

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/>



Climate Informatics
network

How Wealth Informatics can help
with your financial freedom?



Xinformatics

xinfor
XIU TOU

Biomedical Informatics
Computer Applications in Health Care
and Biomedicine

AstroInformatics2012

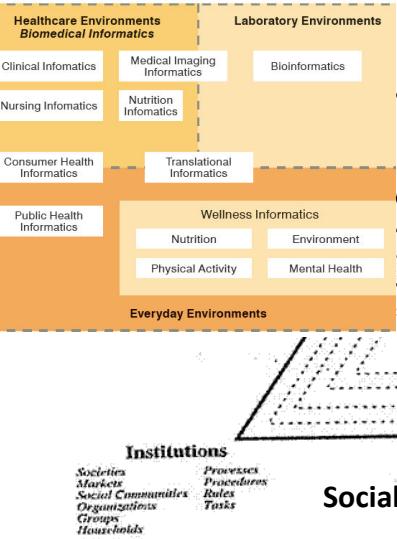
Redmond, WA, September 10 - 14, 2012

RICHARD E. NEAPOLITAN • XIA JIANG

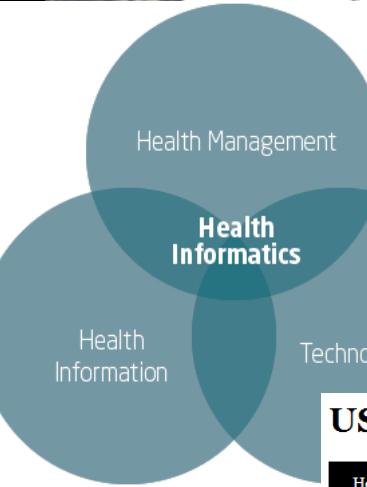
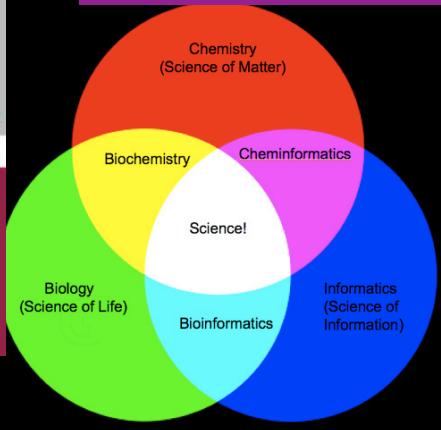
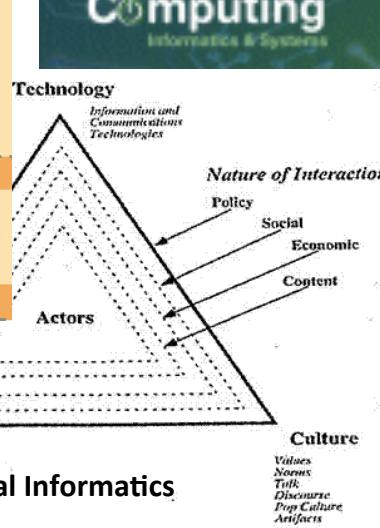
PROBABILISTIC
METHODS
FOR FINANCIAL AND
MARKETING
INFORMATICS



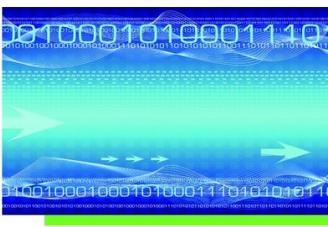
Sustainable
Computing
Informatics & Systems



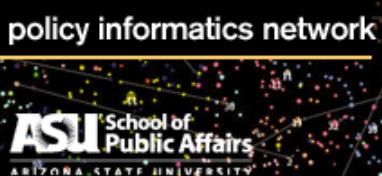
Social Informatics



Opportunities and Challenges
in Crisis Informatics



Noelia Penelope Greer (Ed.)
Business Informatics
Information technology, Management,



USC Center For Energy Informatics

Home Research Publications Smart Grids

GEO Informatics
Knowledge for Surveying, Mapping & GIS Professionals

About the Center

Welcome to the Center For Energy Informatics (CEI) at USC, an Organized Research Unit (ORU) housed in the [Viterbi School of Engineering](#). Energy Informatics is the application of information technologies to energy systems.

Lifestyle Informatics

Applications of Lifestyle Informatics
How is the training classified
Occupation Professions
Further study
Student at the University
Watch the movie
Studying Abroad

Admission and registration
VU Honours Programme

BACHELOR-VOORLEIDINGSDAG
ZATERDAG 3 NOVEMBER

LOOP EEN DAG MEE
MET EEN STUDENT

ENVIRONMENTAL INFORMATICS

Combine body, mind, and soul for a healthier, happier, and more fulfilling life.

ASU School of Public Affairs
Arizona State University

Lifestyle Informatics: Let people live longer, healthier, and more fulfilling lives.

The study Lifestyle Informatics is about studying how people live their lives. This bachelor including applied psychology, ergonomics, and informatics knowledge about language and information processing, how people live their lives better. Lifestyle Informatics: let people live longer, healthier, and more fulfilling lives.

Kmeans Clustering

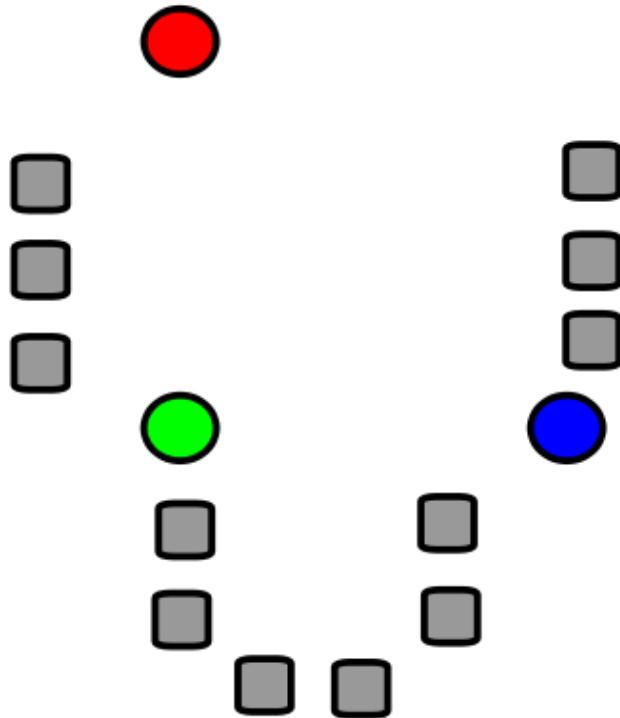
<http://en.wikipedia.org/wiki/Kmeans>

Resources

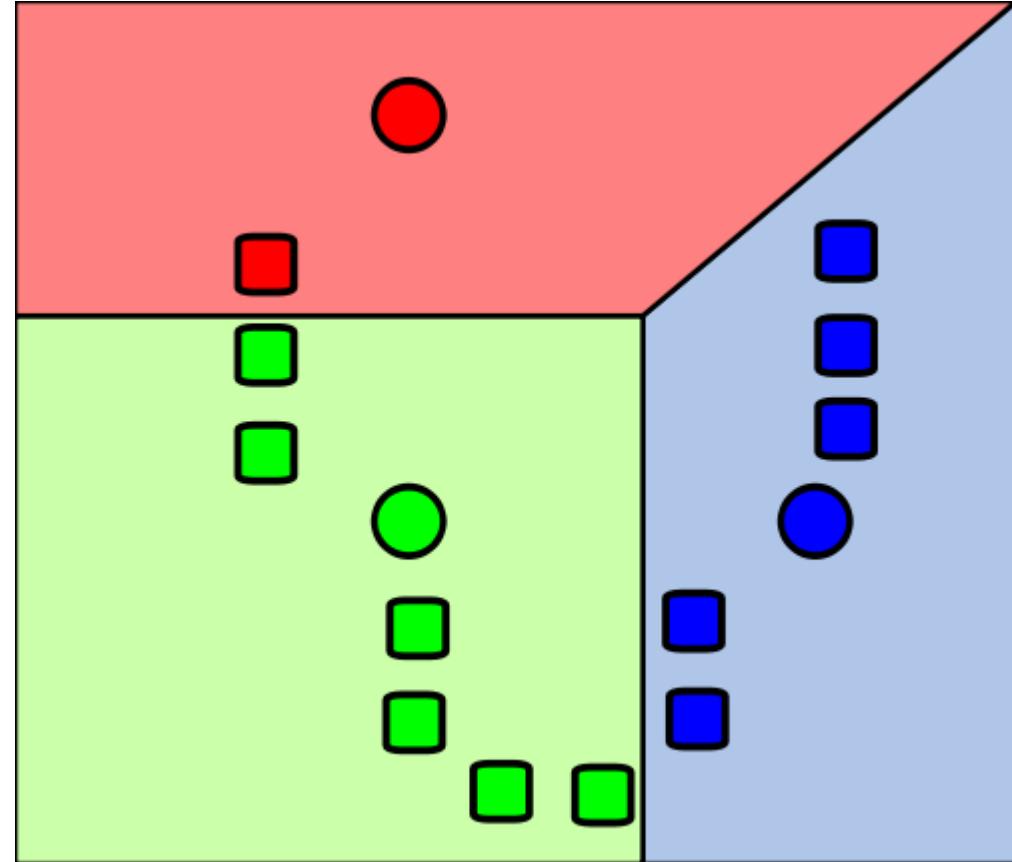
- See
<http://grids.ucs.indiana.edu/ptliupages/publications/Clusteringv1.pdf>
for advanced clustering method improving k means
- This lesson has no code but we will use **PlotViz** program introduced earlier
- And Wikipedia on k means
- Files used are
 - **clusterFinal-M3-C3Dating-ReClustered.pviz**
 - **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

