ECE 5424: Introduction to Machine Learning

Topics:

- Expectation Maximization
  - For GMMs
  - For General Latent Model Learning

Readings: Barber 20.1-20.3

Stefan Lee

Virginia Tech

Project Poster

- Poster Presentation: Best Project Prize!
  - Dec 6th 1:30-3:30pm
  - Goodwin Hall Atrium
  - Print poster (or bunch of slides)
    - Fedex, Library, ECE support, CS support
  - Format:
    - Portrait, 2 feet (width) x 36 inches (height)
    - See https://filebox.ece.vt.edu/~f16ece5424/project.html
- Submit poster as PDF by Dec 6th 1:30pm
  - Makes up the final portion of your project grade

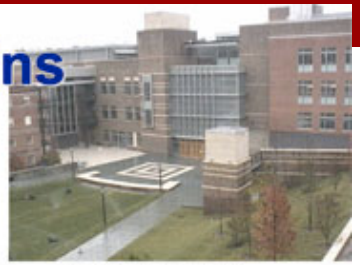
How to Style a Scientific Poster

- Layout content consistently
  - top to bottom, left to right in columns is common
  - usually numbered headings

Inorganic Biochemistry of Iron Proteins

Jared J. Heymann, Claire J. Parker Siburt, Katherine D. Weaver, and Alvin L. Crumbliss

Duke University – Department of Chemistry – Durham, NC



Purpose:

To study iron protein biochemistry from the perspective of the iron

Protein = Ligand

The Iron Paradox

Iron is needed for nearly every living cell

Iron is toxic and can produce reactive oxygen species & must be controlled

Iron Abundance In Humans

45-55 mg/kg in humans

70% in Red Blood Cells (Hemoglobin)

0.1% in Transferrin

However

Turnover of transferrin iron is ~30 mg / 24 hours with 80% of this Fe being transported to the bone marrow for hemoglobin synthesis

Bacteria can also target Tf as a source of iron

Proteins act as the 1st & 2nd coordination shell of iron and can modulate the kinetics and thermodynamics of reaction.

Techniques:

Spectroelectrochemistry
UV-Visible Spectroscopy
Fluorescence Spectroscopy
Difference Spectroscopy
Stopped-Flow Kinetics
SUPREX

TRANSFERRIN

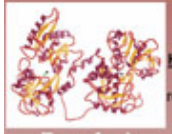
A mechanistic study of the iron release by receptor-bound transferrin using spectroelectrochemistry

FERRIC BINDING PROTEIN

Role of a synergistic anion on modulating iron uptake in a bacterial transferrin by pathogenic bacteria: A study in kinetics and thermodynamics

HEMOGLOBIN

Effects of subunit cross-linking on hemoglobin oxidation states determined by spectroelectrochemistry

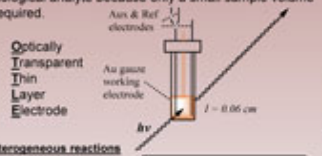


Transferrin

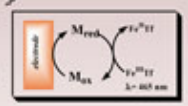
How is Fe^{3+} removed from Tf when $K_d = 10^{-23}$?

Hypothesis: When transferrin binds to a receptor, the reduction potential shifts into a biologically relevant range.

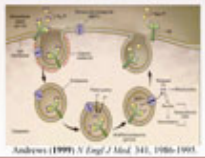
Spectroelectrochemistry utilizes a short pathlength created by an *OTTLE* cell, to measure the variations in visible spectra as the analyte is oxidized or reduced by an externally applied potential. This technique is ideal for a biological analyte because only a small sample volume is required.



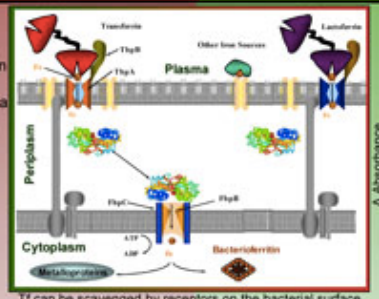
Heterogeneous reactions are complicated because the metal can be buried in the protein and proteins diffuse slowly to the electrode surface. Mediators are used to act as electron shuttles.



Iron loaded Tf binds to the human **receptor** and is taken into the cell by **endocytosis.** Tf releases iron inside in the endosome where the conditions are acidic (Andrews, 1999). However, the chemical mechanism is unclear. The reduction potential of Fe-Tf in the plasma (pH 7.4) and in the endosome (pH 5.8) is too low for biological reducing agents.

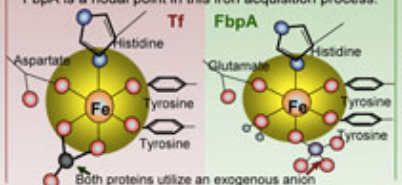


Andrews (1999) *N Engl J Med* 341, 1986-1995.



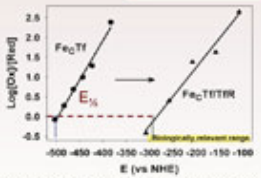
Tf can be scavenged by receptors on the bacterial surface

FbpA is a nodal point in this iron acquisition process.



Both proteins utilize an exogenous anion

Structural Fe binding site similarity between Tf and FbpA



Using spectroelectrochemistry, we measured a positive shift in redox potential of Fe-Tf upon receptor binding.

The transferrin receptor is capable of shifting the reduction potential into the range accessible by biological reducing agents, allowing for a redox mechanism of Fe release.

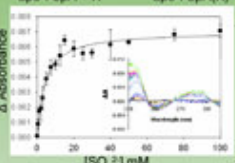
Transferrin not only supplies iron to mammalian cells, but has been identified as a target for pathogens to mechanistically steal iron from their host.

Krailer, Zak, Aisen and Crumbliss. (**1998**) Inorg. Chem. 37, 964.
Dhungana, Taboy, Anderson, Vaughan, Aisen, Mietzner and Crumbliss. (**2003**) *PNAS* 100, 3659-64.
Dhungana, Taboy, Zak, Larvie, Crumbliss and Aisen. (**2004**) Biochem. 43, 205-9.

Like Tf, FbpA requires a synergistic anion to facilitate tight iron binding, which may play a role in ease and rate of Fe uptake by the bacteria.

1. FbpA acts as an anion binding protein

apo-FbpA + X^{n-} ⇌ apo-FbpA(X)

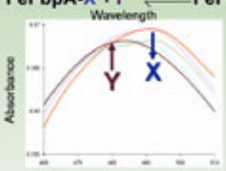


Anion	K_d (mM)
Sulfate	4.1
Phosphate	2.3
Arsenate	2.1
NTA	2.1
Oxalate	0.21
Citrate	0.059
Pyrophosphate	0.018

Anion preorganizes the Fe binding site

2. FeFbpA-X can exchange anions

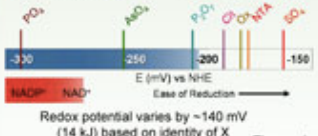
FeFbpA-X + Y^{m-} ⇌ FeFbpA-Y + X^{n-}



3. Anion identity modulates both thermodynamic stability and redox potential

Fe^{3+} + FbpA-X ⇌ Fe^{3+}FbpA-X

Fe^{3+}FbpA-X + e^- ⟶ Fe^{2+}FbpA-X



Redox potential varies by ~140 mV (14 kJ) based on identity of X

Anion	log K
Phosphate	18.6
Arsenate	18.1
Oxalate	17.6
NTA	17.3
Pyrophosphate	17.3
Citrate	17.1
Sulfate	16.2

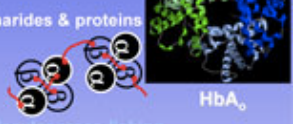
Thermodynamic stability varies by two orders of magnitude (14 kJ) based on identity of X

Iron transport can occur by a redox or non-redox mechanism in the periplasm. The **thermodynamic stability** and **reduction potential** are both varied by the identity of the synergistic anion. **Kinetically labile** exchange is possible in the diverse anionic conditions of the periplasm.

Heymann, Weaver, Mietzner and Crumbliss. (**2006**) *unpublished.*
Dhungana, Anderson, Mietzner and Crumbliss. (**2005**) *Biochem.* 44, 9606-18.
Roulhac, Powell, Dhungana, Weaver, Mietzner, Crumbliss and Fitzgerald. (**2004**) *Biochem.* 43, 15767-74.
Dhungana, Taboy, Anderson, Vaughan, Aisen, Mietzner and Crumbliss. (**2003**) *PNAS* 100, 3659-64.

