X-Informatics Case Study: e-Commerce and Life Style Informatics: Recommender Systems Technologies V: Clustering

June 19 2013

Geoffrey Fox
gcf@indiana.edu

http://www.infomall.org/X-InformaticsSpring2013/index.html

Associate Dean for Research, School of Informatics and Computing
Indiana University Bloomington

2013

Big Data Ecosystem in One Sentence

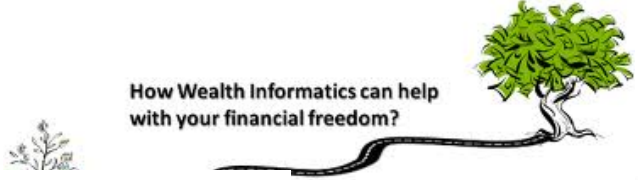
Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics (or e-X)

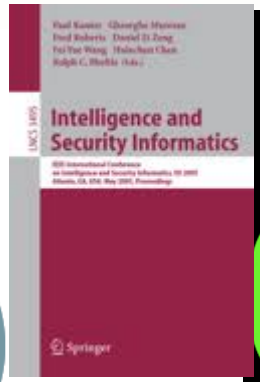
X = Astronomy, Biology, Biomedicine, Business, Chemistry, Climate, Crisis, Earth Science, Energy, Environment, Finance, Health, Intelligence, Lifestyle, Marketing, Medicine, Pathology, Policy, Radar, Security, Sensor, Social, Sustainability, Wealth and Wellness with more fields (physics) defined implicitly

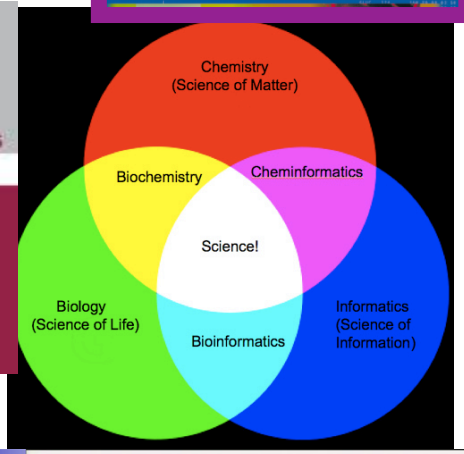
Spans Industry and Science (research)

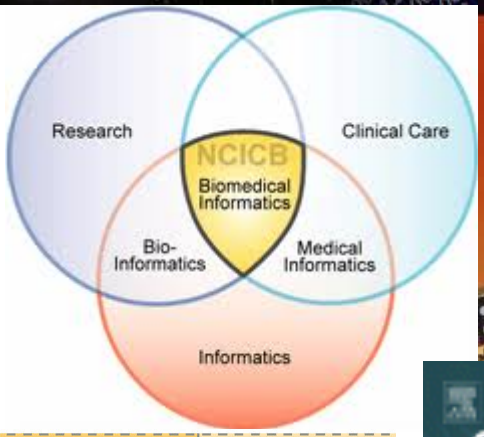
Education: Data Science see recent New York Times articles

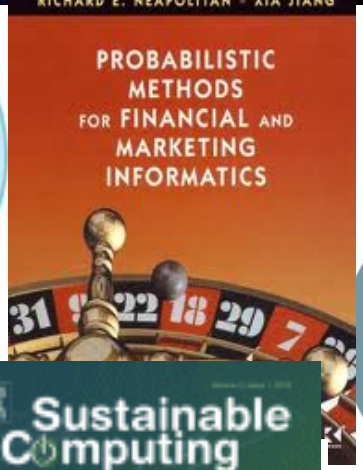
http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/

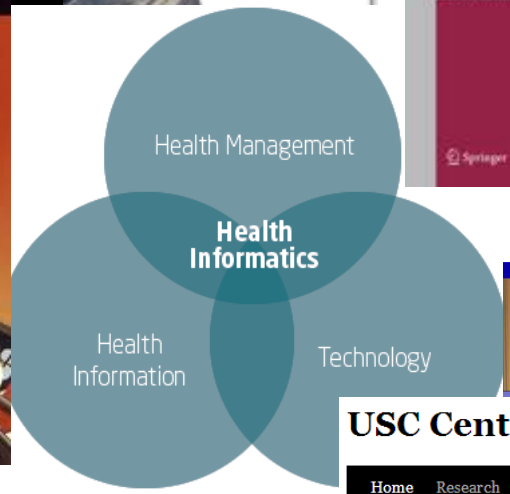


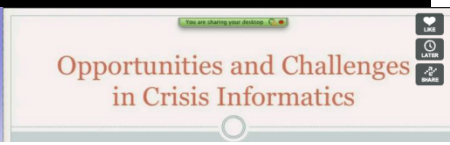


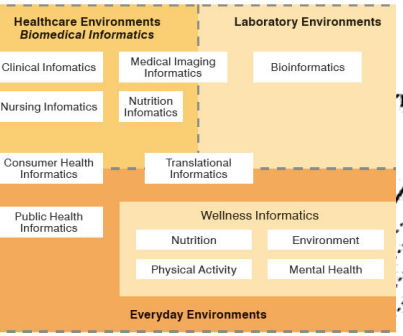


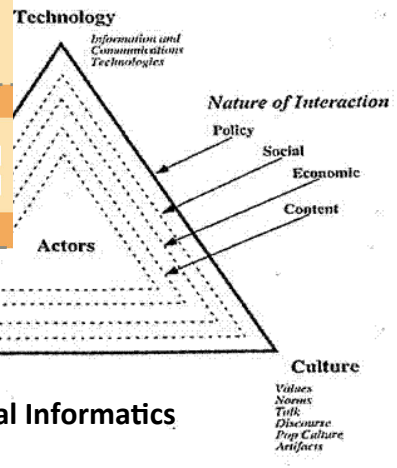








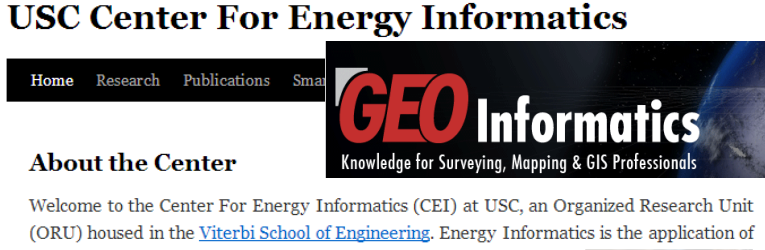


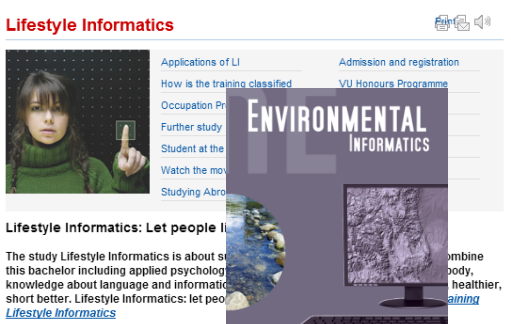


Noelia Penelope Greer (Ed.)