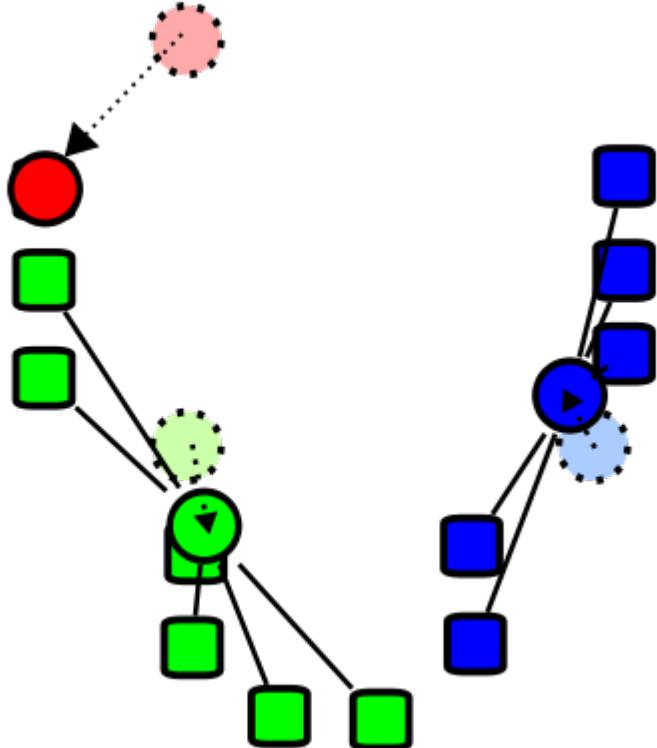
- Choose $k=2$ to 10 million (or more) clusters according to some intuition
- Initialize k cluster centers (this only works in real spaces)
- Then Iterate
 - Assign each point to cluster whose center it is nearest to
 - Calculate new center position as centroid (average in each component) of points assigned to it
- Converge when center positions change $< \text{cut}$
- 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



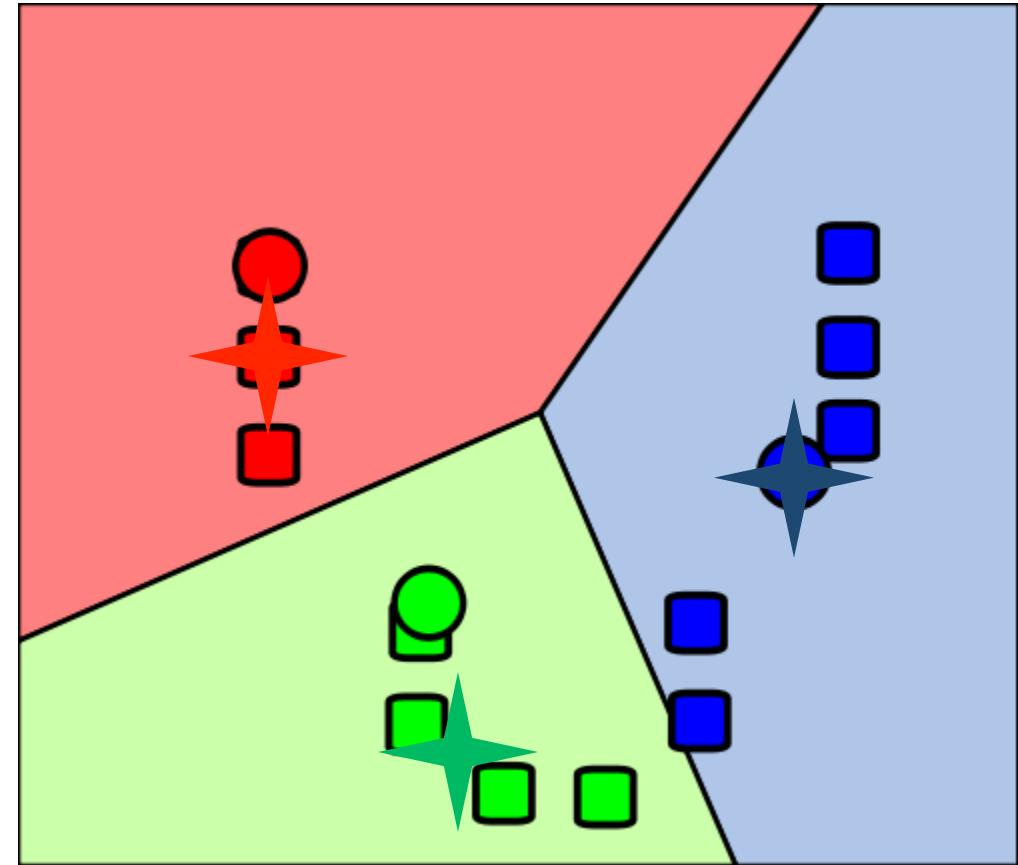
1) k initial "means" (in this case $k=3$) are randomly generated within the data domain (shown in color).



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



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)

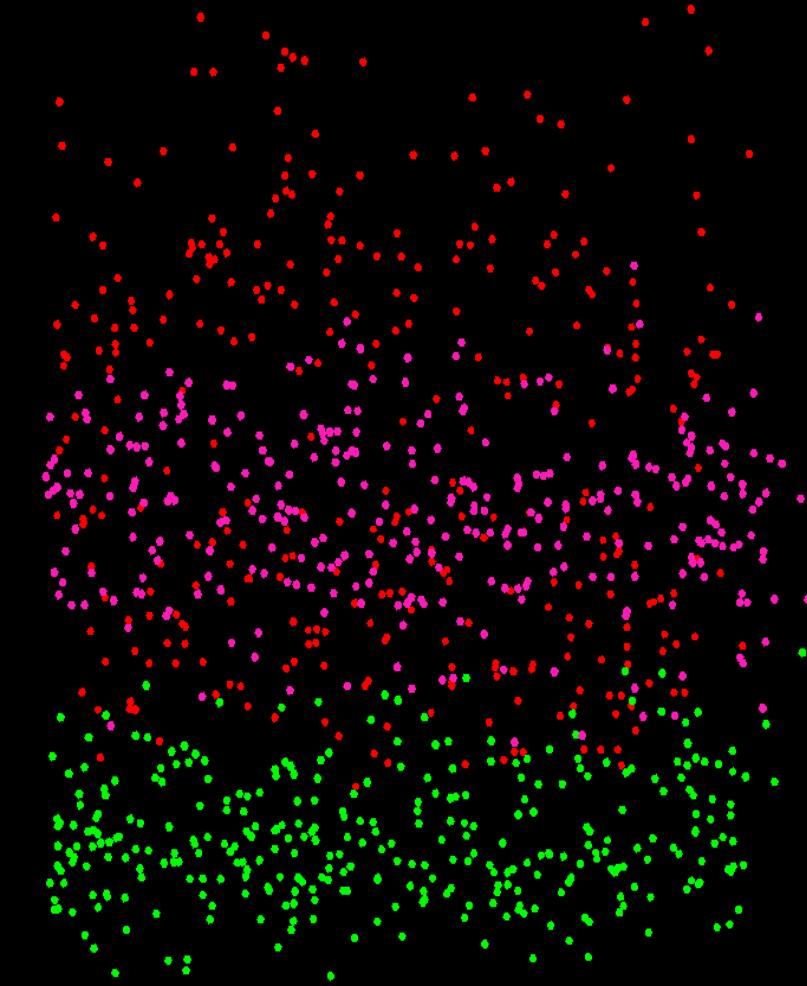
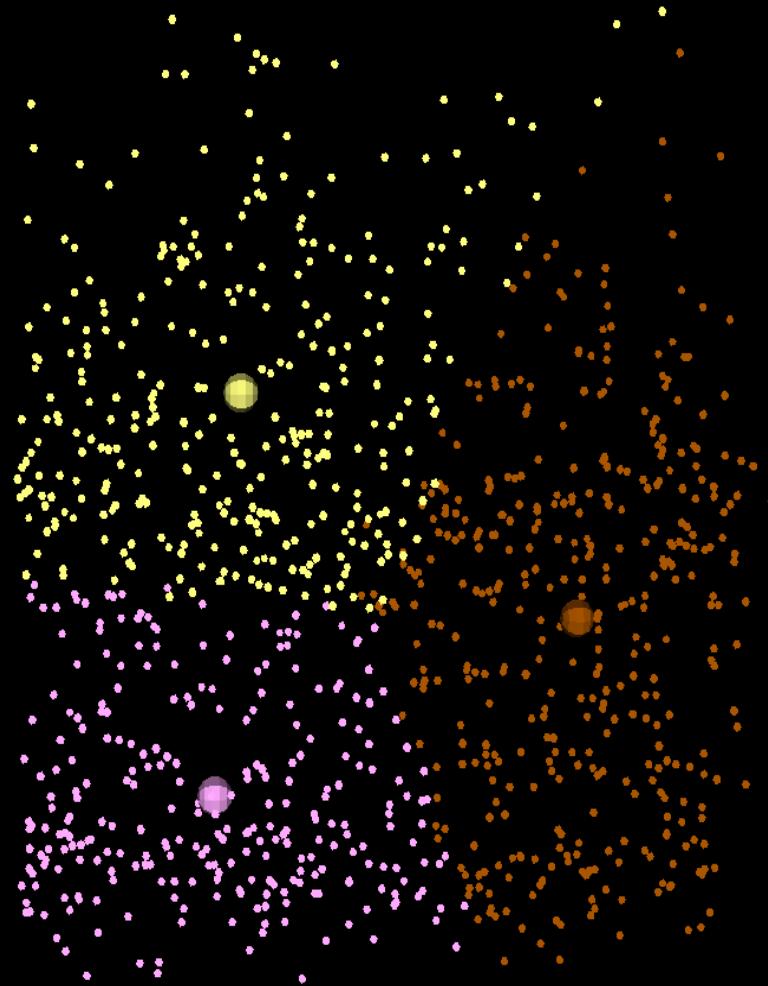
Clustering of Recommender System Example