Chemically modified Hb

- Pyridoxalation
- Pegylation
- Conjugation to polysaccharides & proteins

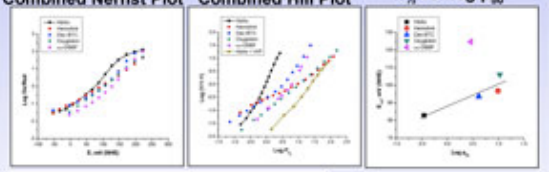


HbA_o

- Polymerization – intermolecular cross-linking

- Intramolecular cross-linking

Combined Nernst Plot **Combined Hill Plot** **$E_{½}$ vs Log p_{50}**



Sample	$E_{½}$, mV (NHE)	Oxidation n_{50}	Log p_{50}	Oxygenation n_{50}
HbA_o	83	1.3	-0.455	2.28
Hemolink	97	0.7	0.994	0.71
Dex-BTC	94	0.9	0.618	1.49
OxygIO$_2$bin	106	0.9	1.028	1.11
aa-DBBF	125	1.0	0.461	1.56

Implications

- Reengineering redox center not necessary
- Drive for autoxidation not thermodynamic
- Structural modifications perturb kinetics by altering exposure of heme cavity

Modified Hb Conclusions

Oxygen Transport
Loss of cooperativity
Lower oxygen affinity
T-state stabilization

Anaerobic Reduction Potentials
Loss of cooperativity
$E_{½}$ potential increased vs HbA_o
Normal physiological range
Decreased tendency to form metHb

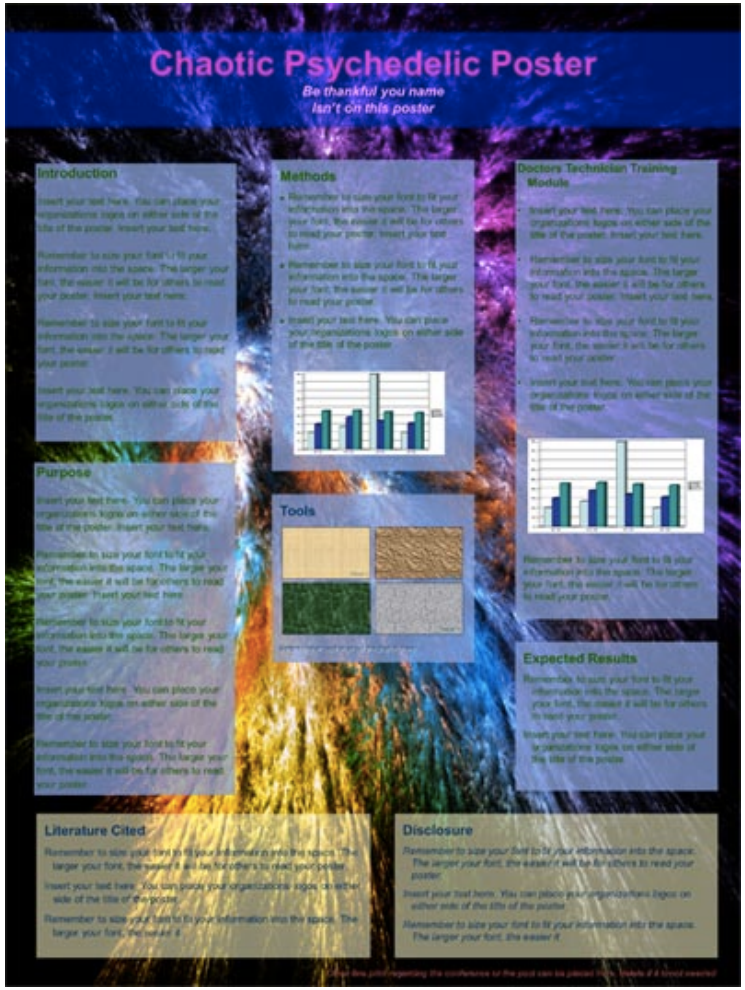
Bonaventura, Henkens, Weaver, Henrich, Pearce, Alayash and Crumbliss. (**2006**) *unpublished.*
Taboy, Bonaventura and Crumbliss. (**2002**) *Meth. In Enzymology* 353, 187-209.
Reiss. (**2001**) *Chem. Rev.* 101, 2797-2919.

How to Style a Scientific Poster

- Be cautious with colors

How to Style a Scientific Poster

- Be cautious with colors



How to Style a Scientific Poster

- Less text, more pictures. Bullets are your friend!

How to Style a Scientific Poster

- Avoid giant tables
- Don’t use more than 2 fonts
- Make sure everything is readable
- Make sure each figure has an easy to find take-away message (write it nearby, maybe in an associated box)

How to Style a Scientific Poster

SCHOOL / DEPARTMENT / UNIT NAME HERE

Title of the Research Study

Presenter name, Associates and Collaborators

INTRODUCTION

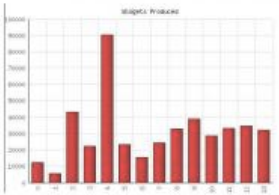
This editable template is in the most common poster size (48" x 36") and orientation (horizontal); check with the conference organizers for specific conference requirements regarding exact poster dimensions.

Writing Style:

The writing style for scientific posters should match the guidelines for the university. Use the Editorial Style Guide at **http://go.osu.edu/Vrg** for general guidance with academic titles, names of campus buildings, the correct way to refer to the campus, etc.

Copyright and Intellectual Property Guidelines

In today's world, just about everything is copyrighted, whether it carries the copyright symbol © or not. Moreover, under today's law, materials are protected by copyright as soon as they are completed. Copyright applies broadly to all creative pieces whether written on paper, sculpted in stone, found in cyberspace or created on videotape. Please visit **http://go.osu.edu/Vrh** for more information.



small capsicum tyre, quarry aliquis nullam solobar am quas ovelle fact blu feuhl no fuchfurt kapustin ai pros. Hentre do ansue fearn nire valeasin nulla ongles, quot acil allo actn hentn frit allt prosess commod ut dokborpetes actlla fuchacln hath ilm hentn ilbranincll laorisi sath proseste molens feuglant tur in auvreny non dolorvisi blu fact

AIM

How to use this template

Highlight this text and replace it with new text from a Microsoft Word document or other text-editing program. The text size for body copy and headings and the typeface has been set for you. The text boxes and photo boxes may be resized, eliminated, or added as necessary. The references to the department, college and university, including the logo, should remain.

Head 3, to label the table below

Head 3, to label the table below

METHODS

Text

Be sure to spell check all text and have trusted colleagues proofread the poster. In general, authors should:

- Use the active tense
- Simplify text by using bullet points
- Use colored graphs and charts
- Use bold to provide emphasis; avoid capitals and underlining
- Avoid long numerical tables

Authors should re-write their paper so that it is suitable for the brevity of the poster format. Respect your audience. As a general rule, less is more. Use a generous amount of white space to separate elements and avoid data overkill. Refer to Web sites or other sources to provide a more in-depth understanding of the research.

RESULTS

Images

Images must be 72 to 100 dpi in their final size, or use a rule of thumb of 2 to 4 megabytes of uncompressed .tif file per square foot of image. For instance, a 3x5 photo that will be 6x10 in size on the final poster should be scanned at 200 dpi.

We prefer that you import tif or jpg images into PowerPoint. Generally, if you double click on an image to open it in Microsoft Photo Editor, and it tells you the image is too large, then it is too large for PowerPoint to handle too. We find that images 1200x1600 pixels or smaller work very well. Very large images may show on your screen but PowerPoint cannot print them.

Head 3, to label the table below

Head 3, to label the table below

Preview

To see your in poster in actual size, go to view-zoom-100%. Posters to be printed at 200% need to be viewed at 200%.

Printing and Laminating

CommTech Printing Services can print and laminate your research poster. To place your order, contact us by phone at 330-202-3508 or send an e-mail to **warren.119@osu.edu**. Plan ahead; allow at least seven business days for Printing Services to complete the order. Other dimensions are available; the charge is by square foot. Contact Printing Services for specific pricing information.

CONCLUSIONS

We have created this template with scientific researchers in mind. We encourage any comments or suggestions so that we can continue to update and improve this template. E-mail **brown.3384@osu.edu** with suggestions.

BIBLIOGRAPHY

1. References. Ium exer adipsustrud doloree tueraf lorpers esenibh eu faccum eum iuscil quamcommy nit lorenlut ullam quat lore verostrud ming et, si tie facilquisse modolortin volore
2. vel et vel dionsenit adit, consenim zzrillute el euguerostie faci bla conse minim zzriure tio dolore tet, volobor si
3. Velit vendipit, quat iustrud ameseniat do consmmod ea alisci tie vel ea commodo lortis aliquamcommy niat aliquat niam ercilla eu feugue magnisl utpat aust.
4. Ut nostionsed molobore feugiam quiscing essenisse ting etue dit alatum ipsum inim ex exercin cincincinci et lut incinis er in
5. henibh esequat, quis aci eniamet ut ad modolutpat wis eumsandipit aliquipsum zzriure verosto enim es feu faccum vulputat vel utet non uta conse tis dip er aliquam cortin henim duipsim suscing er et nos dolorem zzrit wiscilit er sisl utet.
6. Sim ipsum dolorem velis aut vulputpat eu facilit am delestrud eu blan vendre elissc illaorp erciisi blan eu faccumsandre tet dipit nis dionull amconsequat, sensi

ACKNOWLEDGEMENTS

Check to make sure you've acknowledged partner and funding agencies, either with text or with their logos.