Business Informatics

Information technology, Management,





Kmeans Clustering

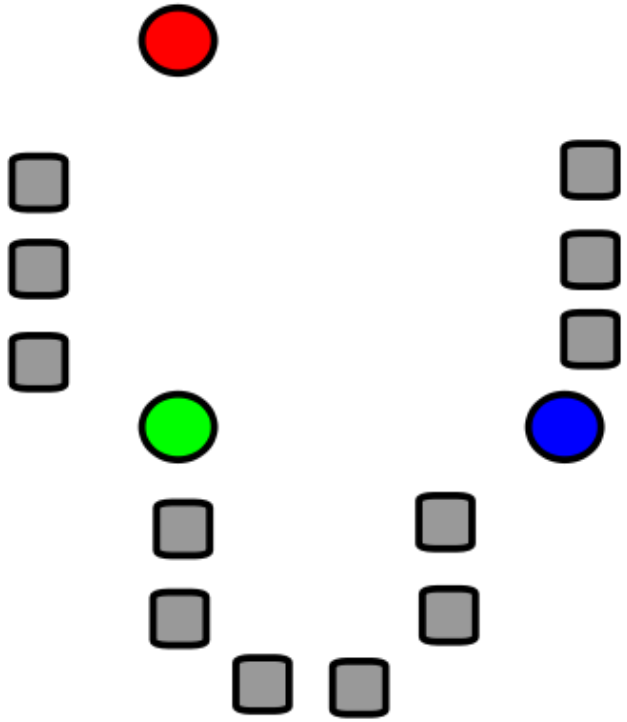
http://en.wikipedia.org/wiki/Kmeans

Resources

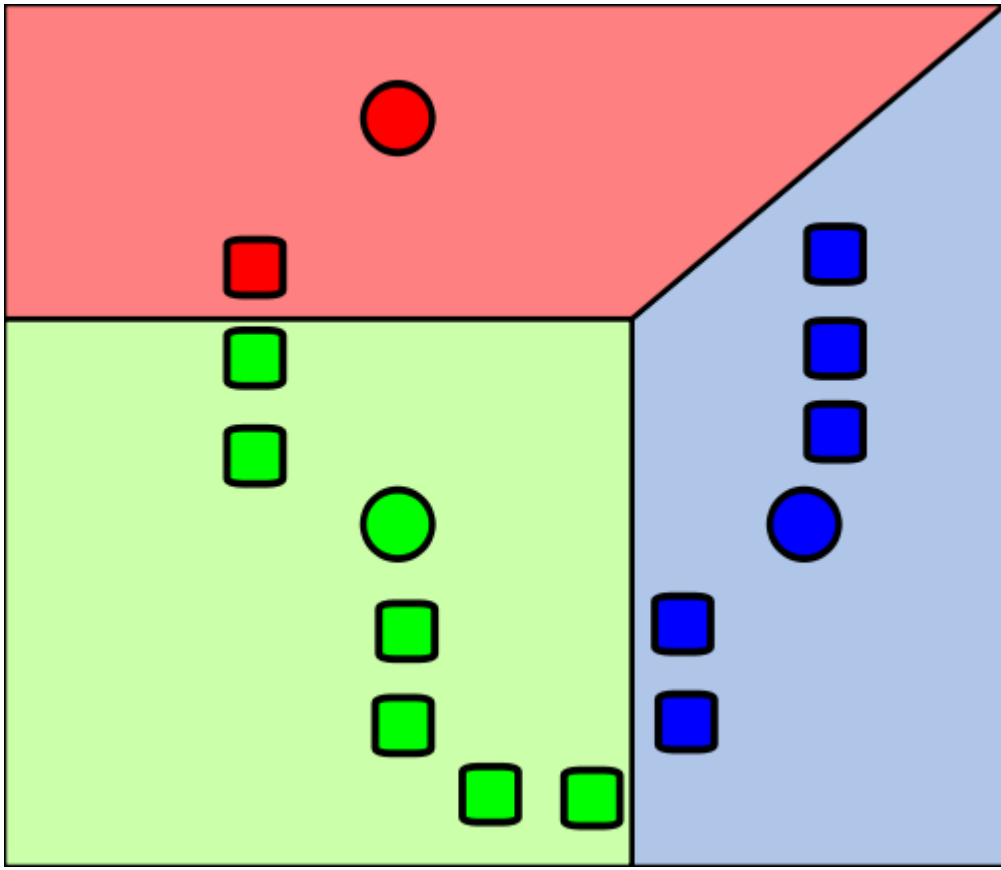
- See http://grids.ucs.indiana.edu/ptliupages/publications/Clusteringv1.pdf for advanced clustering method improving k means
- This lesson has no code but we will use PlotViz program introduced earlier
- And Wikipedia on k means
- Files used are
- clusterFinal-M3-C3Dating-ReClustered.pviz
- clusterFinal-M30-C28.pviz
- These were found by advanced clustering based on vector file DatingRatingforPviz.txt of 1000 points discussed earlier with PlotViz version DatingRating-OriginalLabels.pviz
- fungi_LSU_3(15)_to_3(26)_zeroidx.pviz is used to demonstrate genomic clusters in PlotViz

Kmeans Clustering

- Choose k=2 to 10 million (or more) clusters according to some intuition
- Initialize k cluster centers (this only works in real spaces)
- Then Iterate
  - Assign each point to cluster whose center it is nearest to
  - Calculate new center position as centroid (average in each component) of points assigned to it
- Converge when center positions change < cut
- Lots of variants that are better but k means used as simple and fast
  - Also methods for when only some distances available and no vector space
- Sometimes gets wrong answer – trapped in a local minima

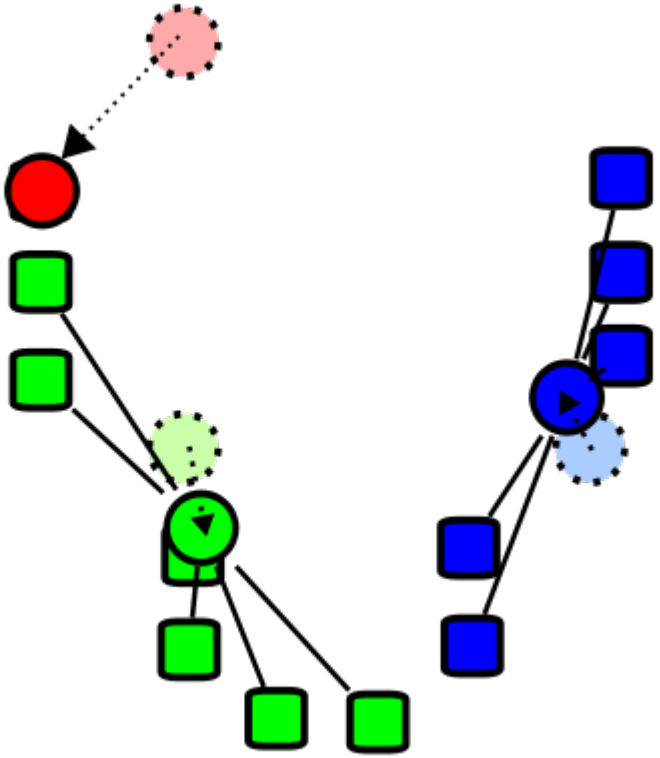


1) *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

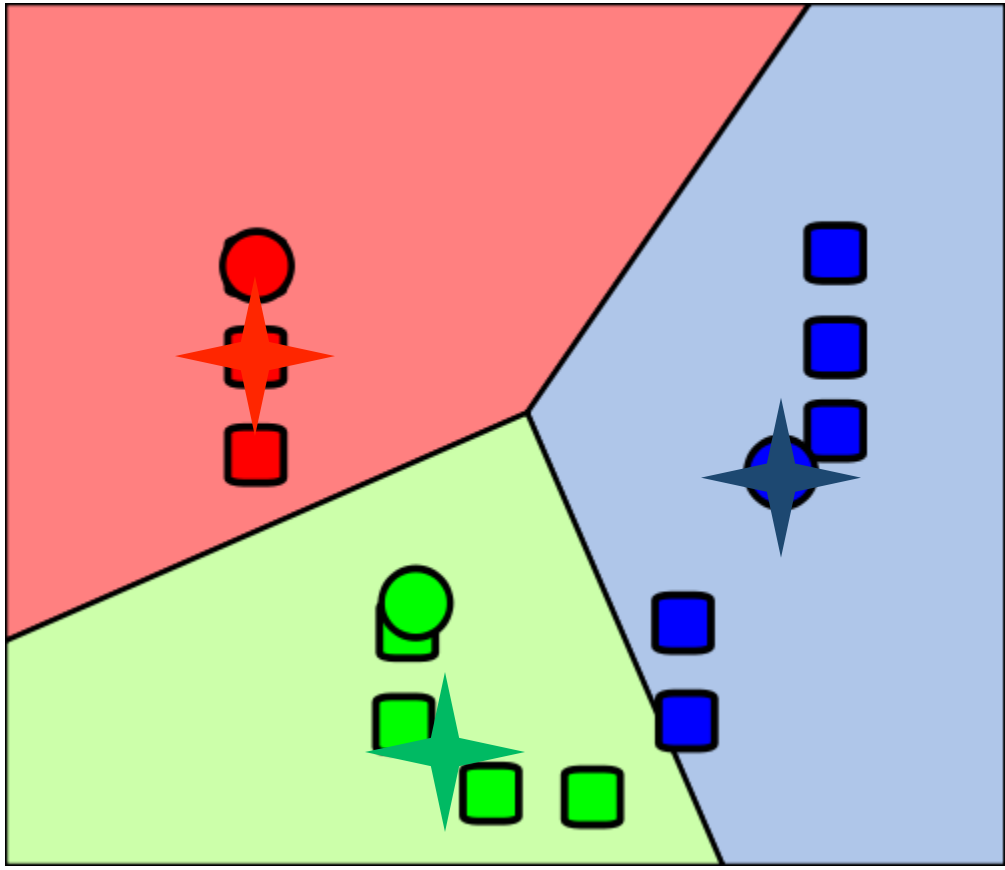


2) k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

Wikipedia



3) The centroid of each of the k clusters becomes the new mean.



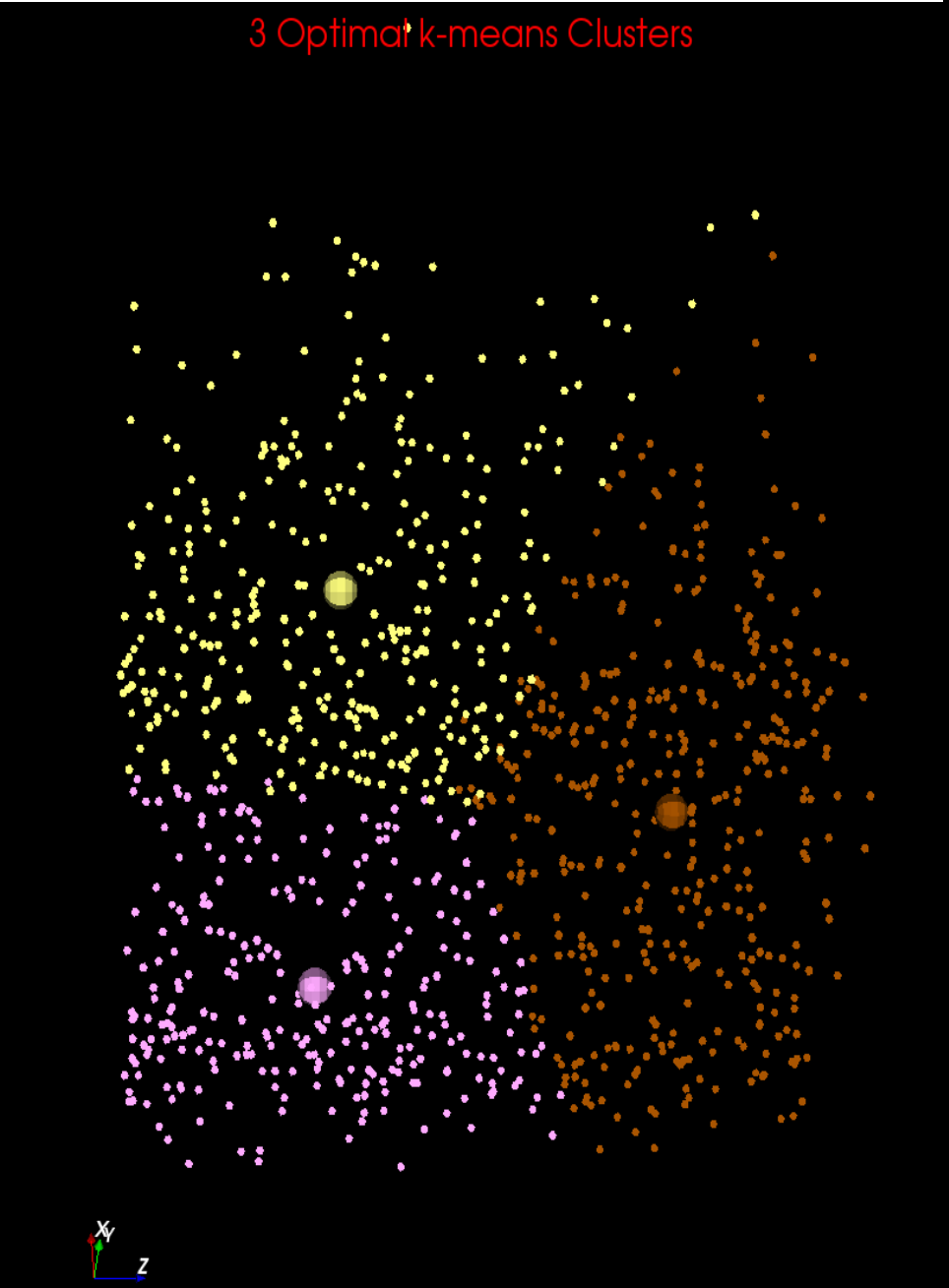
4) Steps 2 and 3 are repeated until convergence has been reached. But red and green centers (circles) are clearly in wrong place!!! (stars added by me about right for centers)