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

3 Optimal k-means Clusters

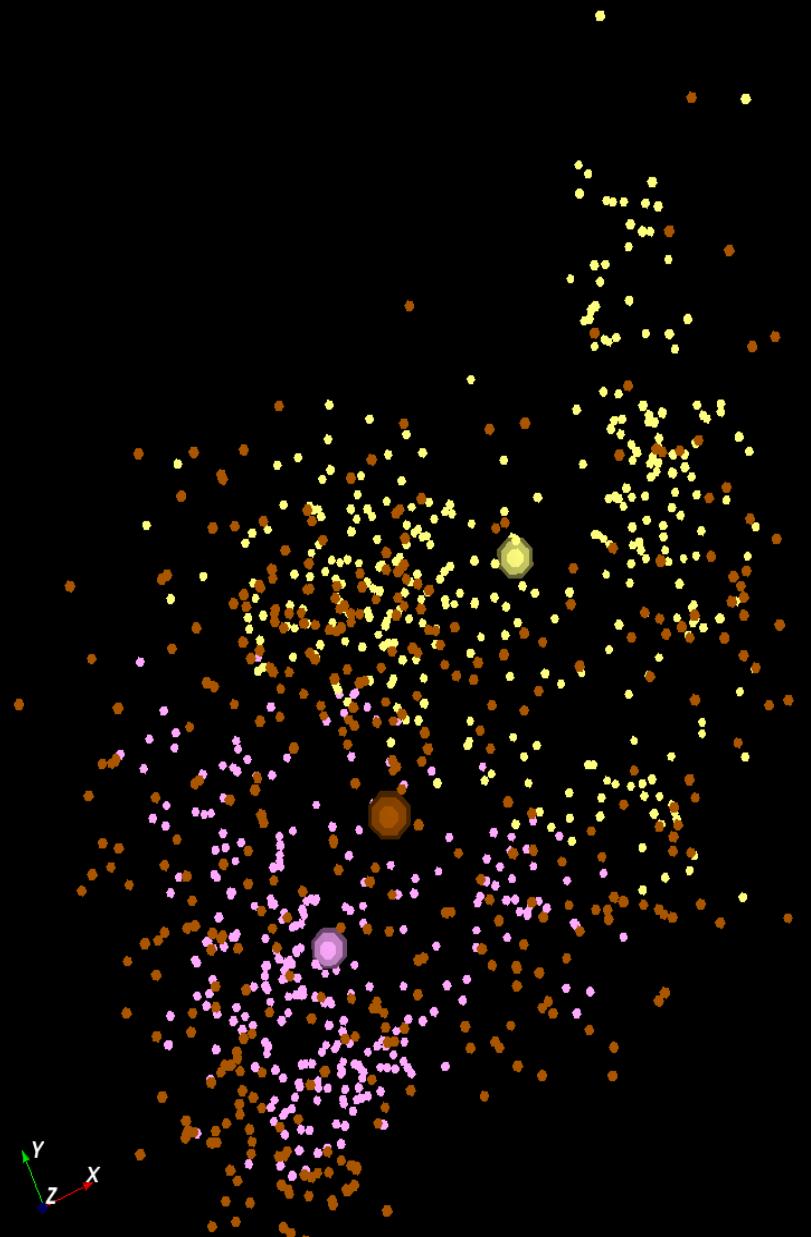
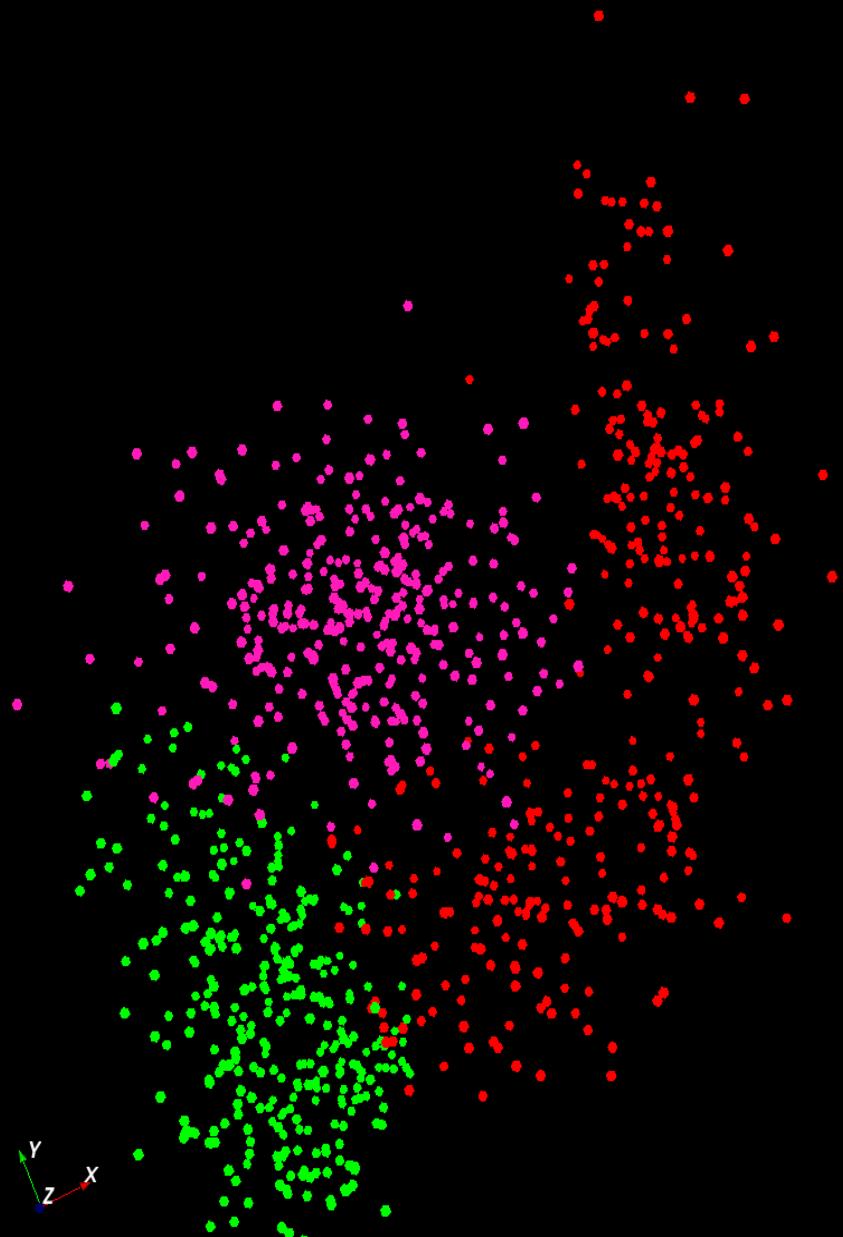
Original labels (Clustering)