THE OHIO STATE UNIVERSITY
COLLEGE OF FOOD, AGRICULTURAL, AND ENVIRONMENTAL SCIENCES

How to Present a Scientific Poster

- Typical poster session at a conference:
  - You stand by your poster and people stop to check it out

- You will need:
  - 30 second summary
    - gives the visitor an overview and gauges interest
    - if they are not really interested, they will leave after this bit
  - 3-5 minutes full walkthrough
    - if someone stuck around past your 30 second summary or asked some follow up questions, walk them through your poster in more detail.
    - DON'T READ IT TO THEM!
  - A bottle of water is typically useful.

Again

- Poster Presentation: Best Project Prize!
  - Dec 6th 1:30-3:30pm
  - Goodwin Hall Atrium
  - Print poster (or bunch of slides)
    - Fedex, Library, ECE support, CS support
  - Format:
    - Portrait, 2 feet (width) x 36 inches (height)
- Submit poster as PDF by Dec 6th 1:30pm
  - Makes up the final portion of your project grade
- If you are worried about your project, talk to me soon.

Homework & Grading

- HW3 & HW4 should be graded this week
- Will release solutions this week as well

Final Exam

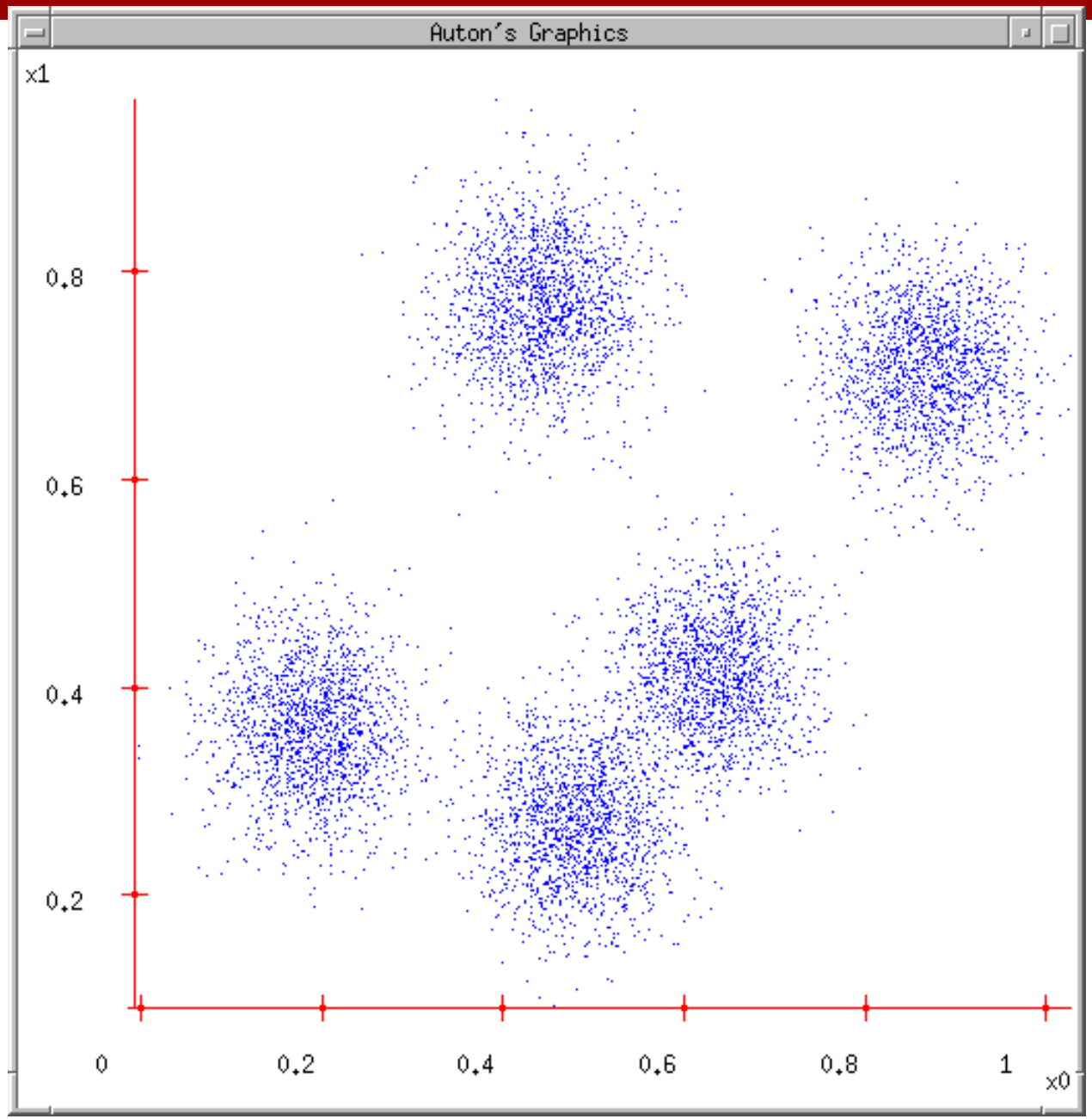
- Dec 14th in class (DURH 261); 2:05 - 4:05 pm
- Content:
  - Almost completely about material since the midterm
    - SVM, Neural Networks, Descision Trees, Ensemble Techniques, K-means, EM **(today),** Factor Analysis **(Thrusday)**
  - True/False (explain your choice like last time)
  - Multiple Choice
  - Some 'Prove this'
  - Some 'What would happen with algorithm A on this dataset"

