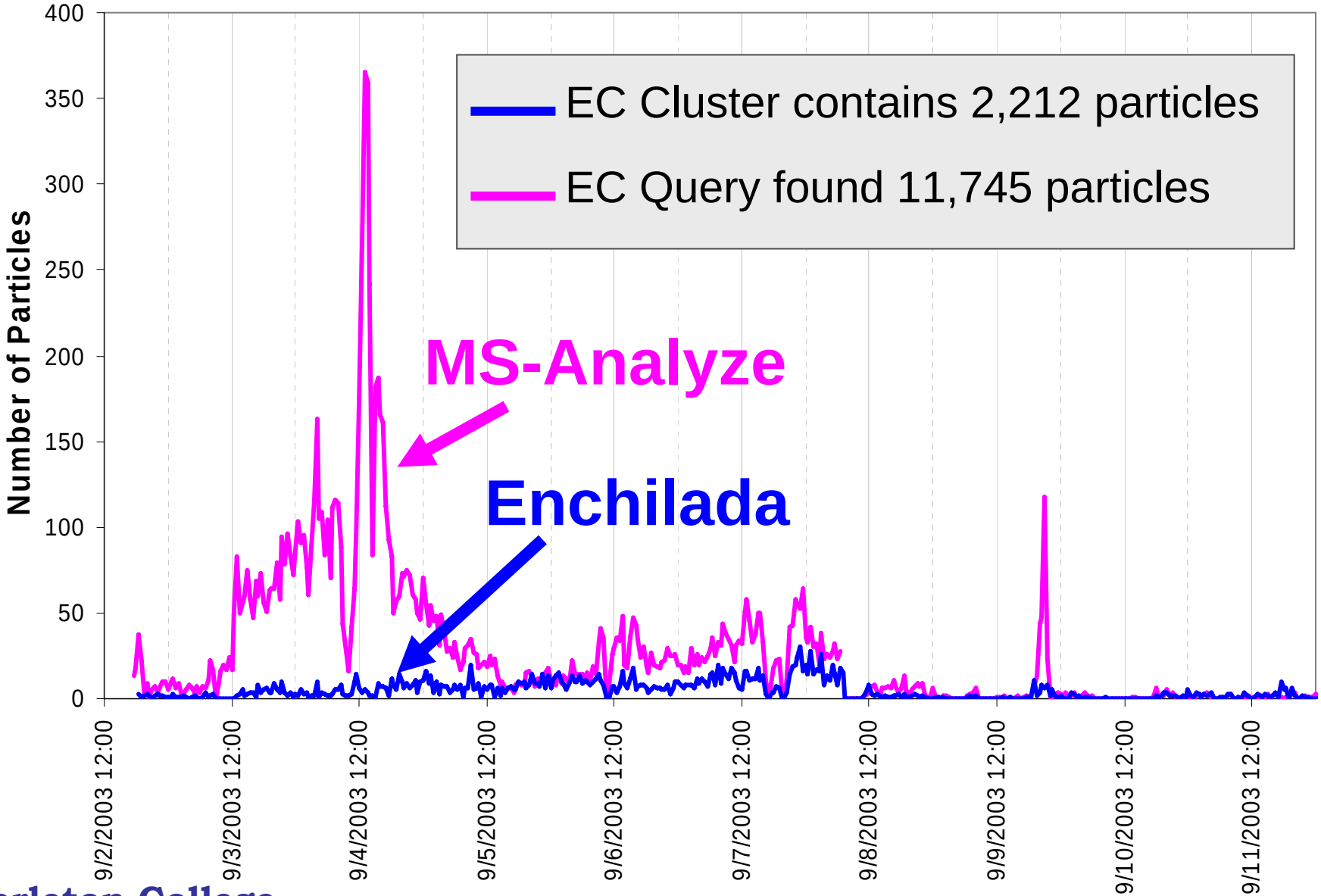
ART-2a centroid
***k*-Means centroid**

Clustering: An EC Case Study

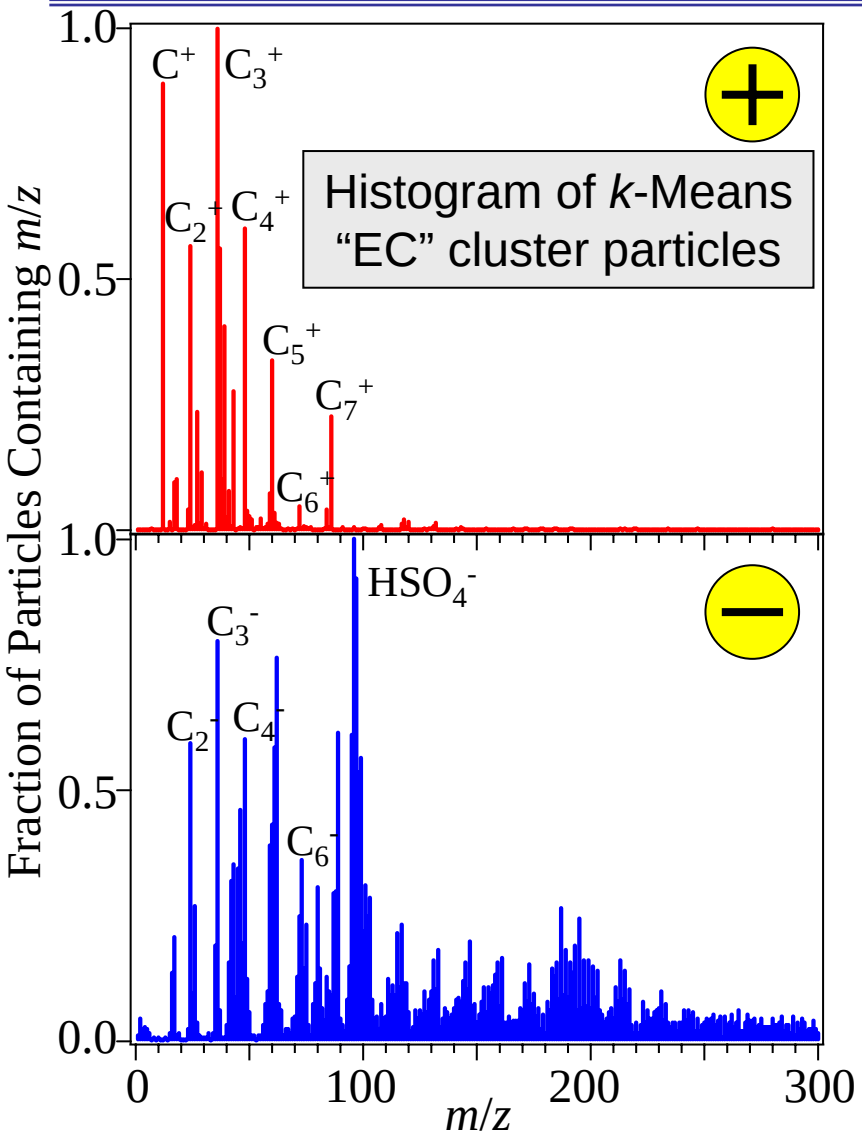


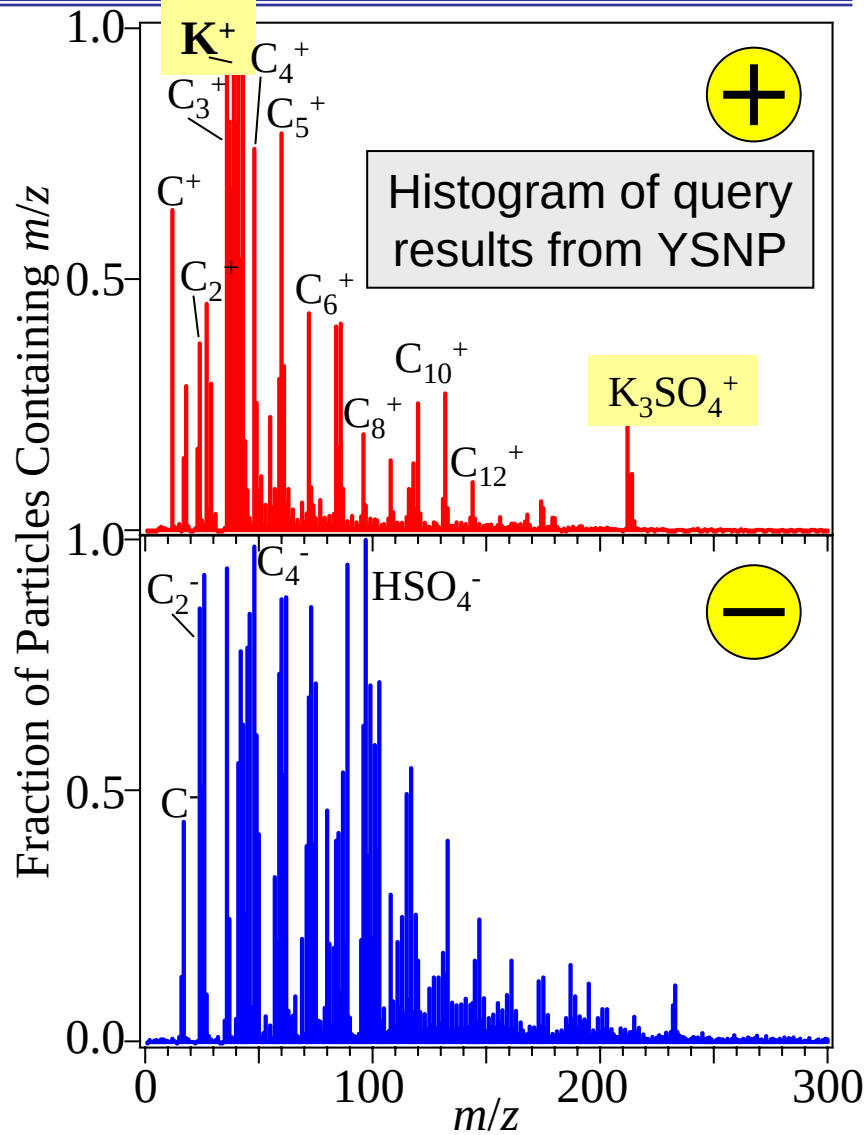


Timeline: Query vs. Cluster



What are the other particles?