Clustering v User Rating I

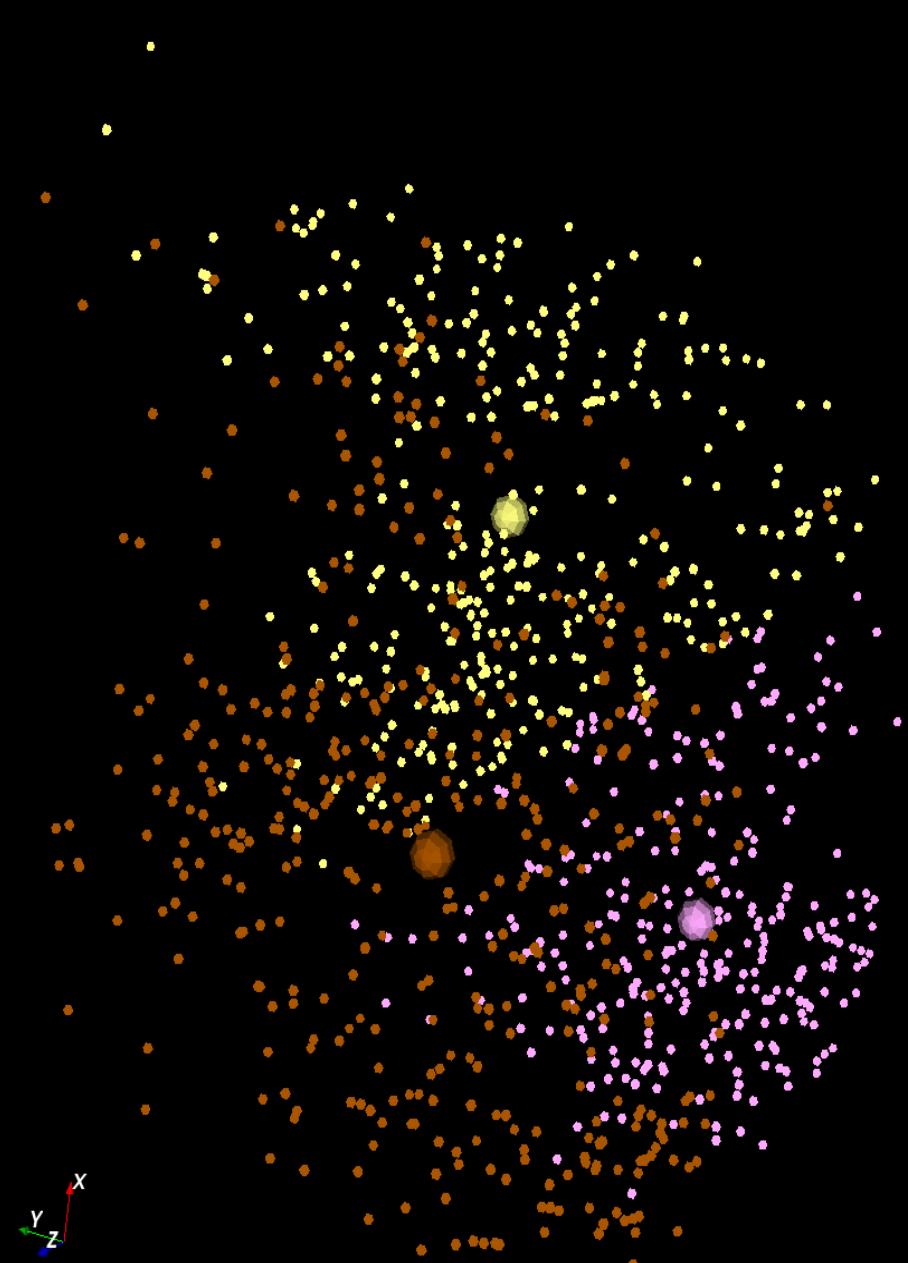
Original labels (Clustering)

3 Optimal k-means Clusters

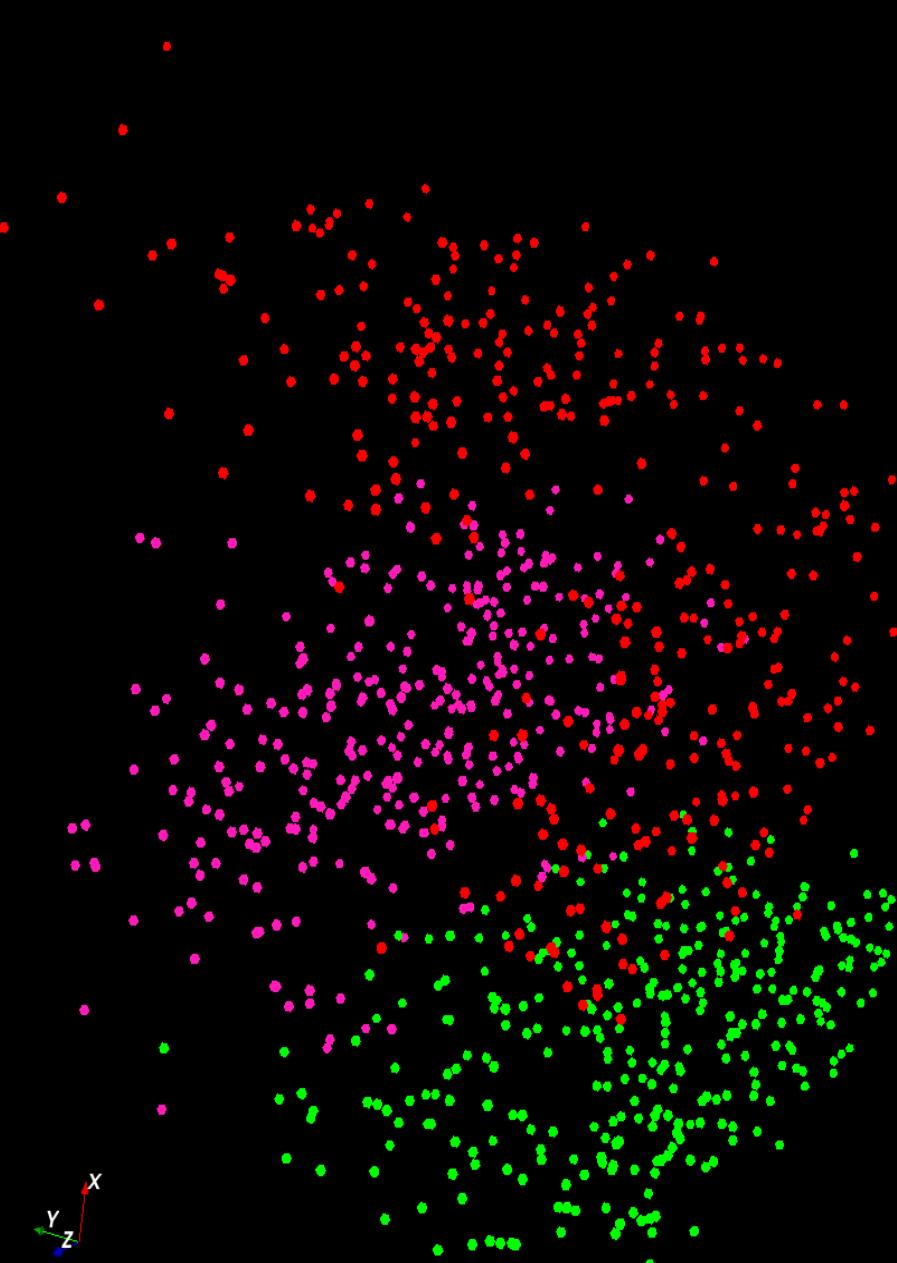


Clustering v User Rating II

3 Optimal k-means Clusters



Original labels (Clustering)