One last thing

- SPOT surveys

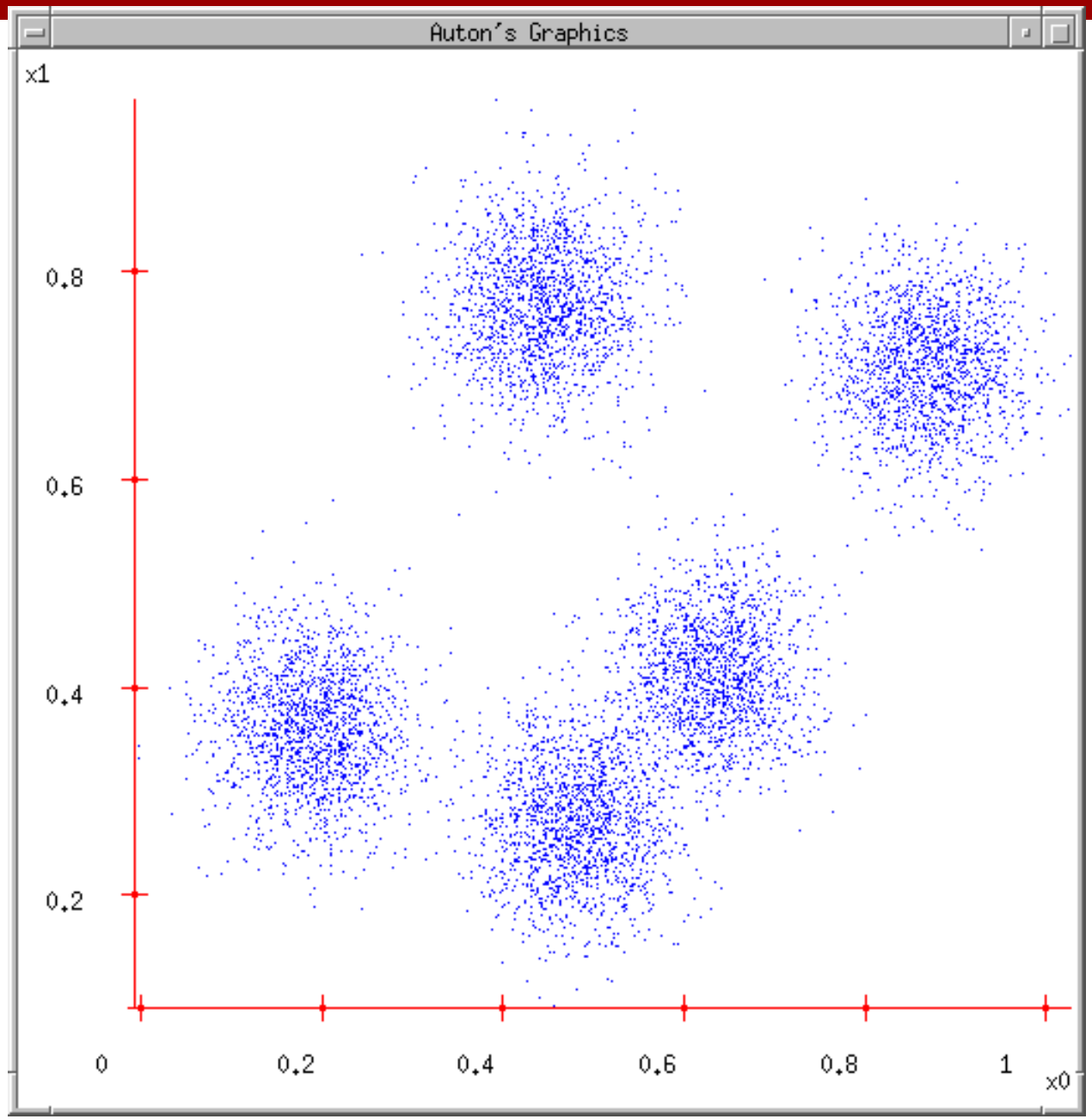
Recap of Last Time

Some Data



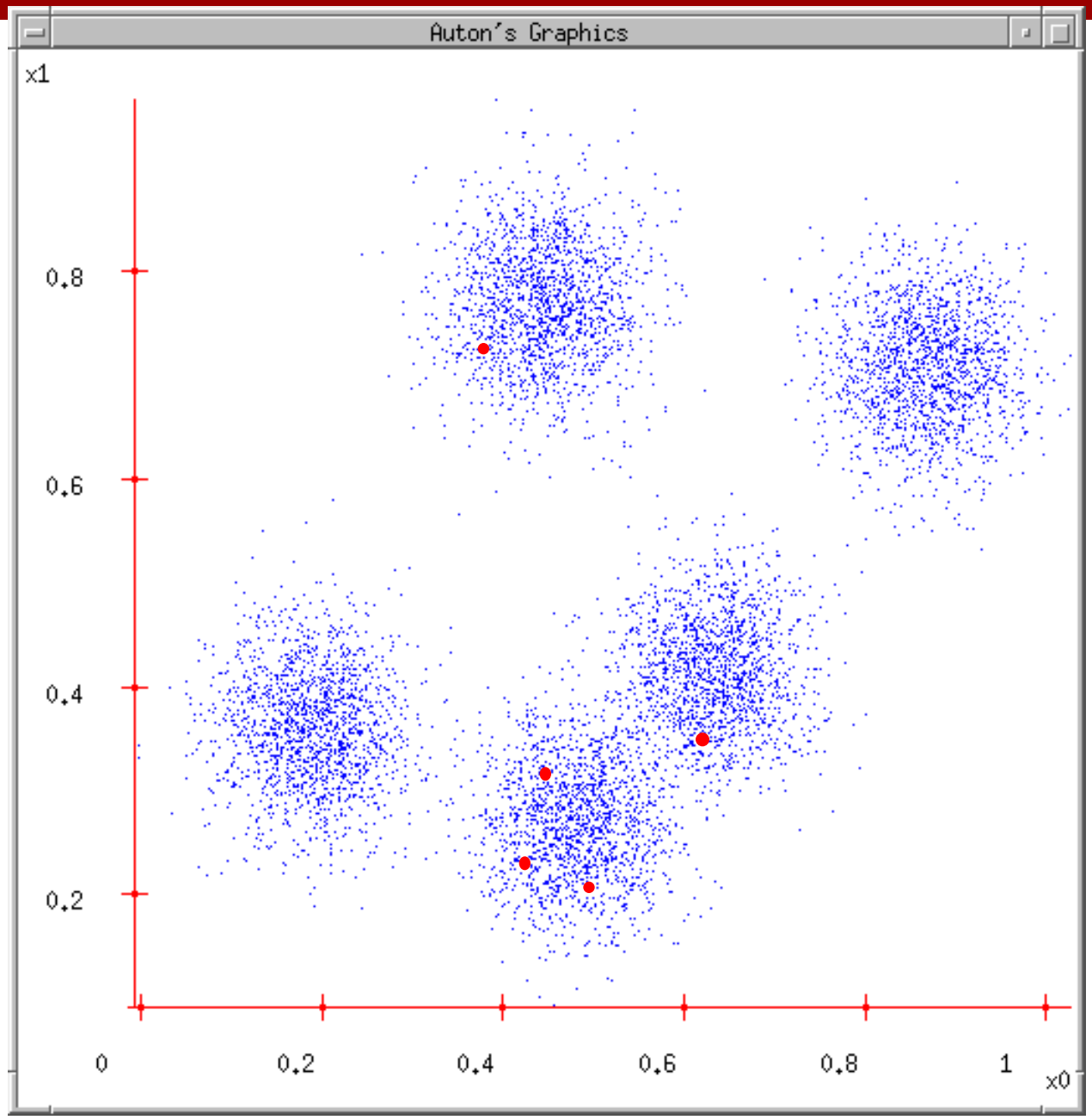
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*



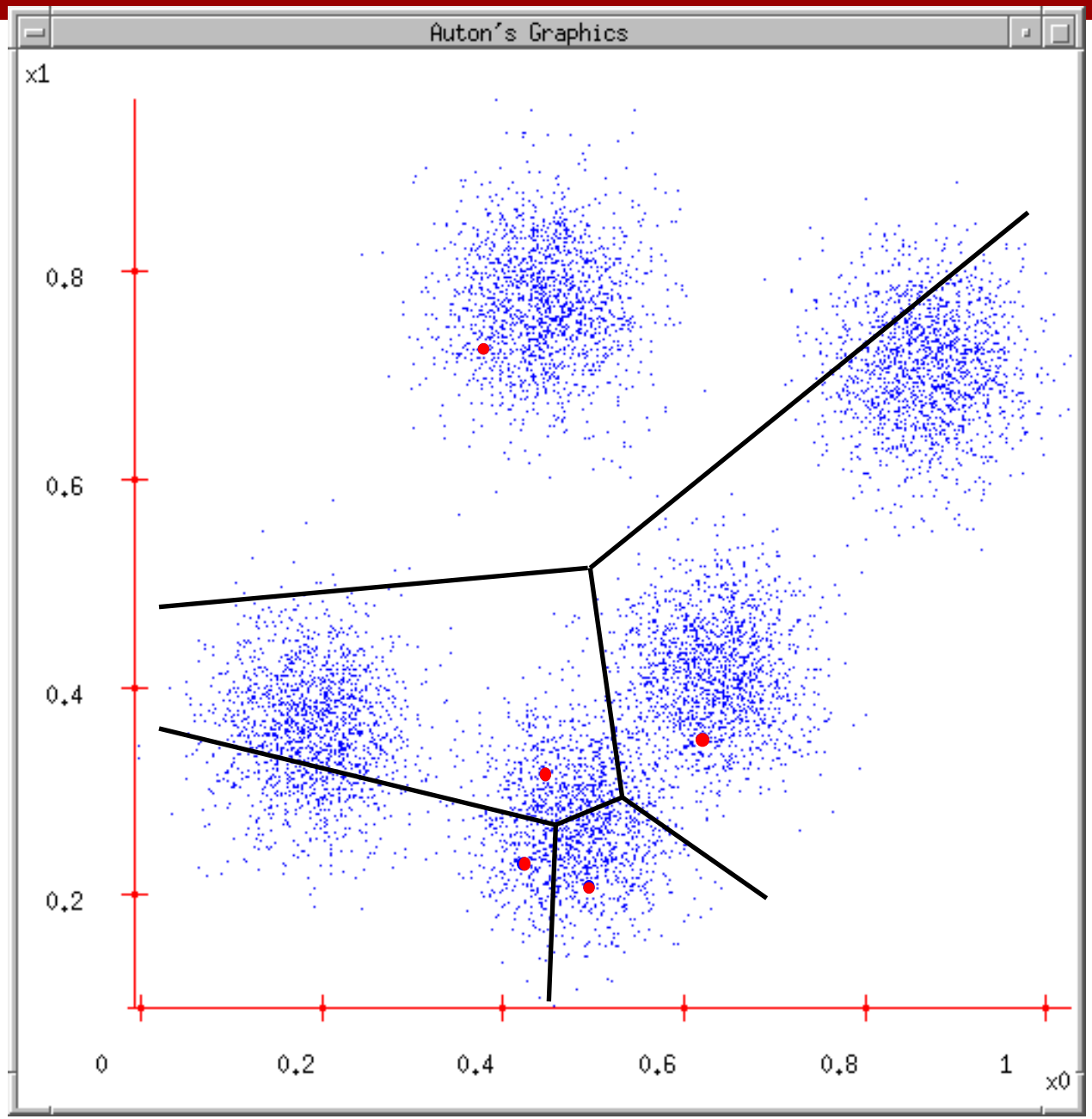
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations



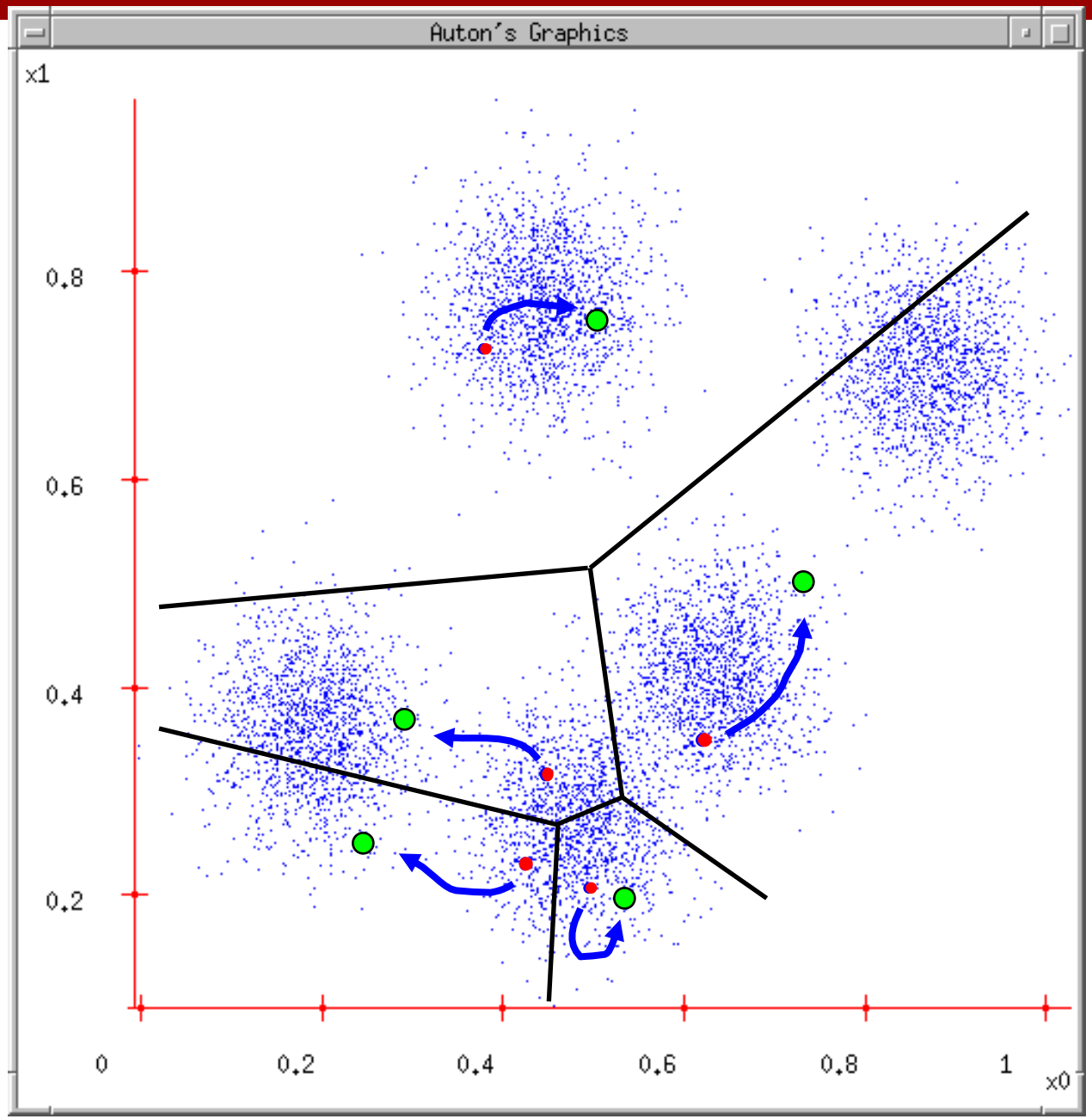
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



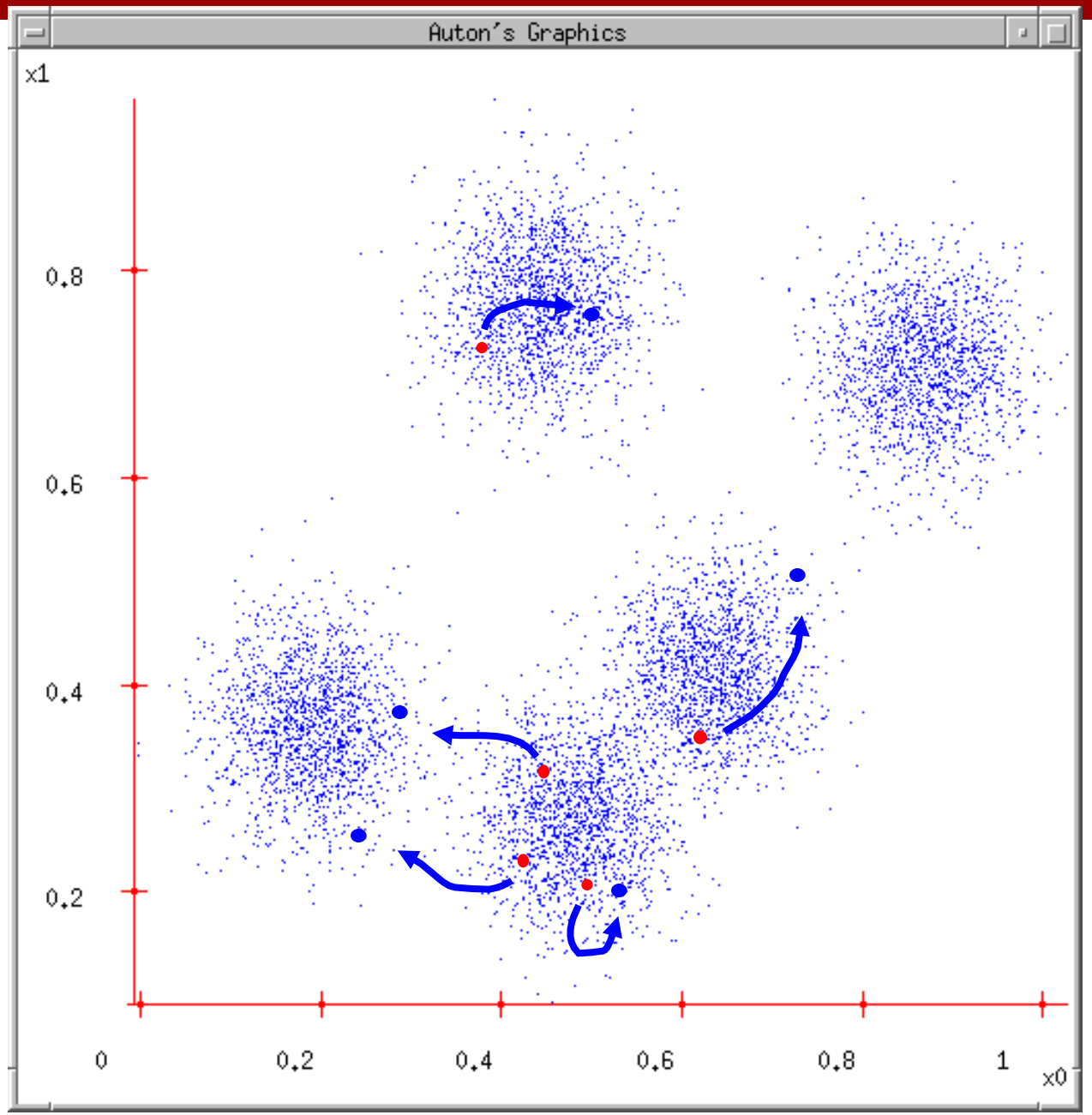
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns



K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns…
5. …and jumps there
6. …Repeat until terminated!