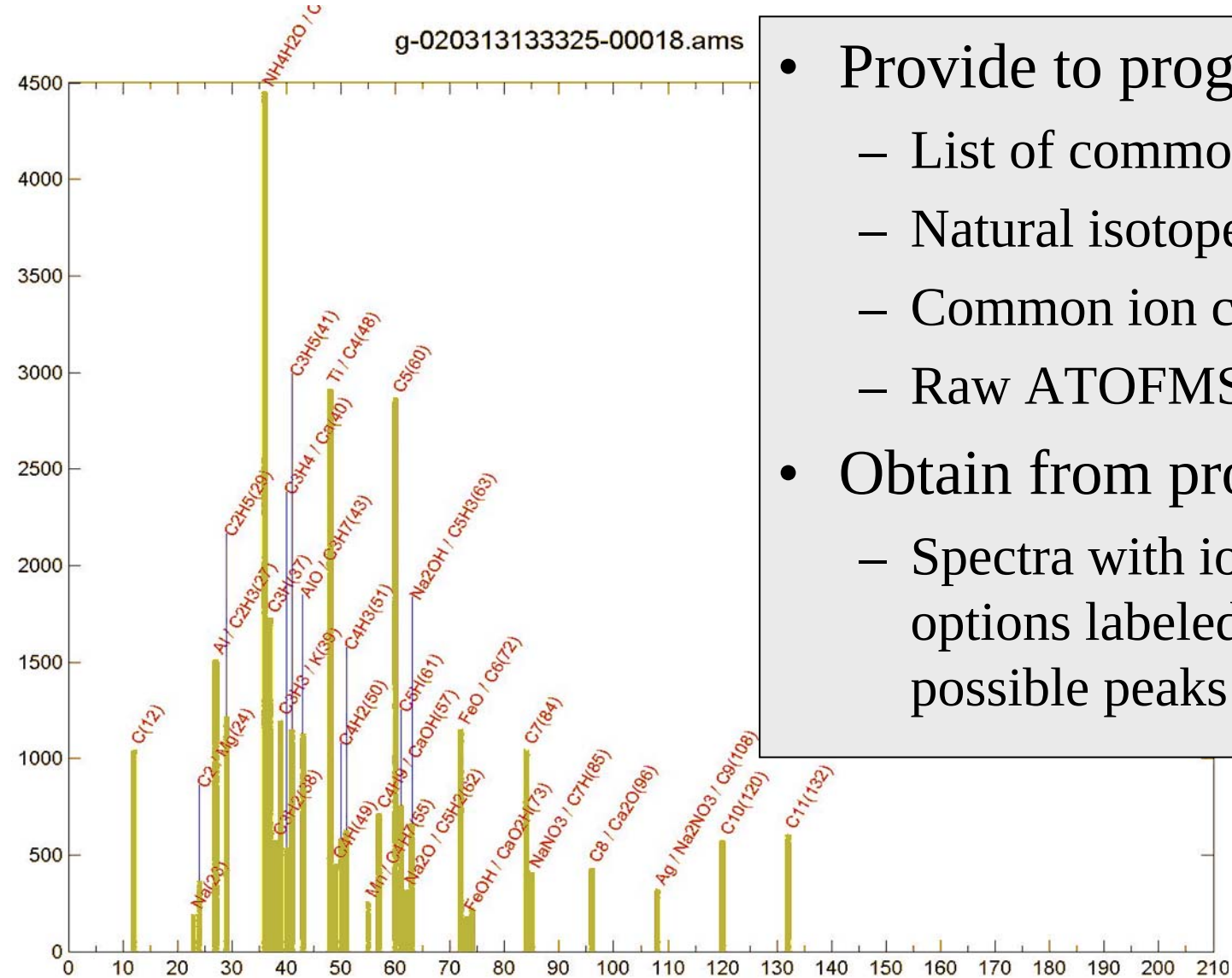
Cluster and Query Results

- What are EC particles?
- Clusters produce a narrow definition of EC and queries use a broad definition of EC
  - Are the queries finding too many particles?
  - Are EC particles distributed among many clusters?
- Is there an EC particle type that the clustering algorithm is not isolating?
  - **YES! Other clusters contain EC and K^+ peaks!**
  - Explore integrating domain knowledge for improved clustering.
- Combine query and cluster perspectives!

Future Work in Clustering

- Finding and implementing algorithms to cluster very large datasets
- Real-time clustering
- Finding clusters that change over time
- Adjustable weighting for data features
- Clustering multiple datatypes together

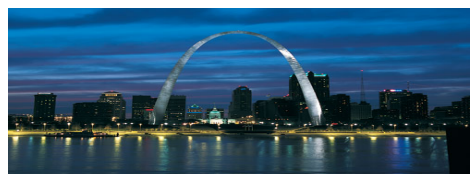
Labeling Mass Spectra



- Provide to program:
  - List of common ions
  - Natural isotope distributions
  - Common ion combinations
  - Raw ATOFMS spectra
- Obtain from program:
  - Spectra with ion composition options labeled above all possible peaks

Conclusions and Future Endeavors

- Enchilada development will continue, and new algorithms will be developed and tested.
- New features will be implemented:
  - Real-time data acquisition and analysis – Enchilada will capture data as it is saved.
  - Support for a variety of data sources, especially other mass spectral data types.
- Many datasets for analysis:
  - Yellowstone National Park (Summer 2003): Natural Geothermal
  - East St. Louis (Winter 2003/04): Industrial Emissions
  - Switzerland (Fall/Winter 2005): Wood Smoke and Vehicles



Acknowledgements

- The National Science Foundation
- Carleton College
- The University of Wisconsin-Madison
- TSI, Inc.
- Others who have contributed:
  - Ben Anderson (Carleton CS)
  - Andy Ault (Carleton Chemistry)
  - Bee-Chung Chen (UW-Madison CS)
  - Zheng (Colin) Huang (UW-Madison CS)
  - Kate Nelson (Carleton CS)
  - Jon Sulman (Carleton CS)