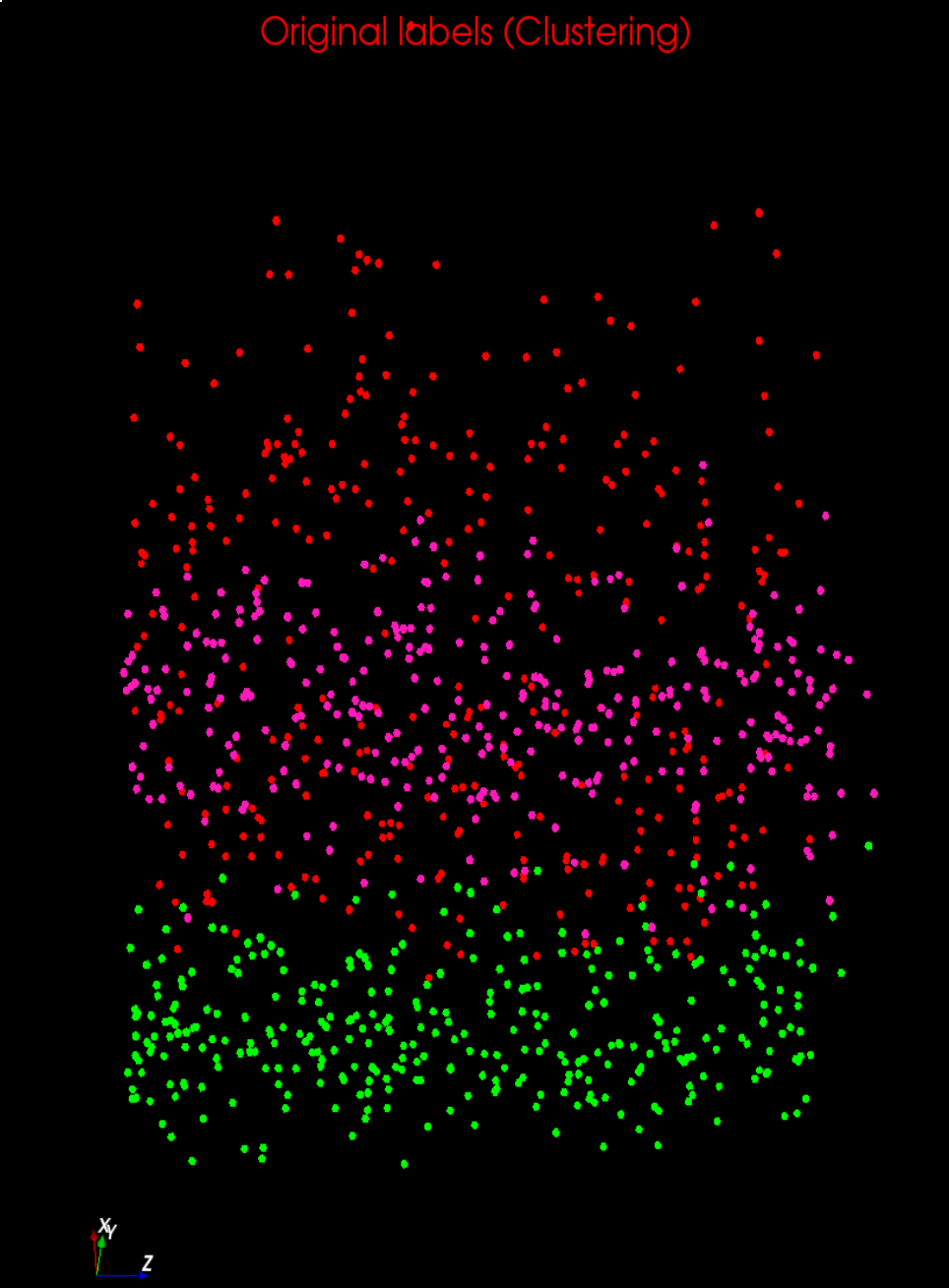
Wikipedia

Clustering of Recommender System Example

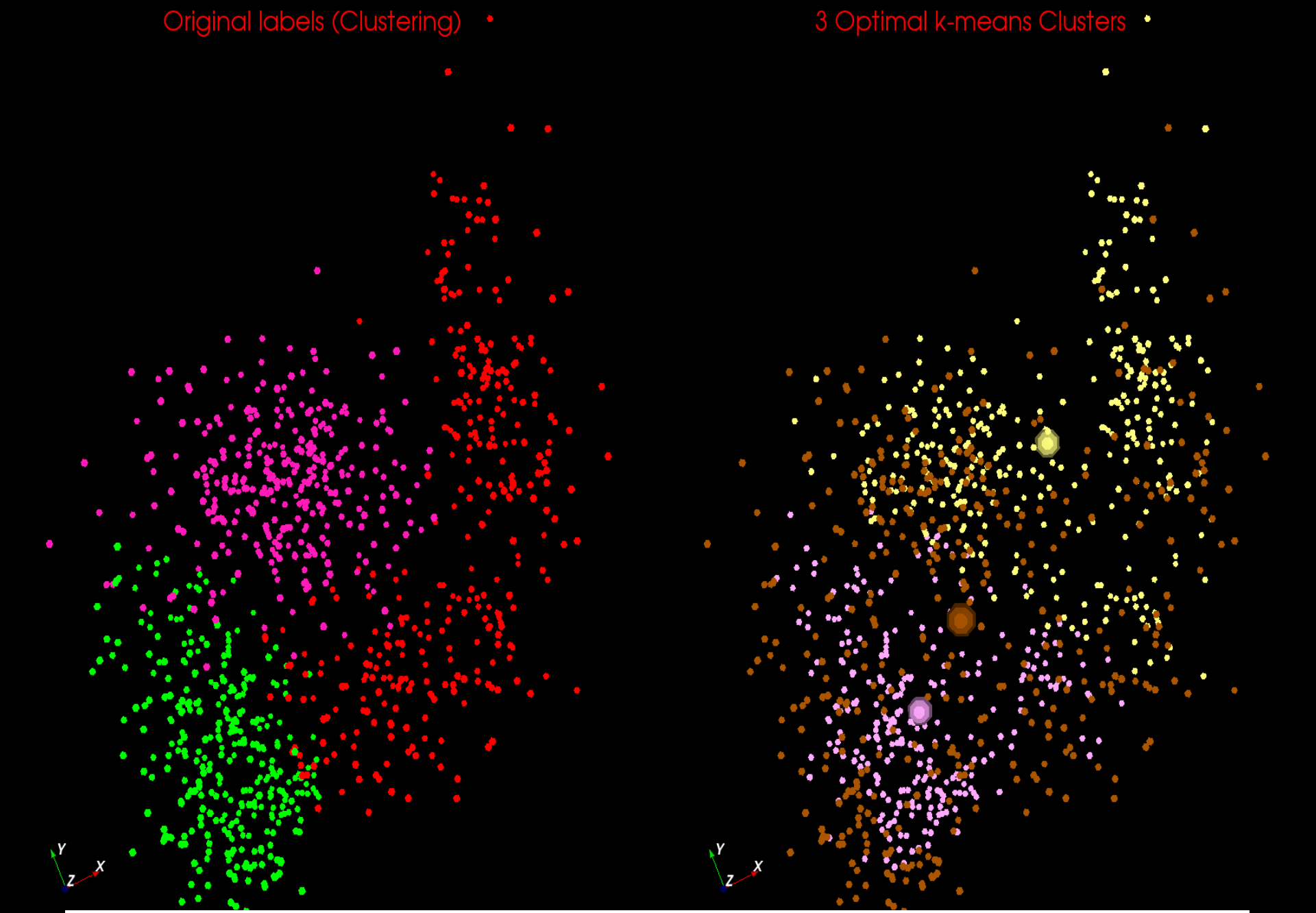
Next Set of Slides

- Compares original labelling of points with a fresh clustering into 3 clusters (using deterministic annealing advanced method in paper but k means similar)
- Fresh clustering has centers marked
- The 1000 points are identically placed in side by side pictures