Clustering v User Rating III

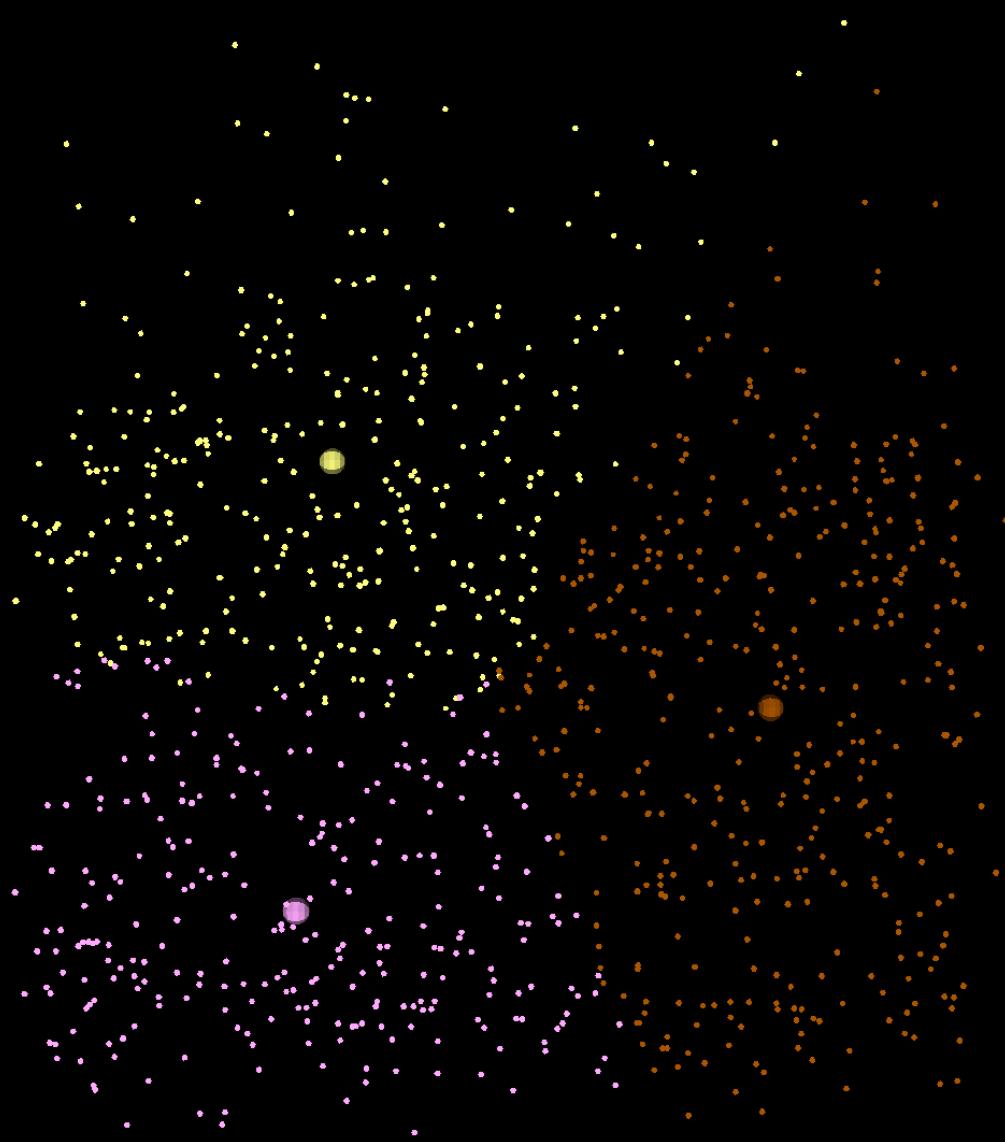
Clustering into more than 3 Clusters

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

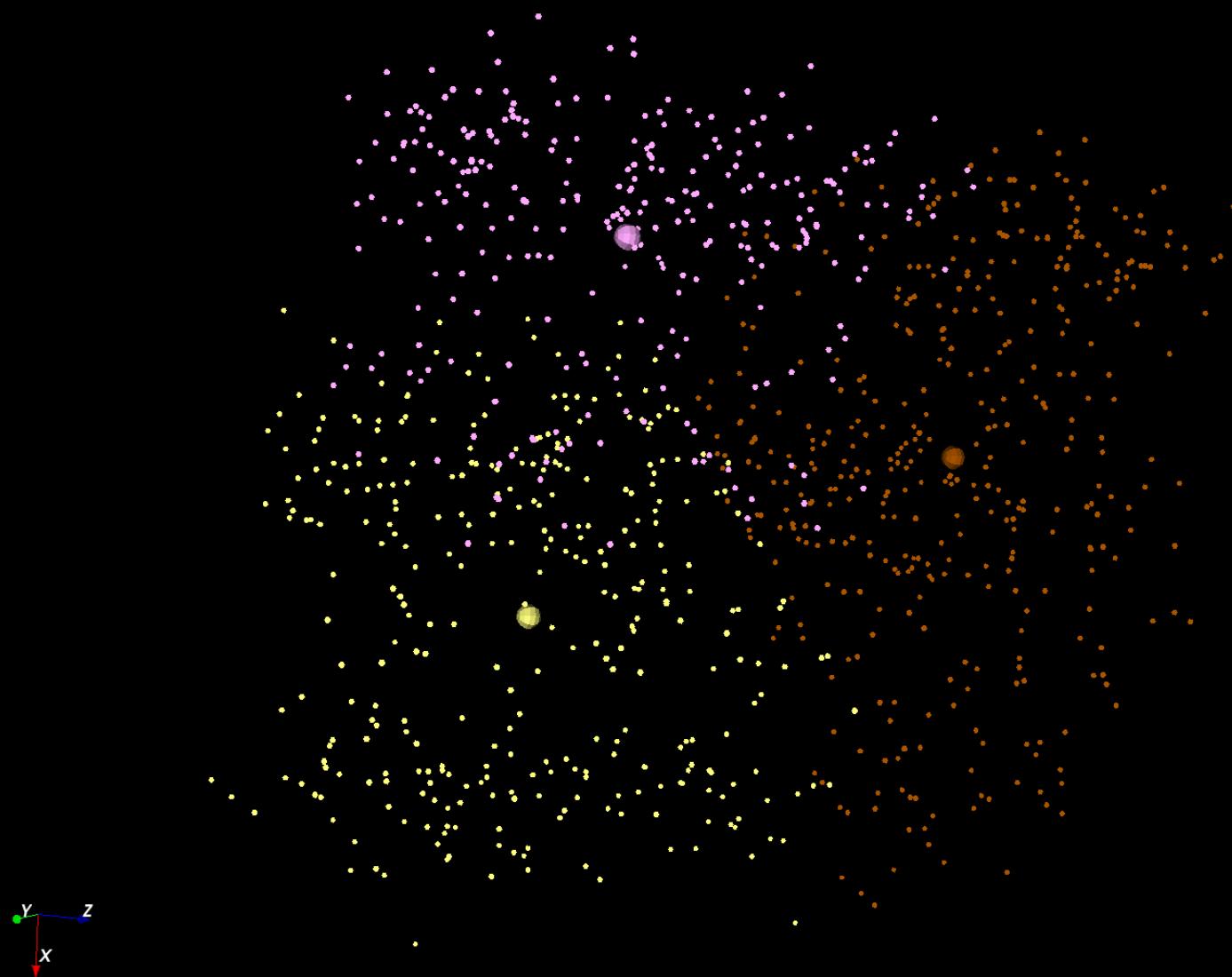
Two views of a 3D Clustering I

3 k means Clusters

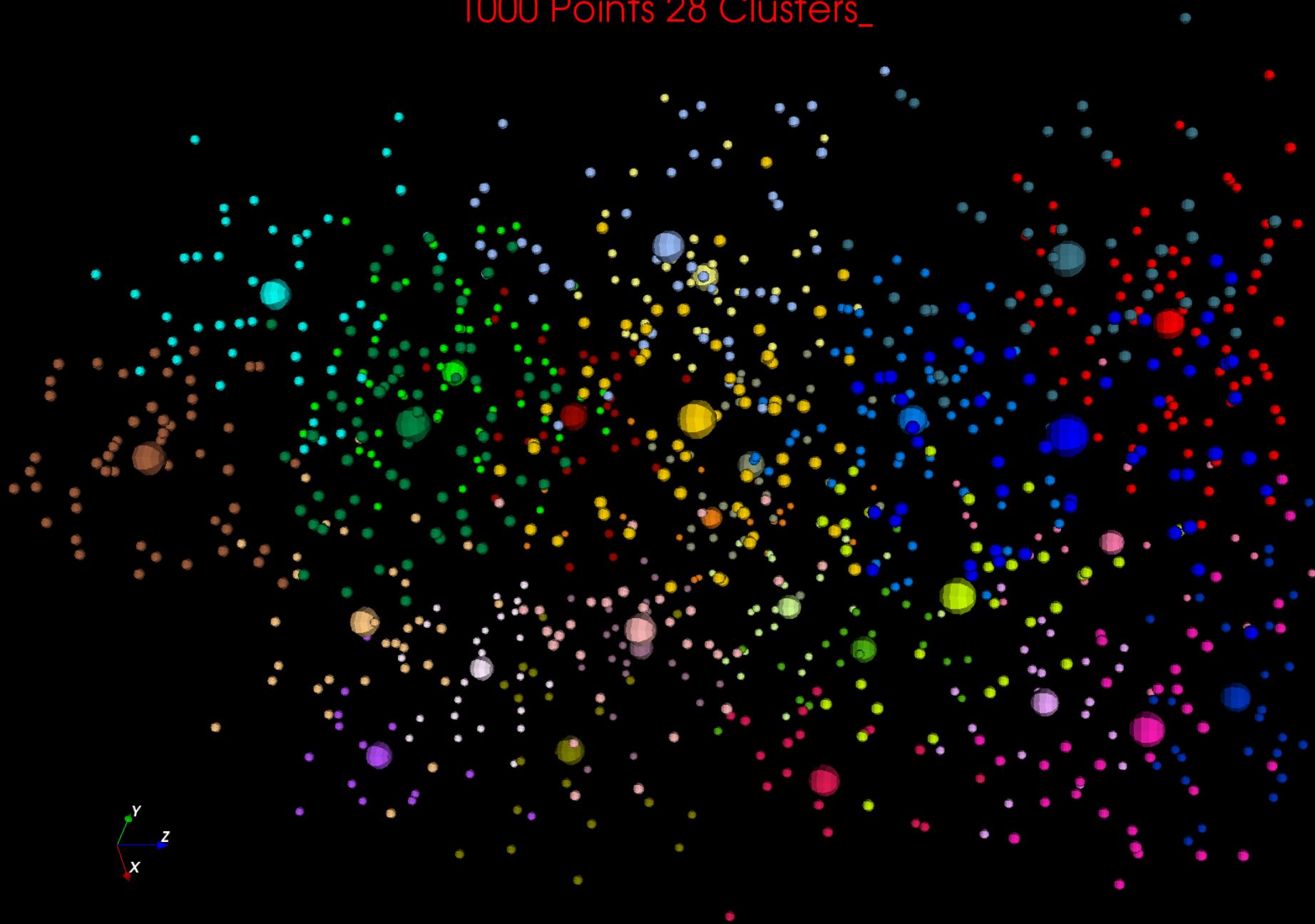


Two views of a 3D Clustering II