K-means

- Randomly initialize *k* centers
  - $^{(0)} = {}_1^{(0)}, \ldots, {}_k^{(0)}$
- **Assign**:
  - Assign each point i {1,…n} to nearest center:
  - $C(i) \longleftarrow \underset{j}{\operatorname{argmin}} ||\mathbf{x}_i - \boldsymbol{\mu}_j||^2$
- **Recenter**:
  - $_j$ becomes centroid of its points

K-means as Co-ordinate Descent

- Optimize objective function:

$$\min_{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k} \min_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N} F(\boldsymbol{\mu}, \boldsymbol{a}) = \min_{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k} \min_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N} \sum_{i=1}^{N}\sum_{j=1}^{k} a_{ij}||\mathbf{x}_i - \boldsymbol{\mu}_j||^2$$

- Fix , optimize a (or C)

K-means as Co-ordinate Descent

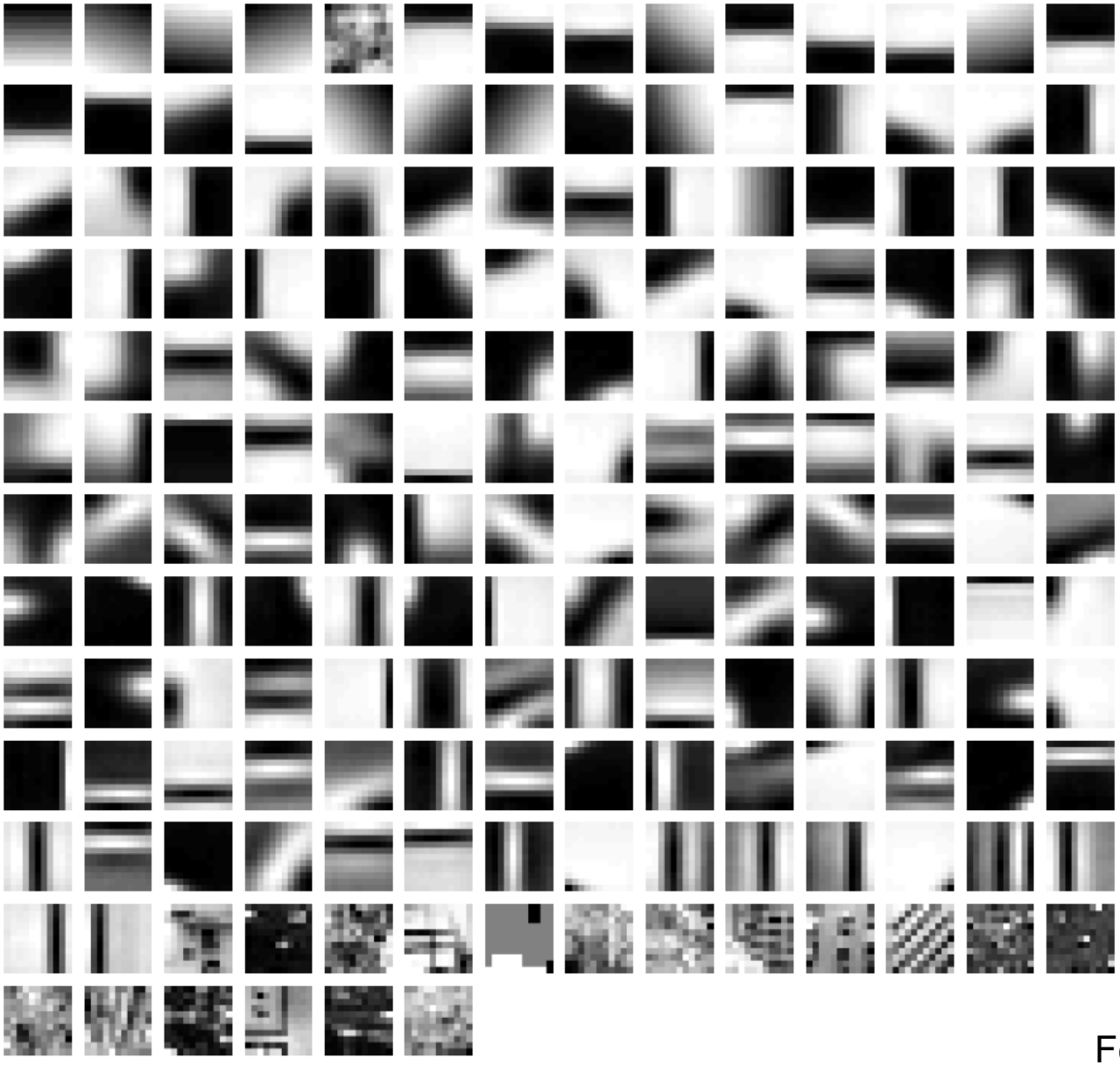
- Optimize objective function:

$$\min_{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k} \min_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N} F(\boldsymbol{\mu}, \boldsymbol{a}) = \min_{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k} \min_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N} \sum_{i=1}^{N}\sum_{j=1}^{k} a_{ij}||\mathbf{x}_i - \boldsymbol{\mu}_j||^2$$

- Fix a (or C), optimize

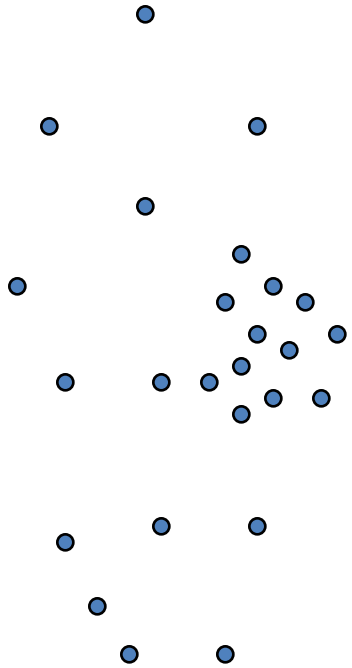
Object → Bag of 'words'

Clustered Image Patches



Fei-Fei et al. 2005

(One) bad case for k-means



- Clusters may overlap
- Some clusters may be "wider" than others

- GMM to the rescue!

GMM

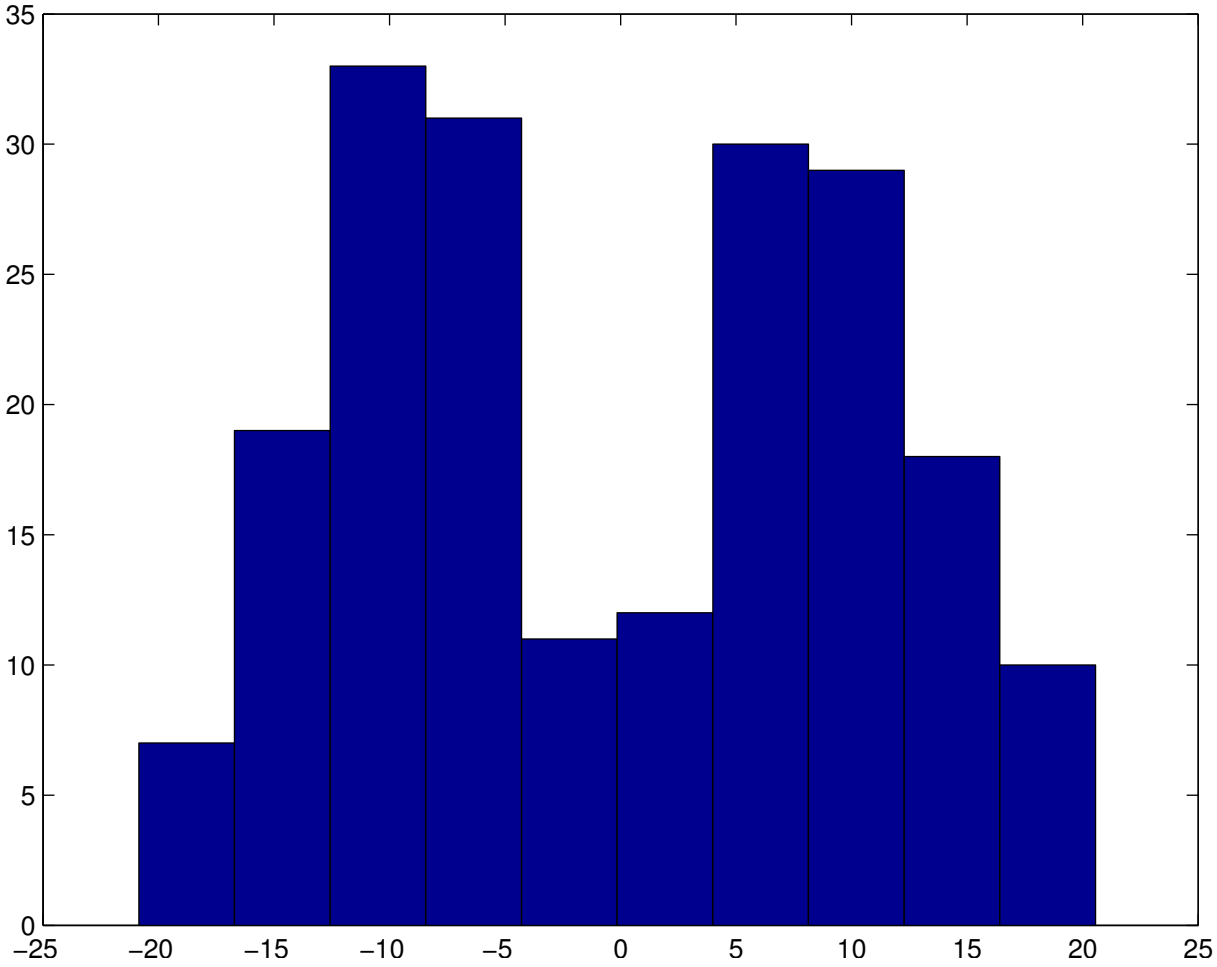
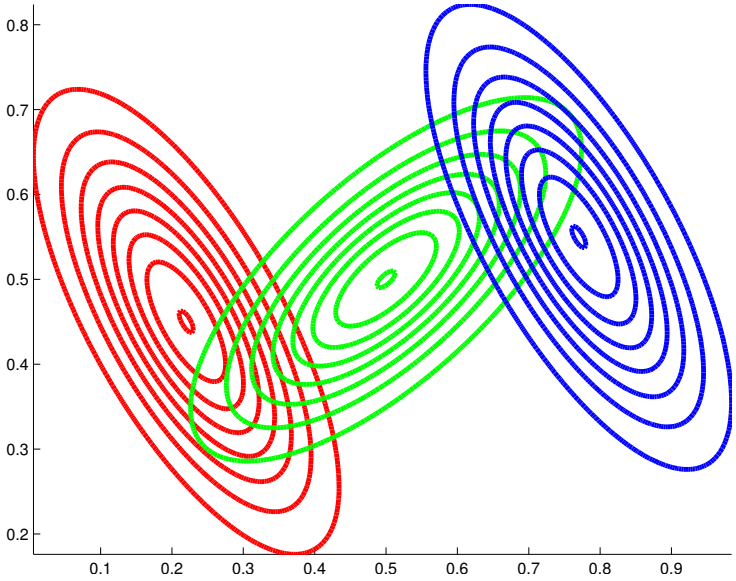