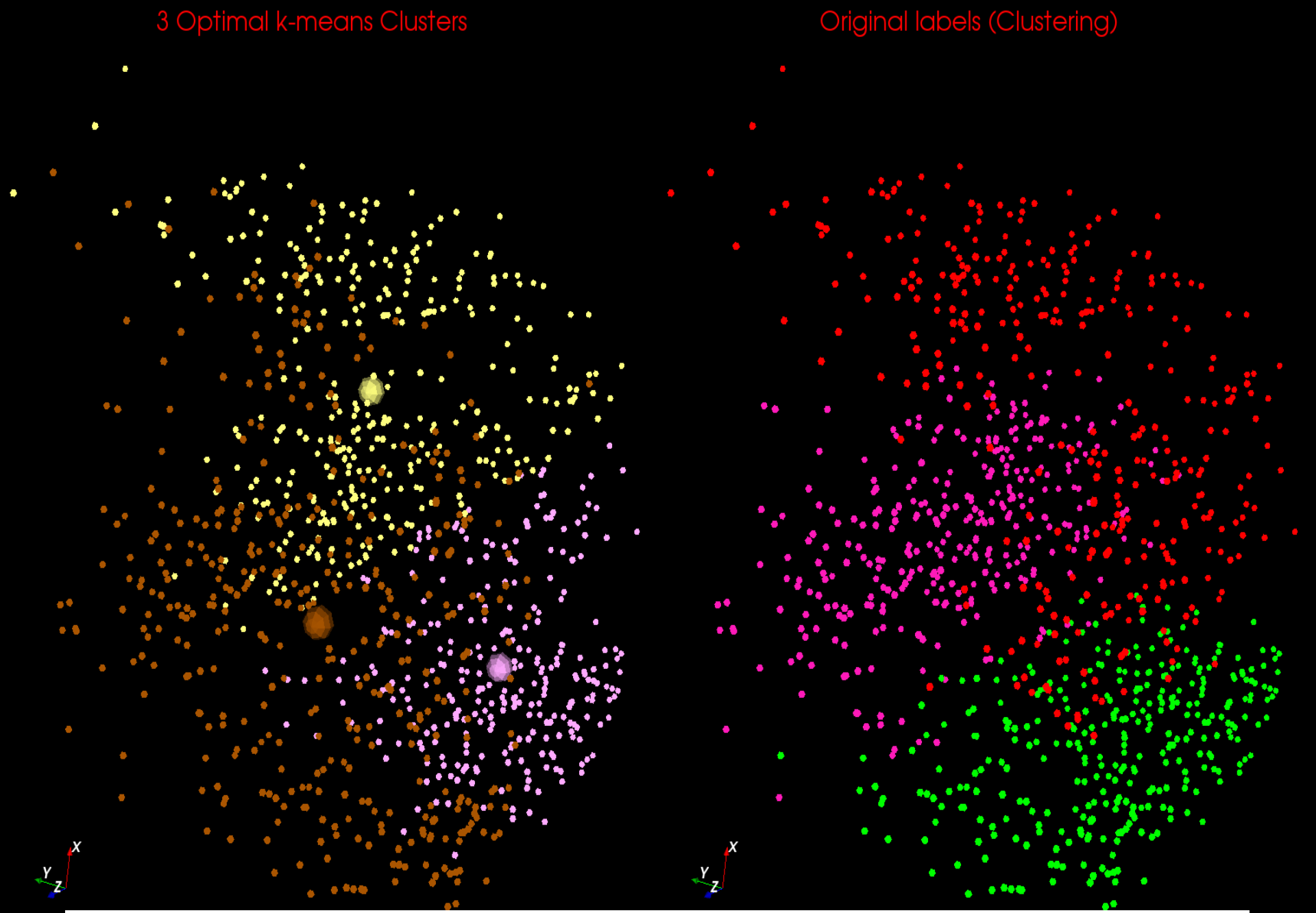




Clustering v User Rating I



Clustering v User Rating II



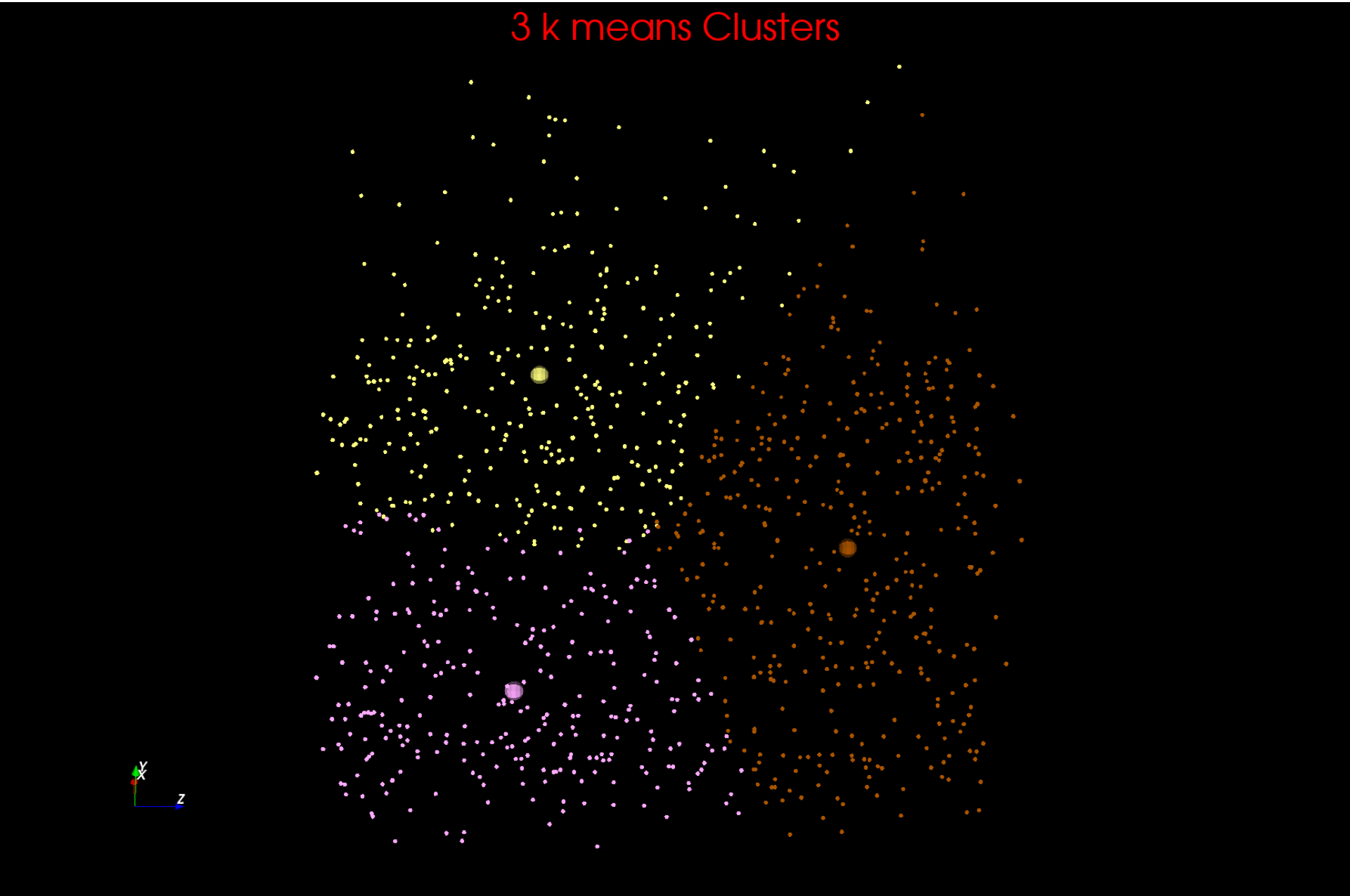
Clustering v User Rating III

Clustering into more than 3 Clusters

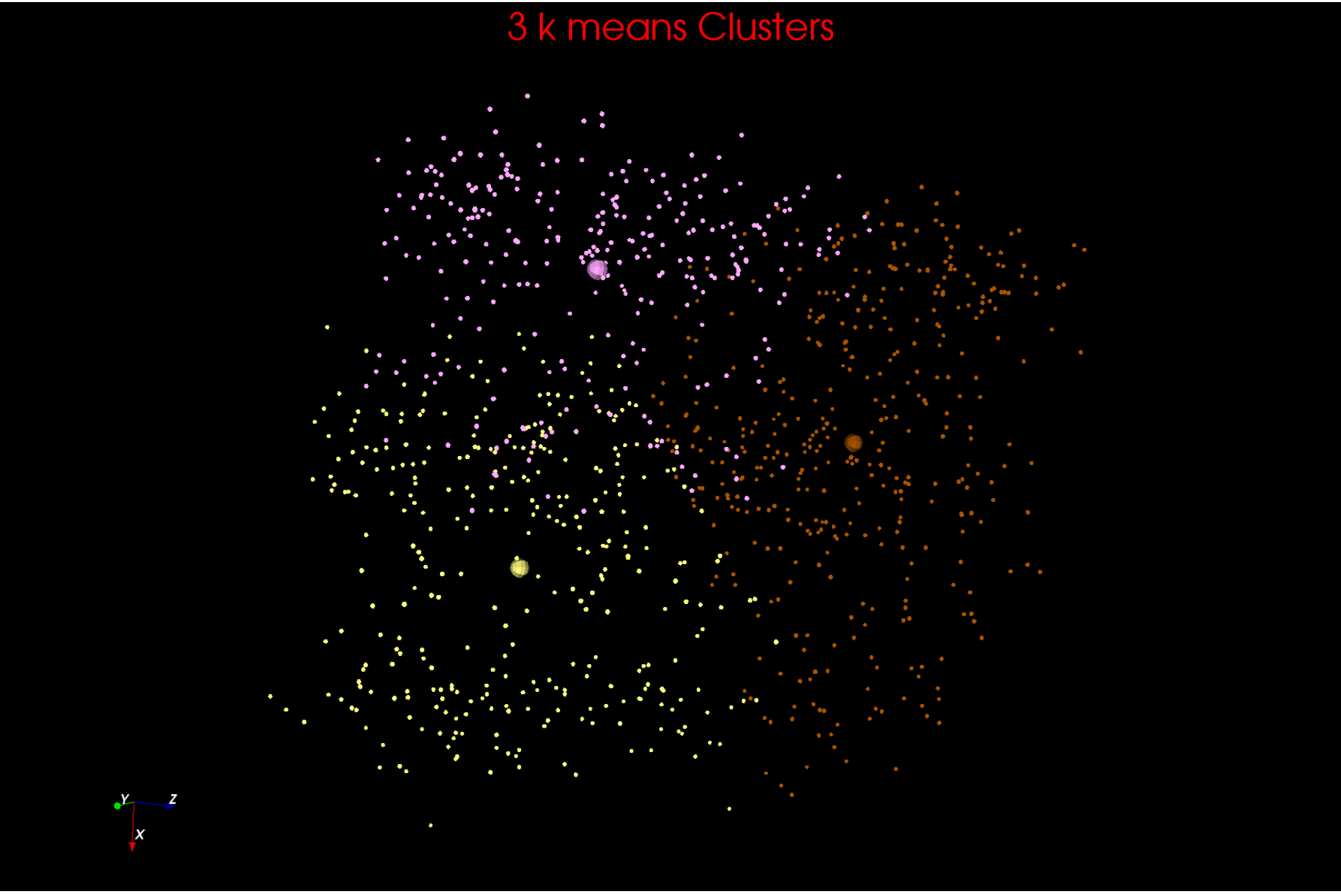
Next Set of Slides