3 k means Clusters



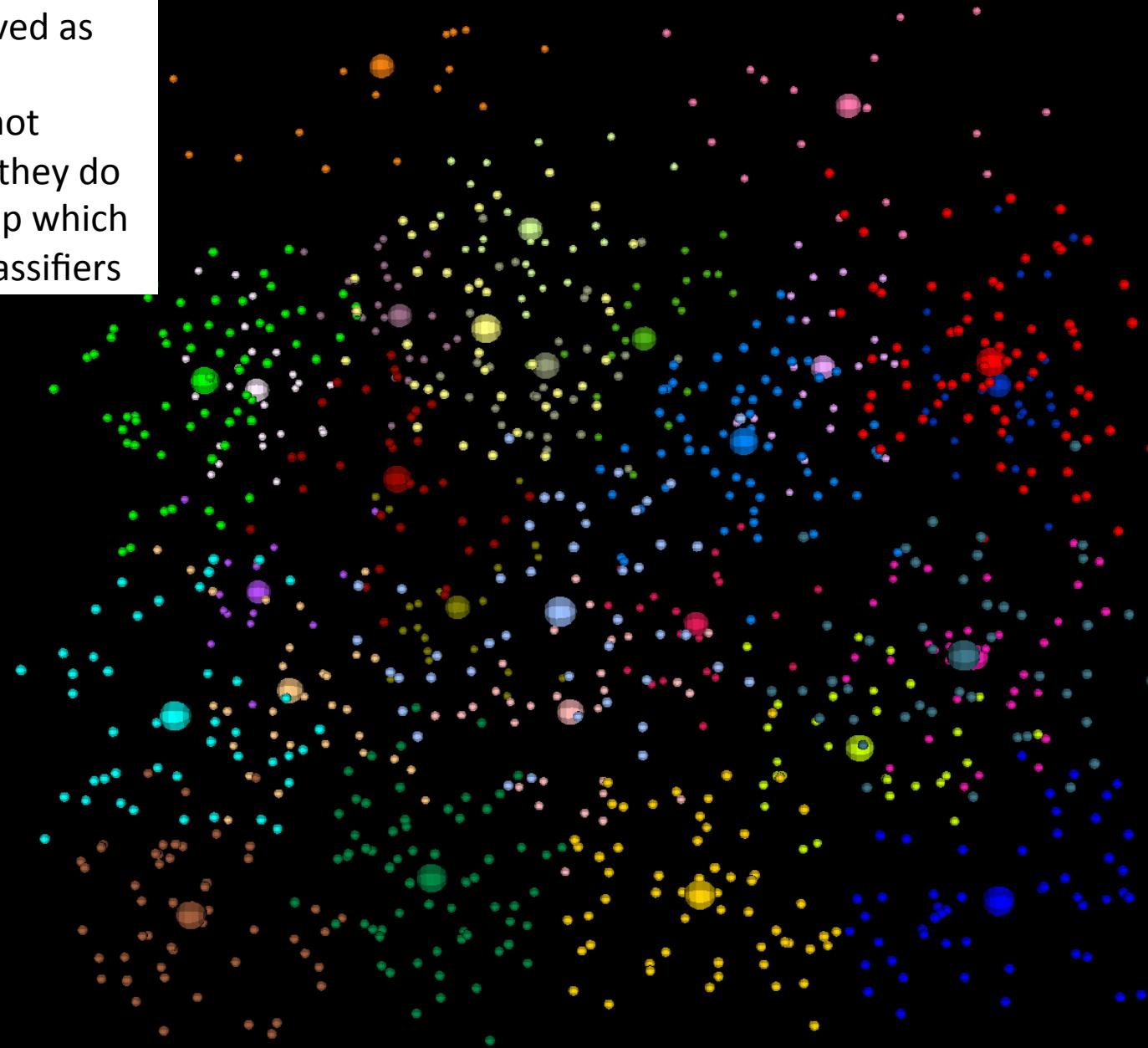
1000 Points 28 Clusters_



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

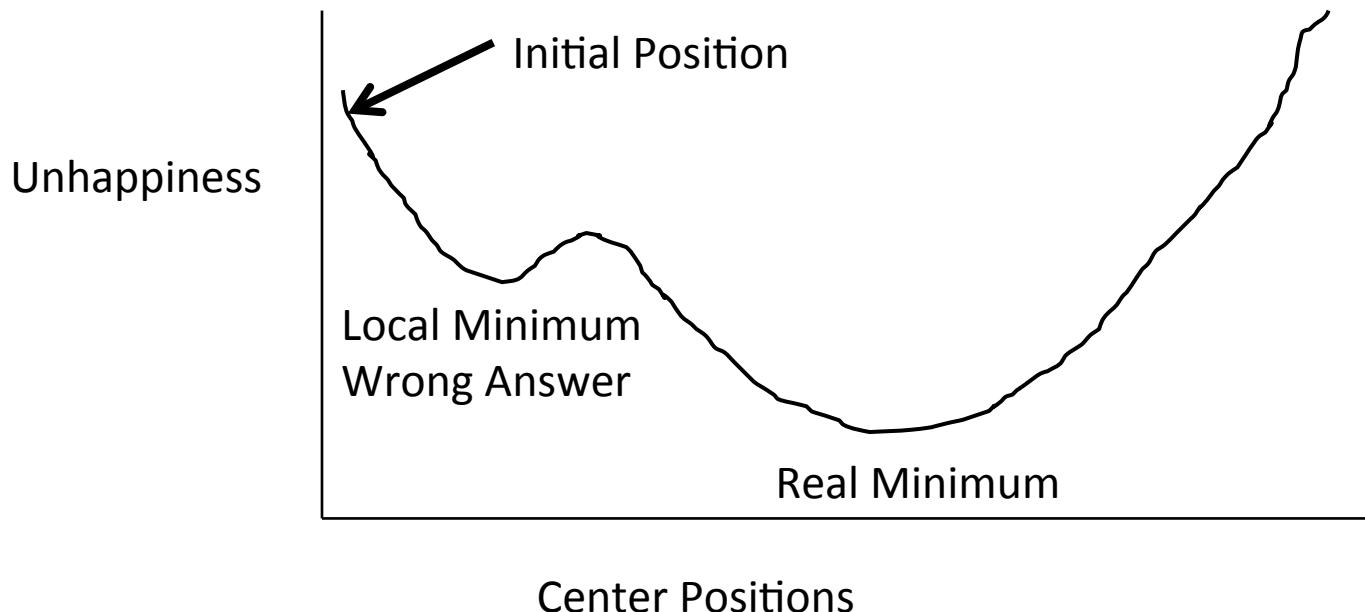
1000 Points 28 Clusters_



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

Why Do I Need to Anneal Beads?