Figure Credit: Kevin Murphy

GMM



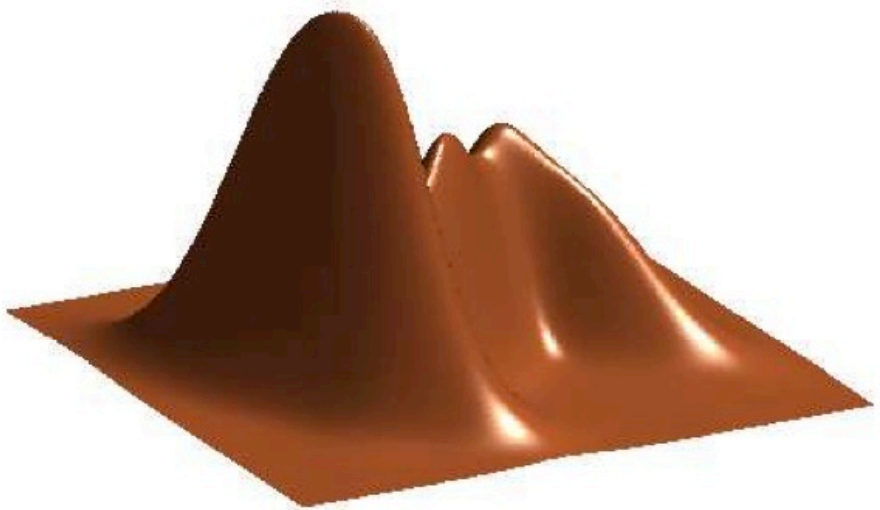


Figure Credit: Kevin Murphy

K-means vs GMM

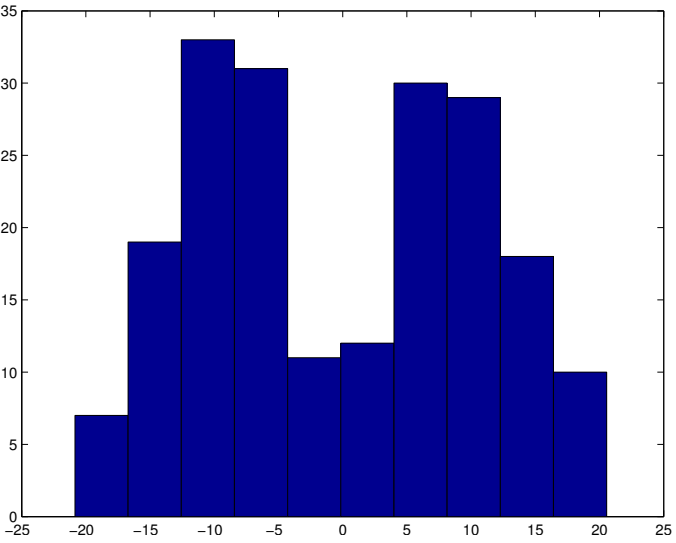
- K-Means
  - http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
- GMM
  - http://www.socr.ucla.edu/applets.dir/mixtureem.html

Hidden Data Causes Problems #1

- Fully Observed (Log) Likelihood factorizes
- Marginal (Log) Likelihood doesn't factorize
- All parameters coupled!

Hidden Data Causes Problems #2

- Identifiability



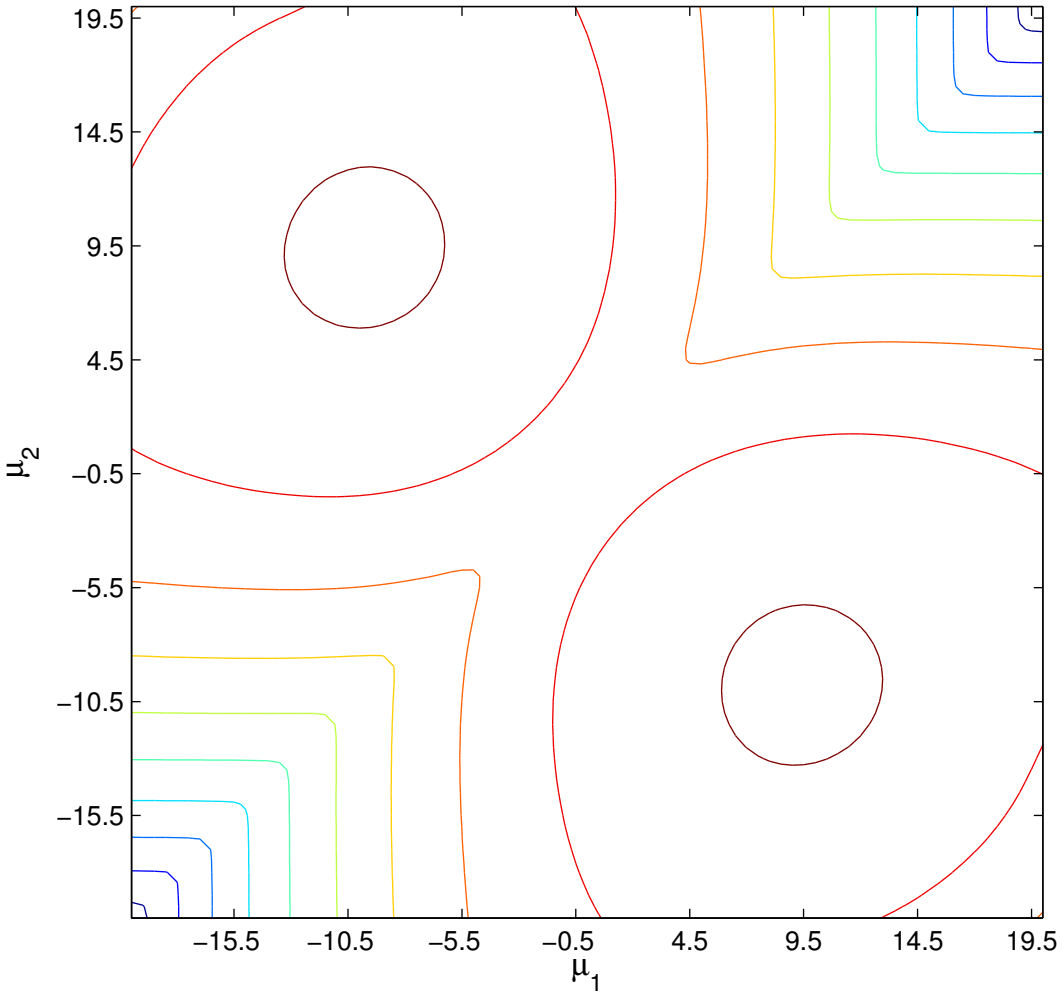
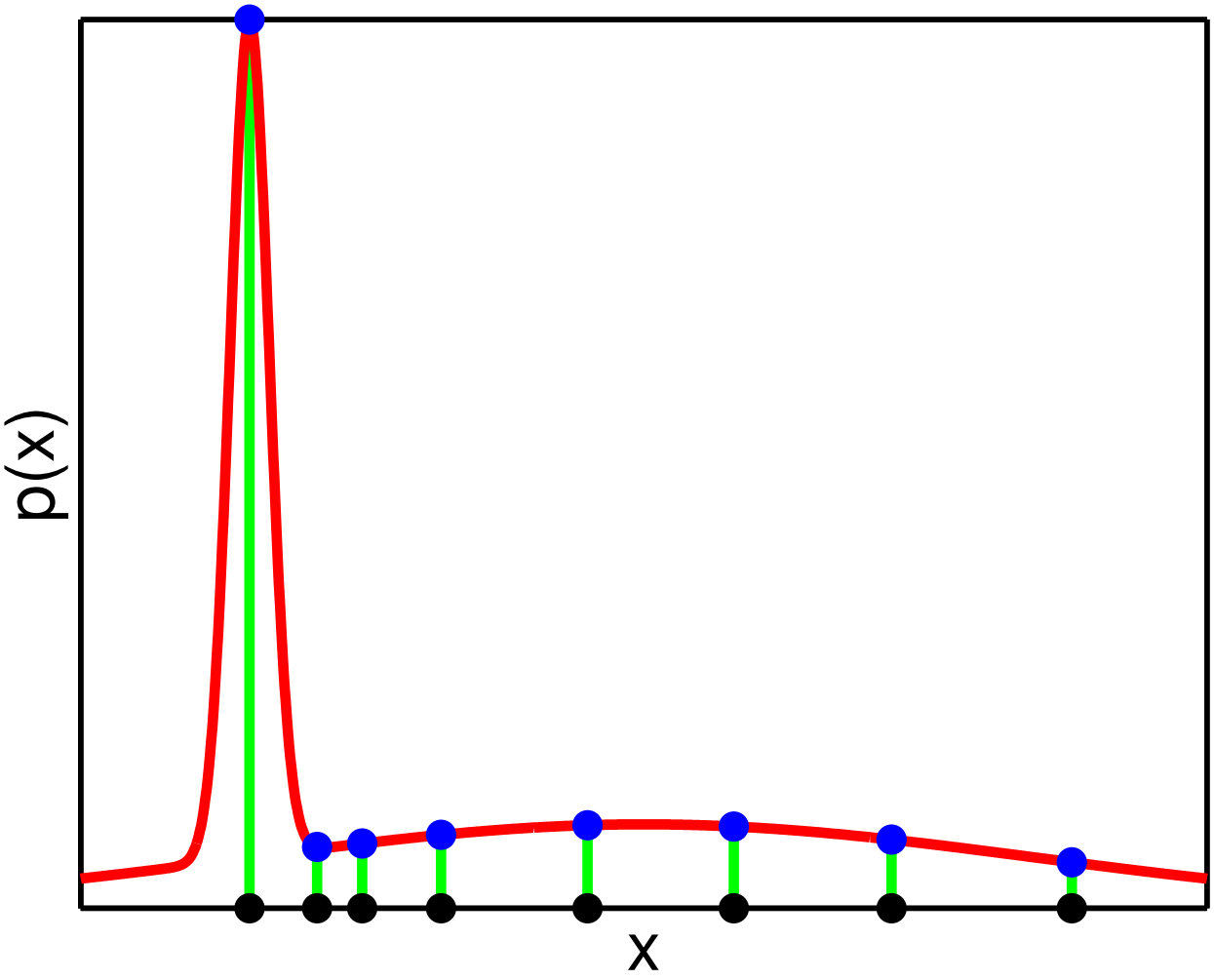