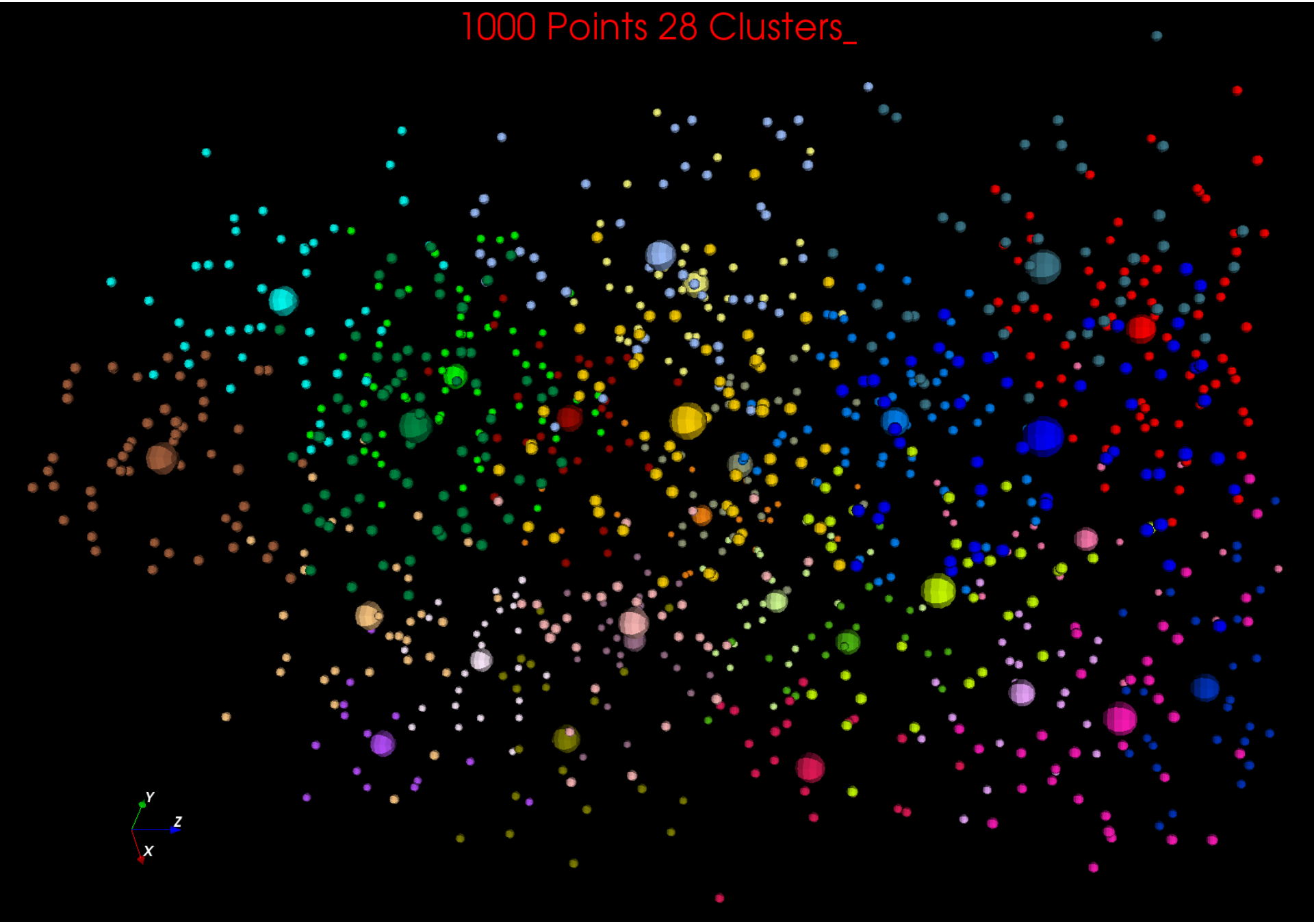
- Looks at fresh clusters into 3 or 28 clusters using deterministic annealing
- We tried for 30 clusters in second case but program removes small clusters so it went down to 28
- Note clustering divides full region into geometrically compact subregions