Copyright © 2003 Richard Downton

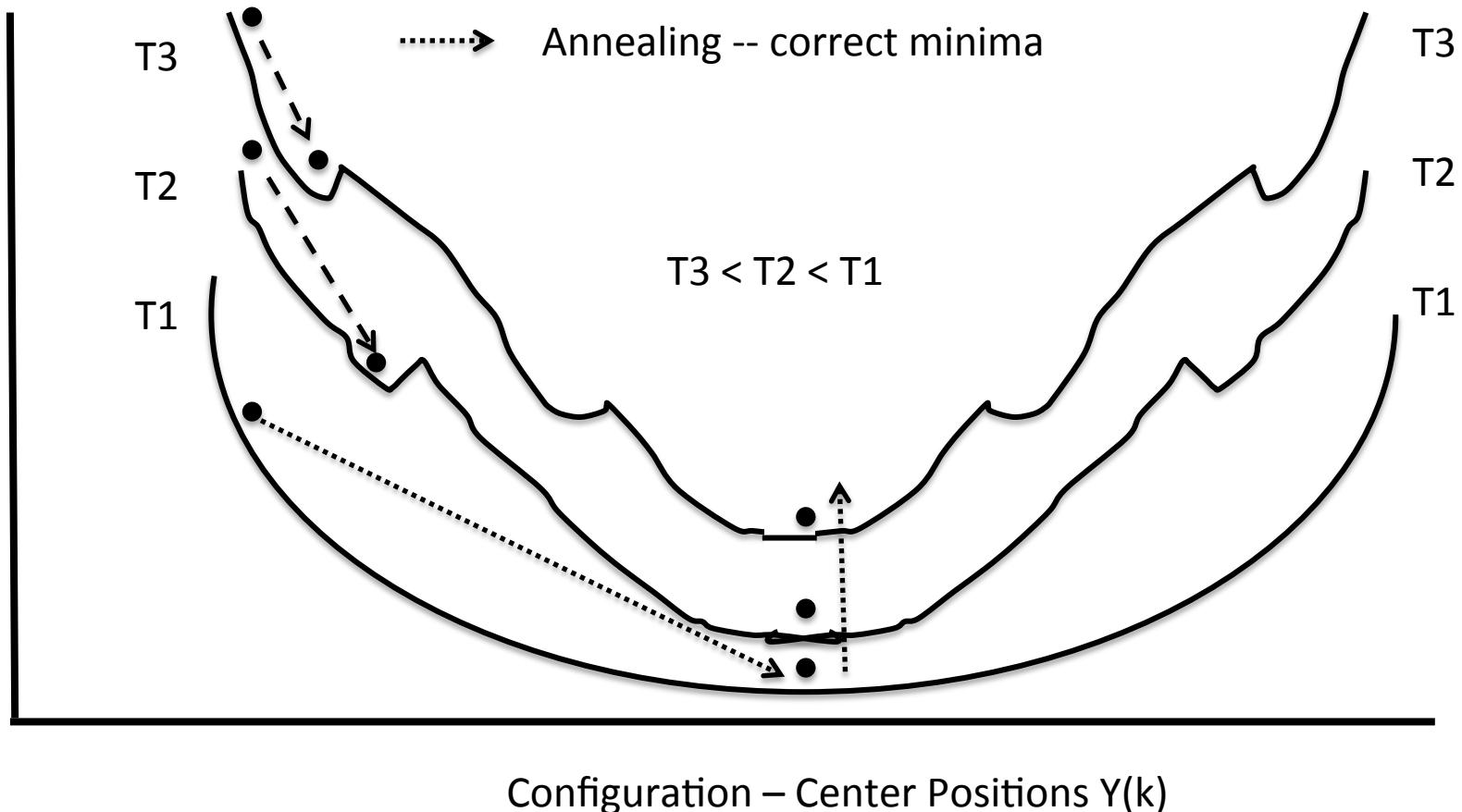


Annealing

Objective Function

---> Fixed Temperature – false minima

.....> Annealing -- correct minima



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

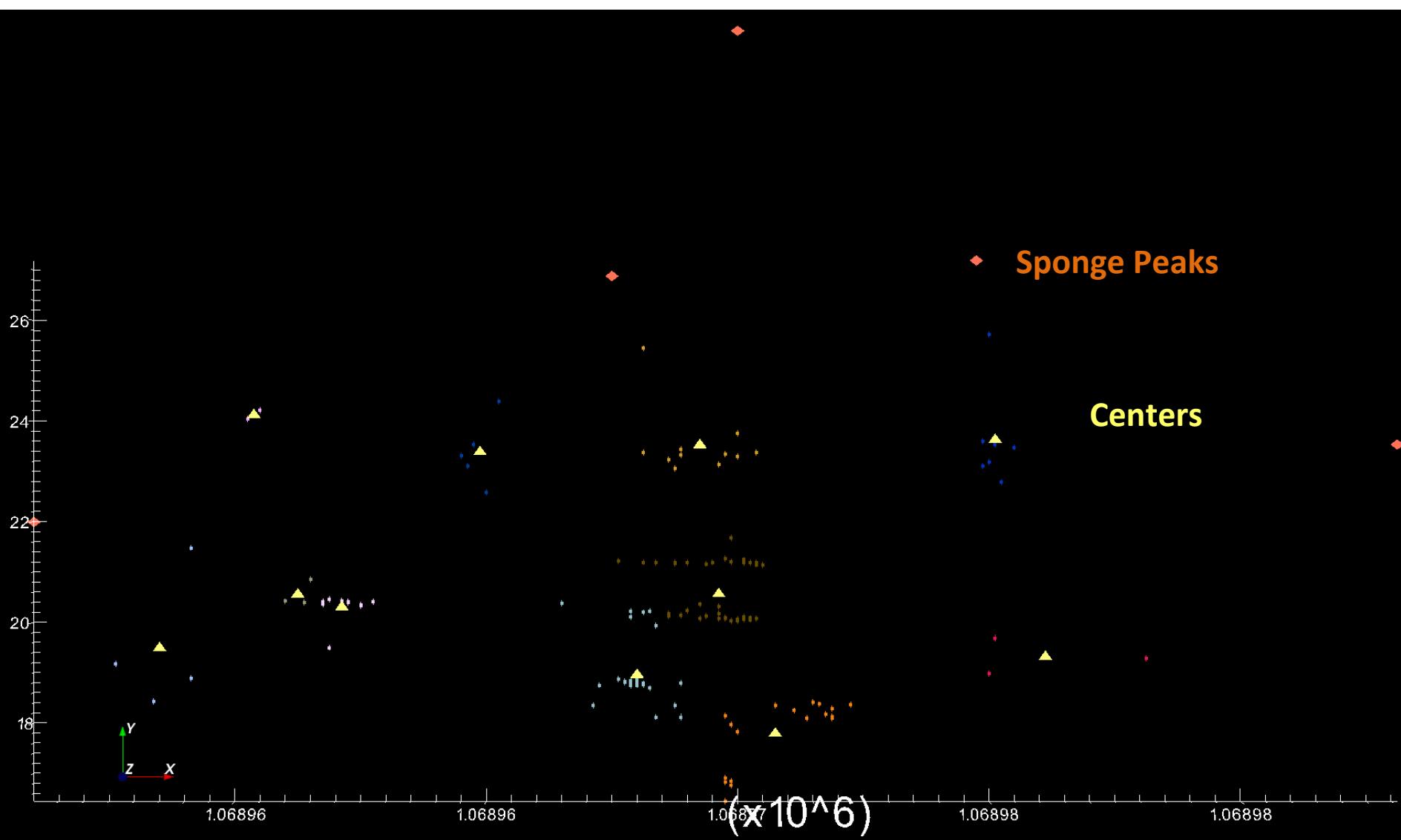
- This is used in several somewhat different ways
- First is taking a set of points in a space and divide into distinct groups with (clear) separation
- Second is dividing points into neighboring points but not groups that are separate
- There are also latent factor and mixture models which use a different 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