Two views of a 3D Clustering I

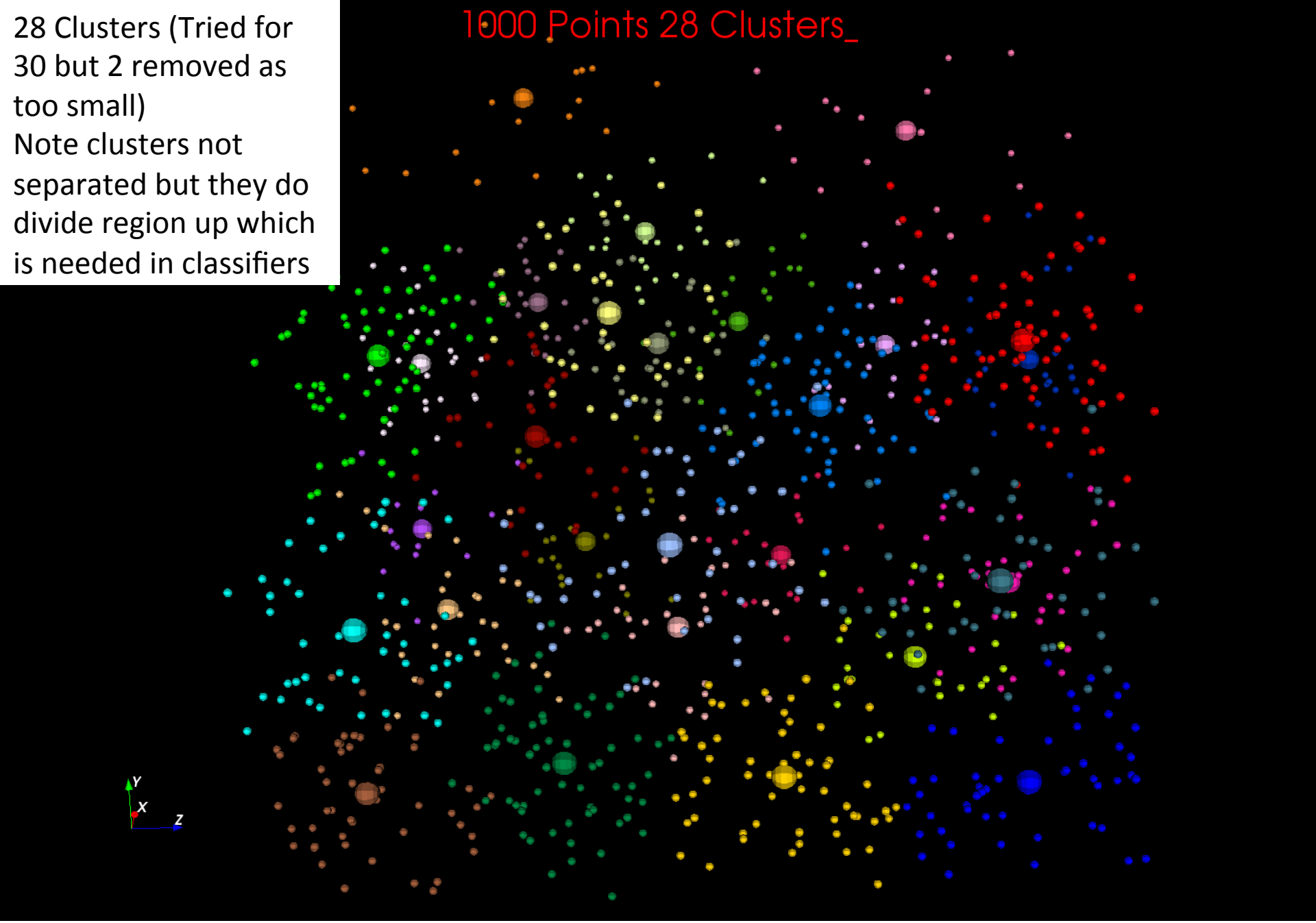


Two views of a 3D Clustering II




1000 Points 28 Clusters_
Y
Z
X

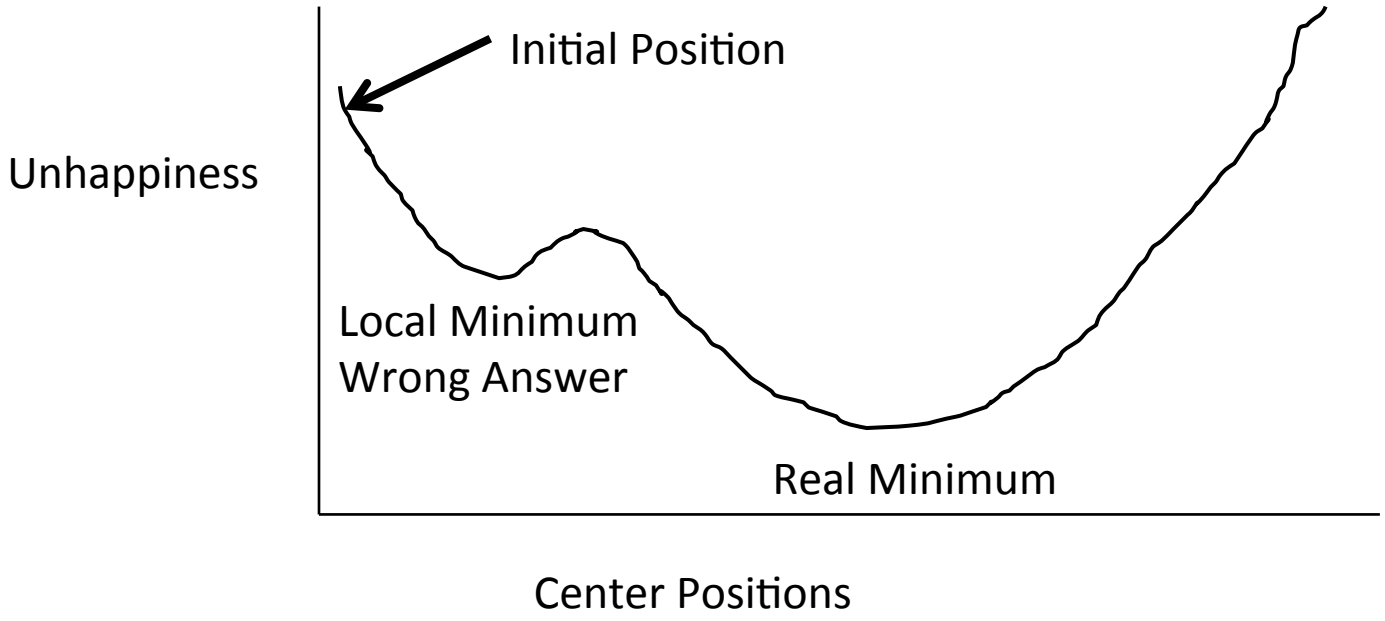
28 Clusters (Tried for 30 but 2 removed as too small)
Note clusters not separated but they do divide region up which is needed in classifiers



Local Optima in Clustering

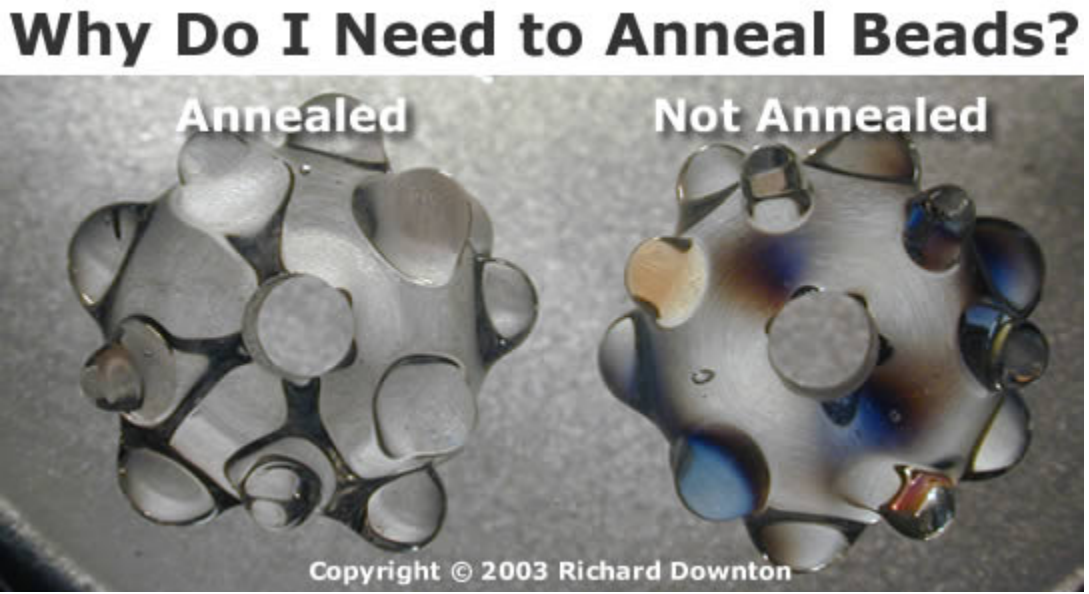
Getting the wrong answer

- Remember all things in life – including clustering – are optimization problems
- Lets say it's a minimization (change sign if maximizing)
- k means always ends at a minimum but maybe a local minimum out of which small changes are stable
- We will see examples of false minima later in course in Kmeans with Python lesson

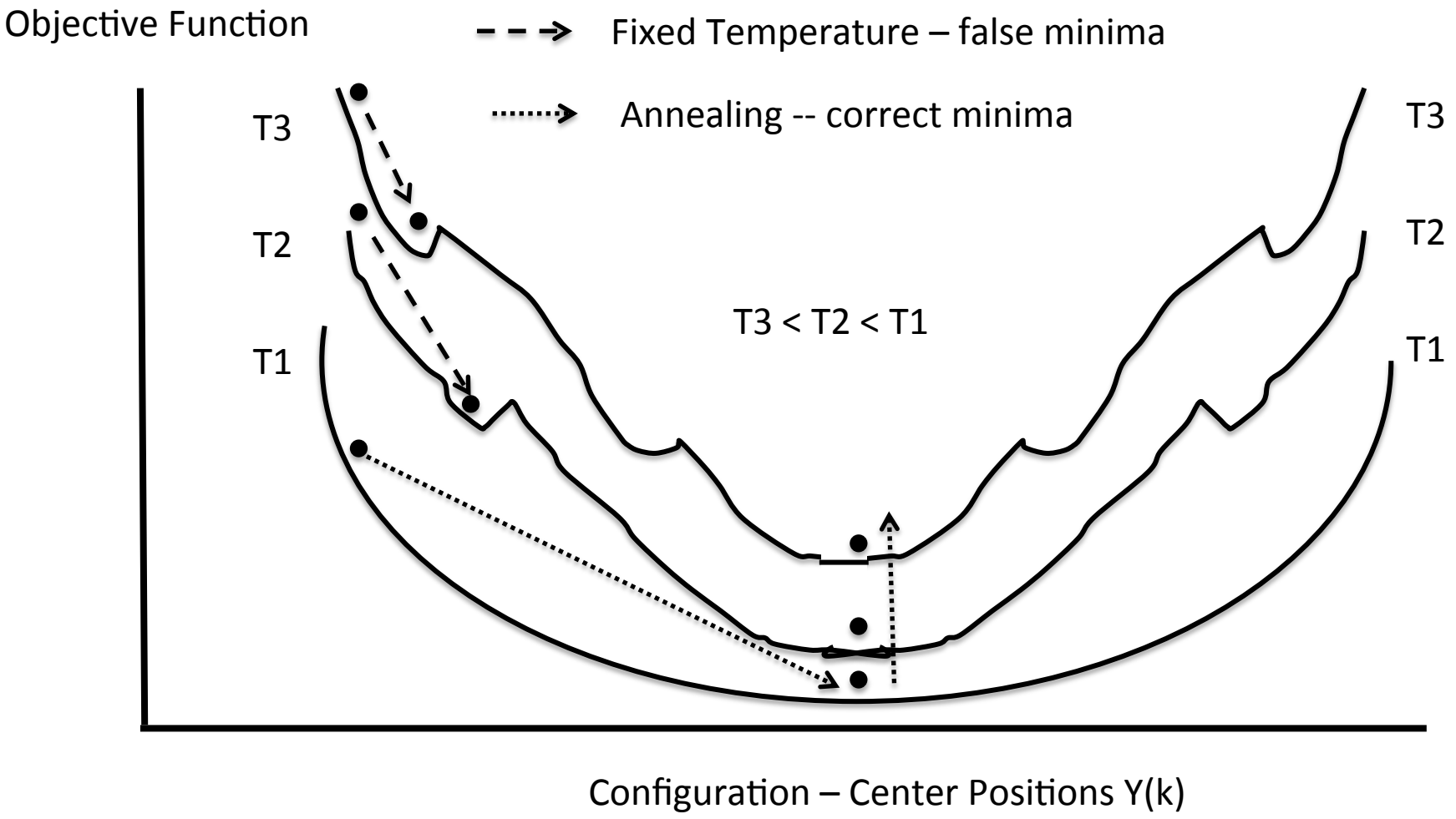


Annealing

- Physics systems find true lowest energy state if you anneal i.e. you equilibrate at each temperature as you cool
  - Allow atoms to move around
- If system cools too fast it finds a false minima
- Corresponds to fuzzy algorithms



Annealing


Objective Function
Fixed Temperature – false minima
Annealing -- correct minima
T3
T2
T1
T3 < T2 < T1
T3
T2
T1
Configuration – Center Positions Y(k)

Greedy Algorithms

- Many algorithms like k Means are greedy
- They consist of iterations
- Each iterations makes the step that most obviously minimizes function
- Set of short term optima; not always global optima
- Same is true of life (politics, wall street)

Clustering in General

Purpose of Clustering

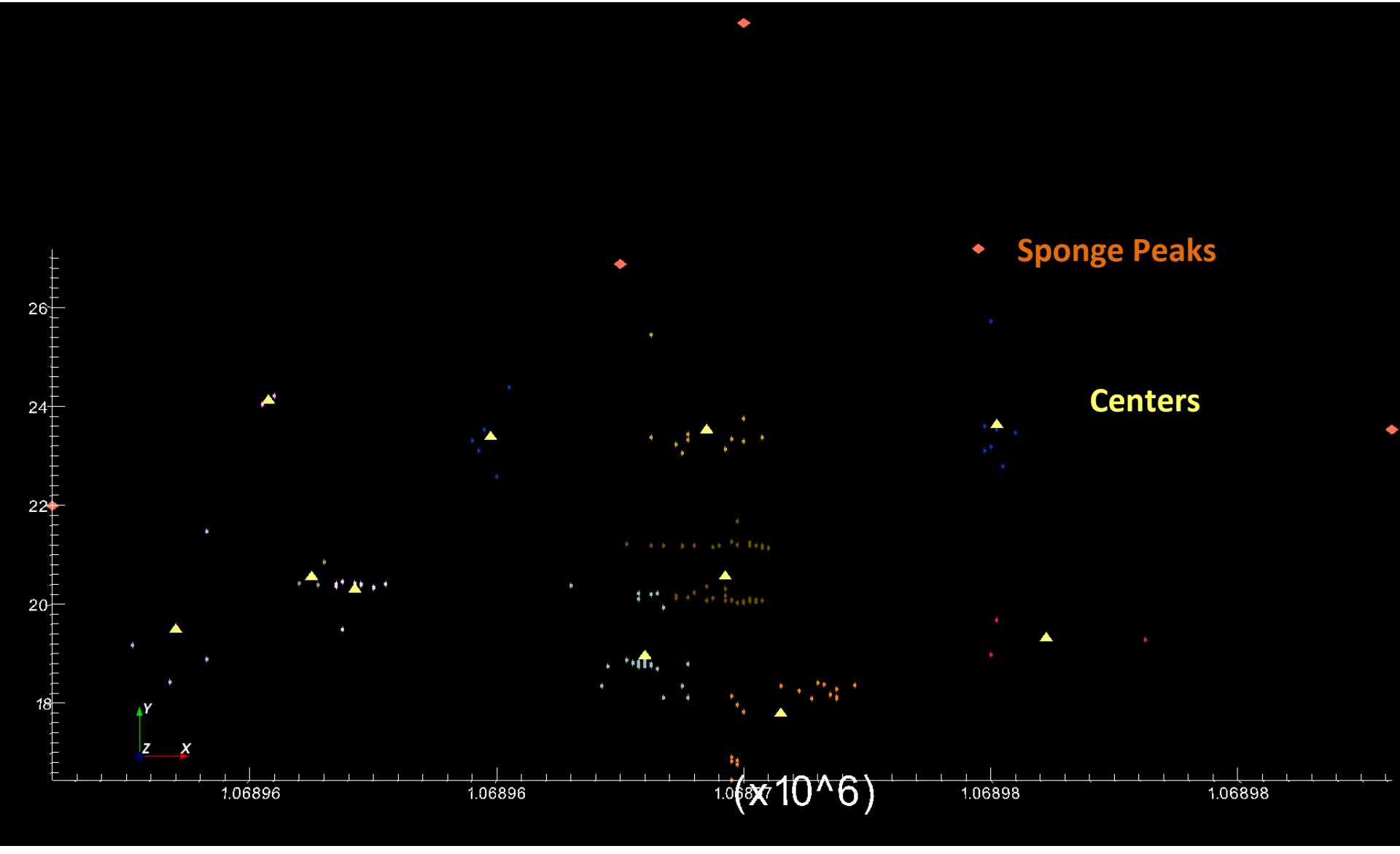
- This is used in several somewhat different ways
- First is taking a set of points in a space and divide into distinct groups with (clear) separation
- Second is dividing points into neighboring points but not groups that are separate
- There are also latent factor and mixture models which use a different definition of cluster that adds probabilities
  - These are used to group “items” together in content based approaches
- Further one has methods/applications that work in
  - Real vector spaces
  - Spaces in which there are no vectors but some or all of distances defined

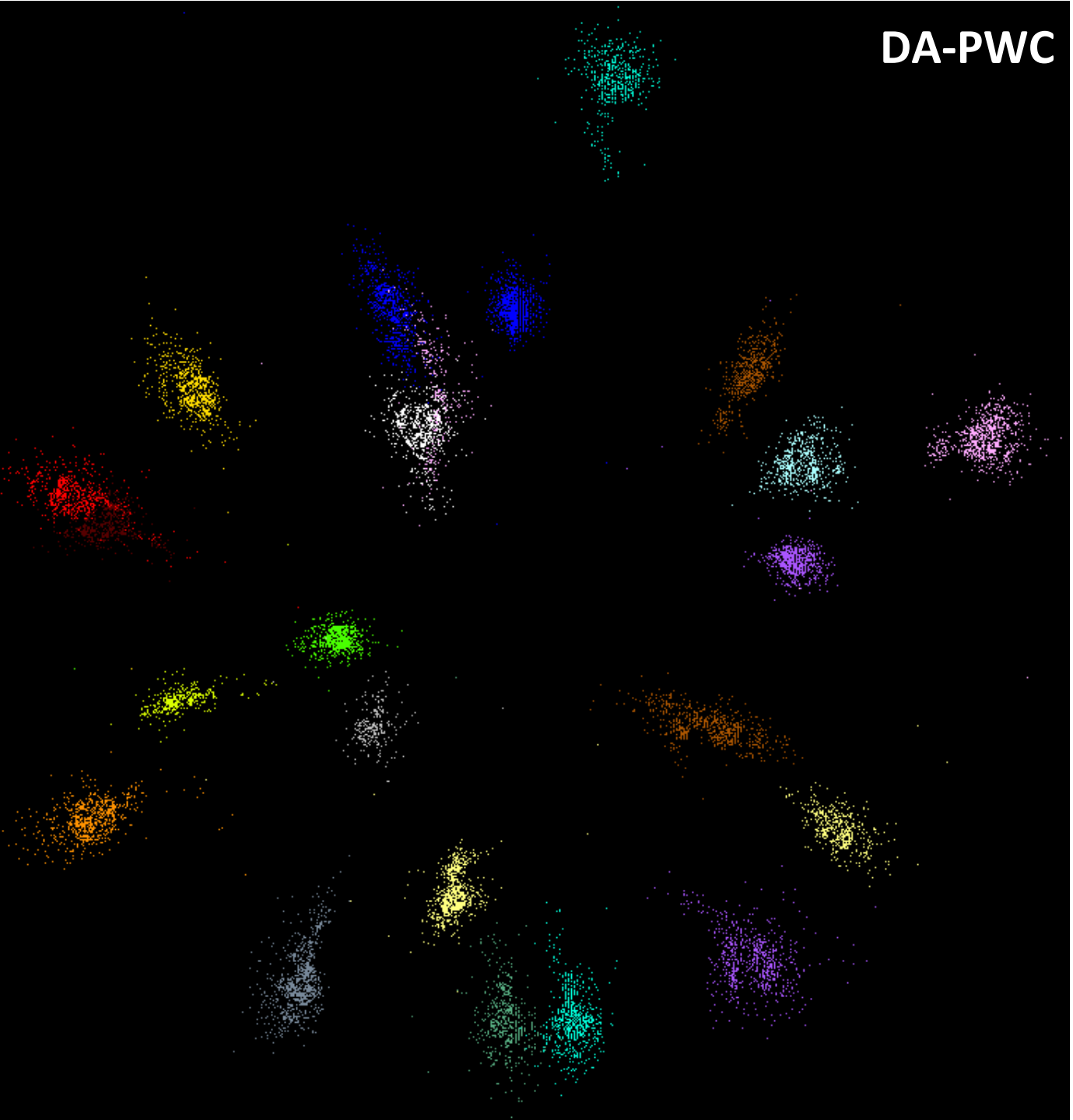
Clustering Examples

- We follow with examples in mass spectrometry peaks and genomics where there are real separated clusters
- The mass spectrometry is in a 2D space
- The genomic case is not a vector space and one clustered without using vectors
  - The pictures show results after projecting to 3 dimensions
- We also have a PlotViz demo

Proteomics 2D DA Clustering T=0.1

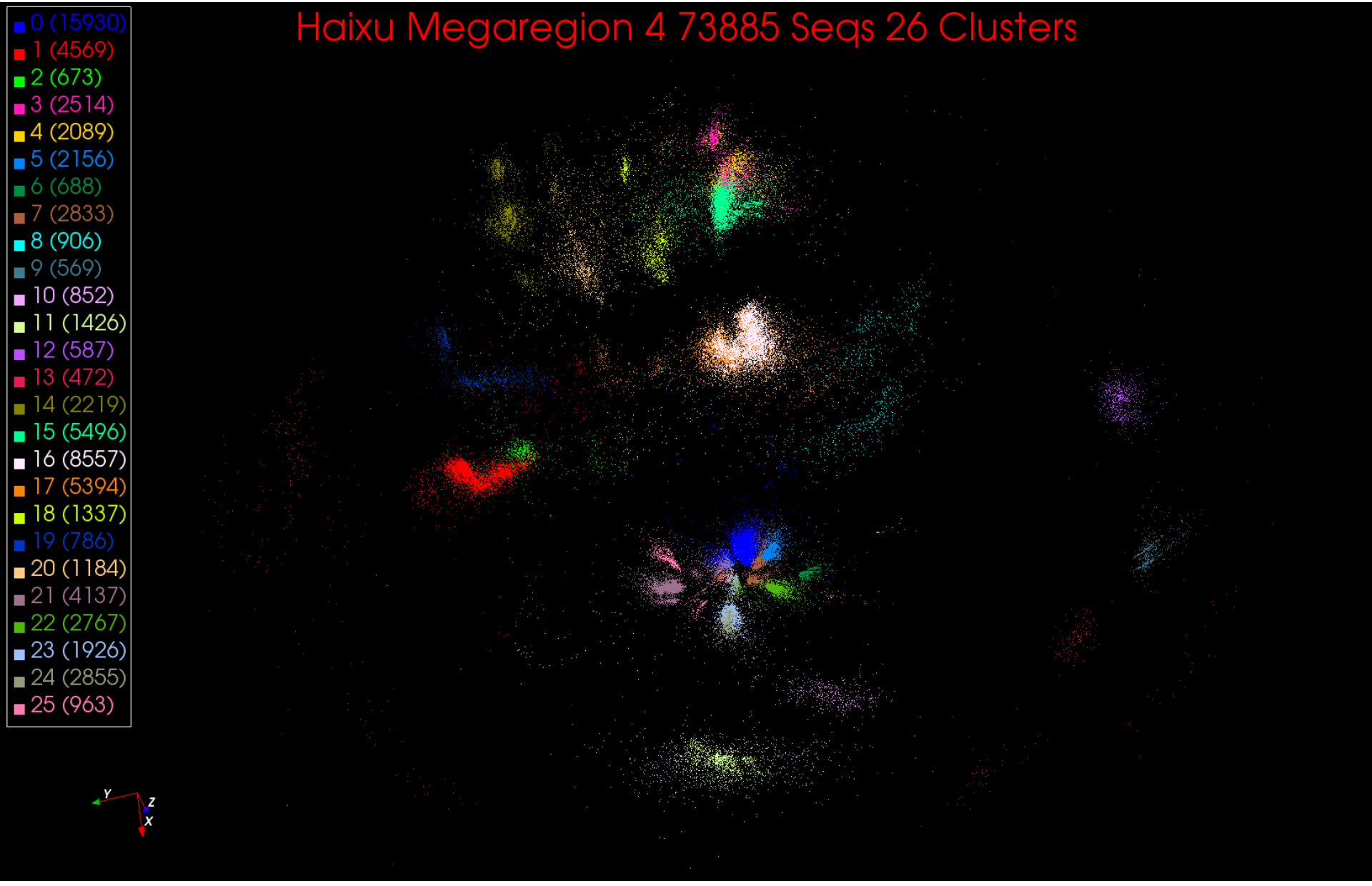
small sample of ~30,000 Clusters Count >=2