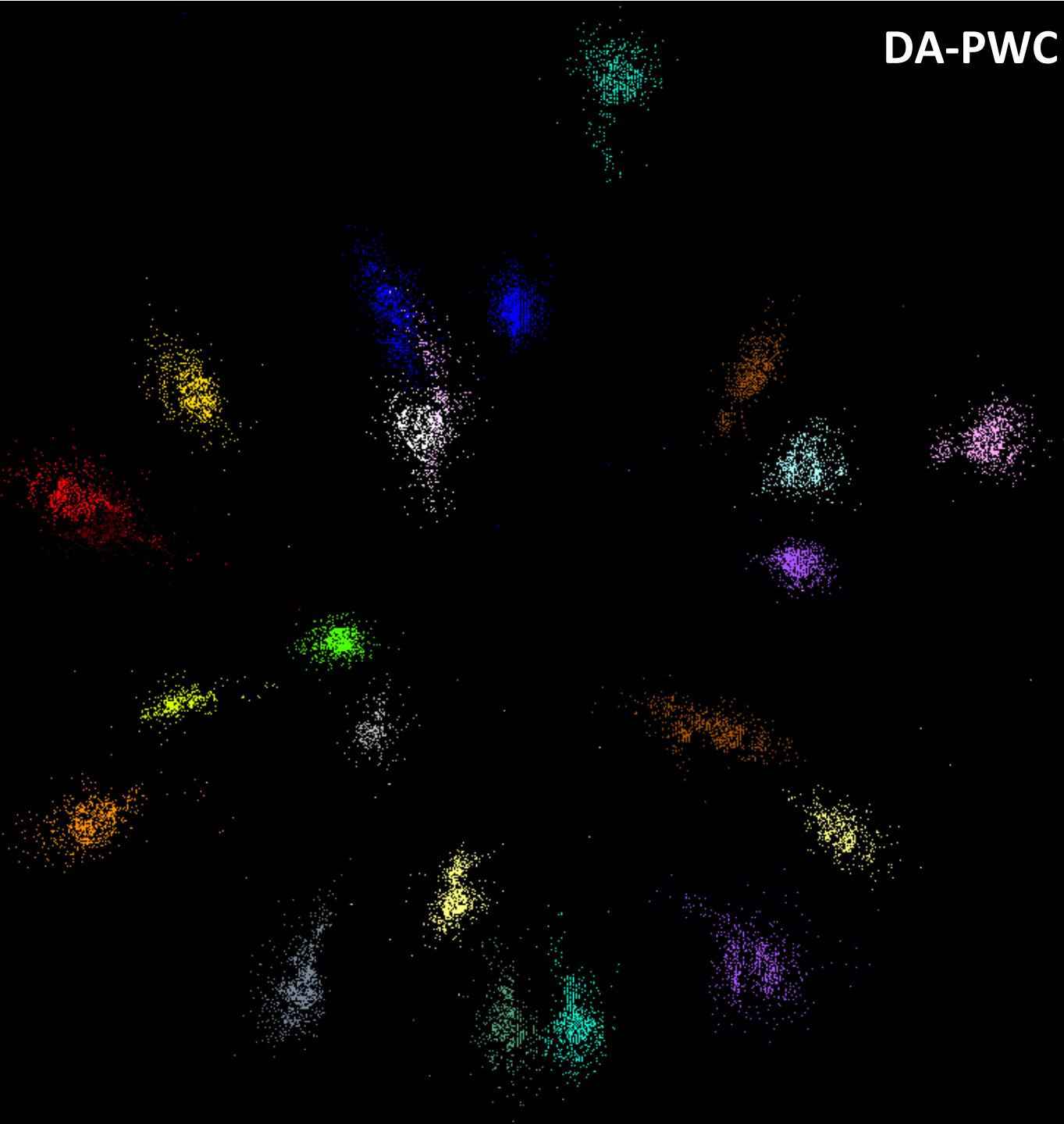
Clustering Examples

- 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





DA-PWC

Metagenomics

26 Clusters in Region 4

Haixu Megaregion 4 73885 Seqs 26 Clusters

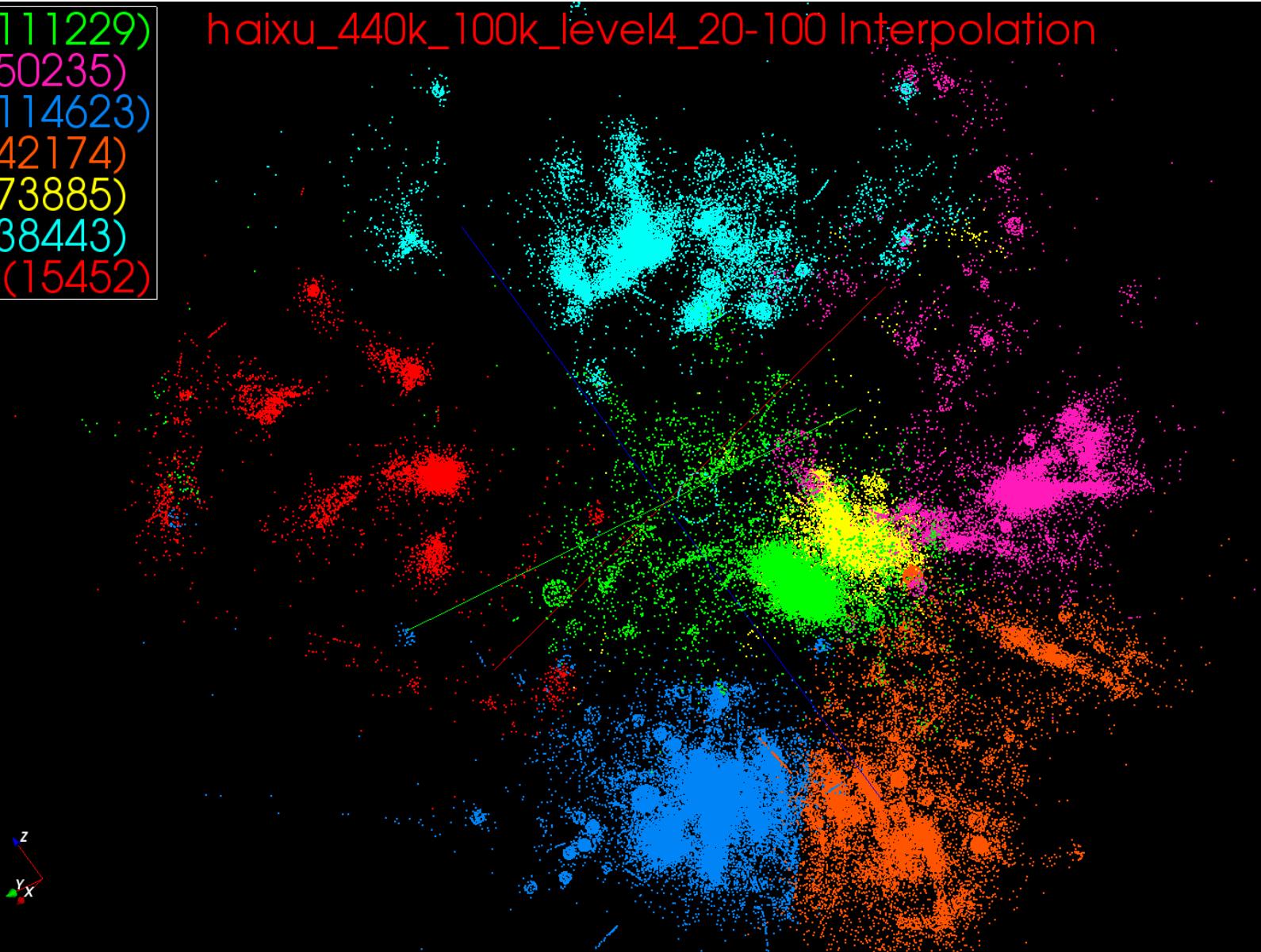
- 0 (15930)
- 1 (4569)
- 2 (673)
- 3 (2514)
- 4 (2089)
- 5 (2156)
- 6 (688)
- 7 (2833)
- 8 (906)
- 9 (569)
- 10 (852)
- 11 (1426)
- 12 (587)
- 13 (472)
- 14 (2219)
- 15 (5496)
- 16 (8557)
- 17 (5394)
- 18 (1337)
- 19 (786)
- 20 (1184)
- 21 (4137)
- 22 (2767)
- 23 (1926)
- 24 (2855)
- 25 (963)



Genomic Data Not fully Clustered

- 2 (111229)
- 3 (50235)
- 5 (114623)
- 6 (42174)
- 8 (73885)
- 9 (38443)
- 10 (15452)

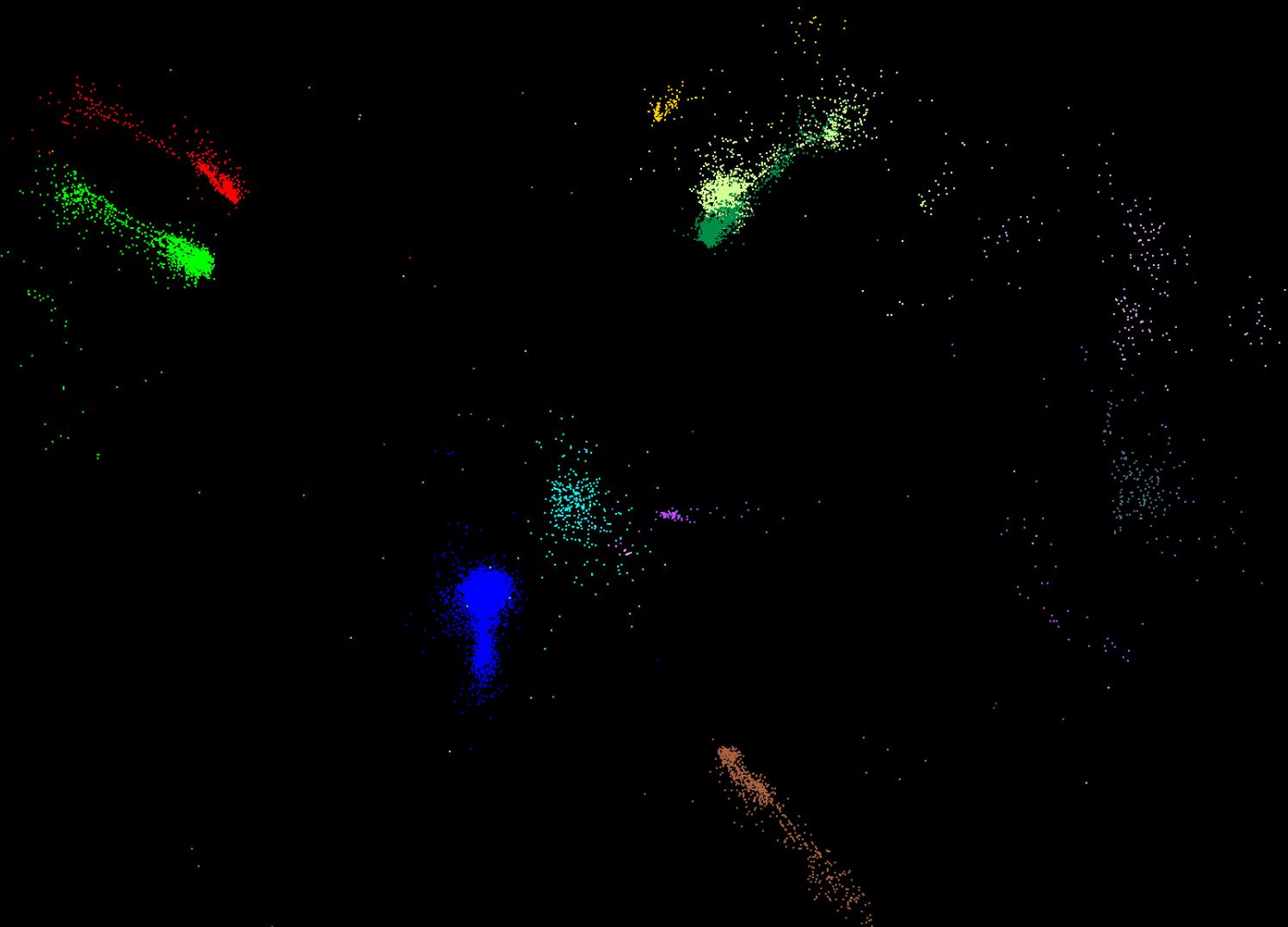
haixu_440k_100k_level4_20-100 Interpolation



13 Clusters in Region 6

Haixu Megaregion 6 15452 Seqs 13 Clusters

- 0 (7384)
- 1 (744)
- 2 (2246)
- 4 (137)
- 6 (1507)
- 7 (753)
- 8 (336)
- 9 (159)
- 10 (210)
- 11 (1797)
- 12 (179)



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