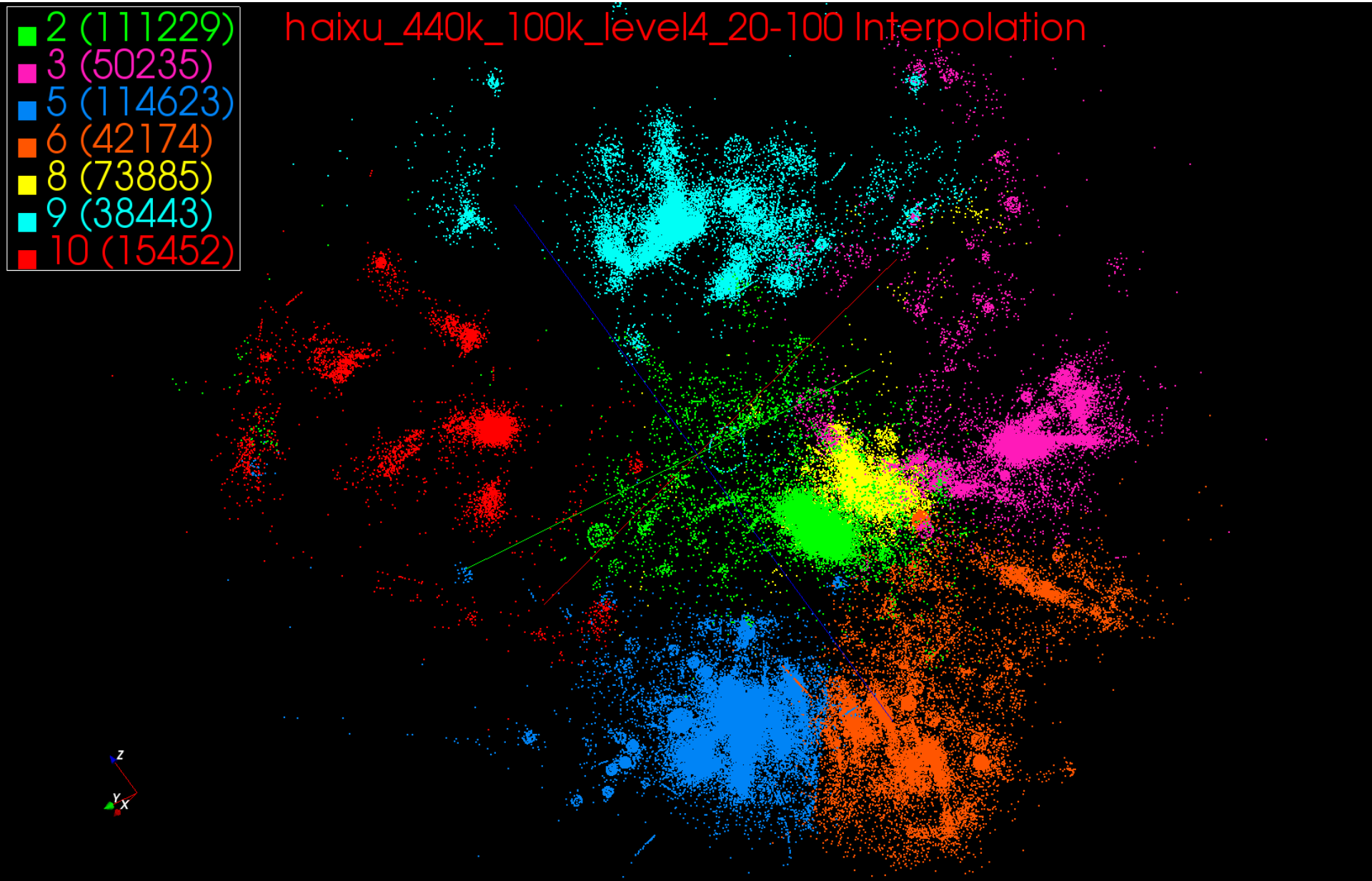


Metagenomics

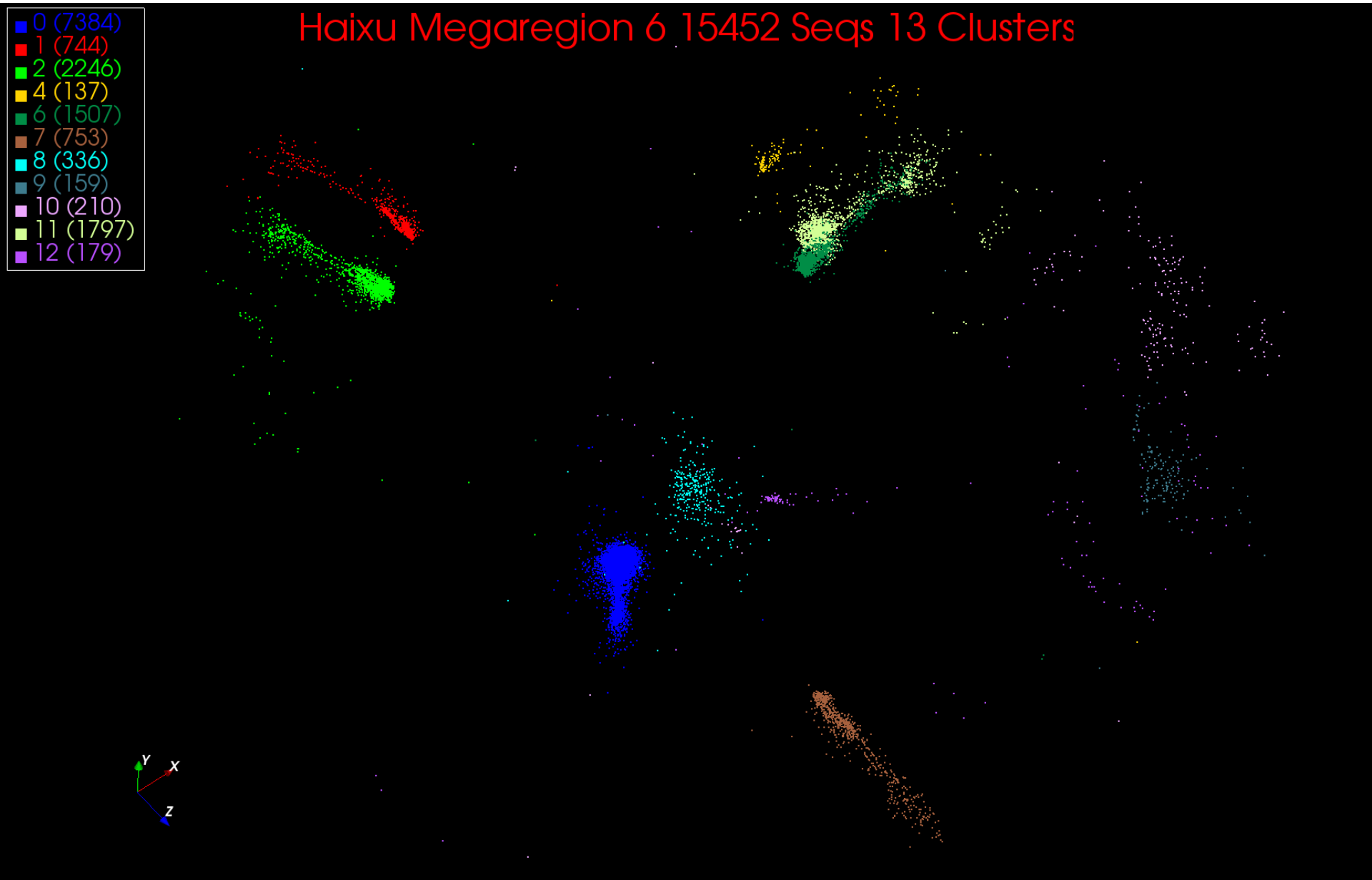
26 Clusters in Region 4



Genomic Data Not fully Clustered



13 Clusters in Region 6



Heuristics

http://en.wikipedia.org/wiki/Heuristics

Heuristics (Wikipedia)

- In computer science, artificial intelligence, and mathematical optimization, a **heuristic** is a technique designed for solving a problem more quickly when classic methods are too slow, or for finding an approximate solution when classic methods fail to find any exact solution. This is achieved by trading optimality, completeness, accuracy, and/or precision for speed.
- Results about **NP-hardness** in theoretical computer science make heuristics the only viable option for a variety of complex optimization problems that need to be routinely solved in real-world applications.
- Clustering is NP hard although collaborative filtering is just “hard” and computationally intractable for real time use
- Roughly NP hard means only (known) exact algorithms are exponential and not polynomial in time to solve

Heuristics II

- Collaborative Filtering (user based) is O(MN) for M customers and N items which is polynomial
- Clustering is harder to estimate as we don't have an exact algorithm
- There is combinatorial estimate K^M where we loop over K assignments for each of M items (K number of clusters)
  - This is exponential
- Note many real world problems are intrinsically "inexact" as
  - Don't have exact data
  - Don't have exact problem
- So heuristics are often/usually just fine and IMHO exponential behavior (NP hard from previous slide) is not so difficult in practice