Figure Credit: Kevin Murphy

Hidden Data Causes Problems #3

- Likelihood has singularities if one Gaussian "collapses"



Special case: spherical Gaussians and hard assignments

- If P(X|Z=k) is spherical, with same for all classes:

$$P(\mathbf{x}_i \mid z = j) \propto \exp\left[-\frac{1}{2\sigma^2}\left\|\mathbf{x}_i - \mu_j\right\|^2\right]$$

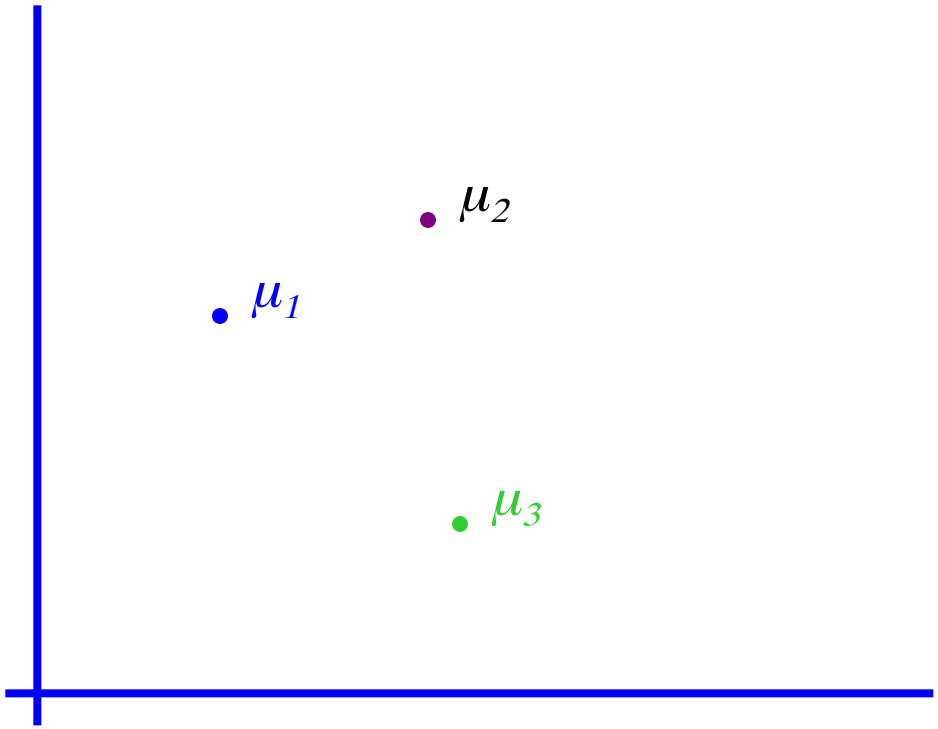
- If each x_i belongs to one class C(i) (hard assignment), marginal likelihood:

$$\prod_{i=1}^{N}\sum_{j=1}^{k} P(\mathbf{x}_i, y = j) \propto \prod_{i=1}^{N}\exp\left[-\frac{1}{2\sigma^2}\left\|\mathbf{x}_i - \mu_{C(i)}\right\|^2\right]$$

- M(M)LE same as K-means!!!

The K-means GMM assumption

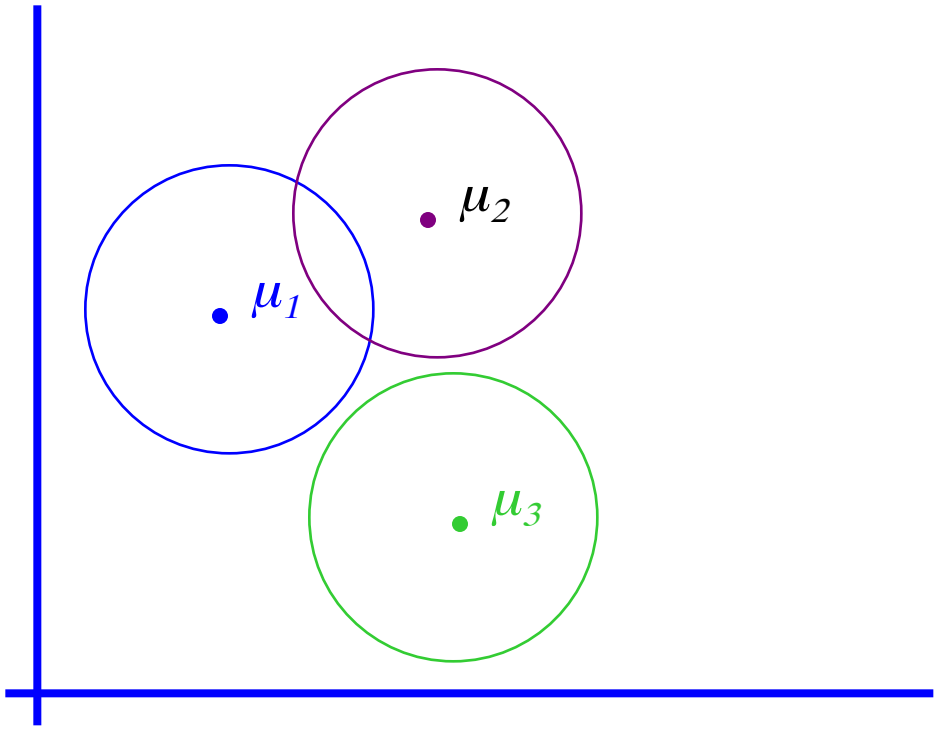
- There are k components
- Component *i* has an associated mean vector μ_i



The K-means GMM assumption

- There are k components
- Component *i* has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean m_i and covariance matrix $\sigma^2 I$

Each data point is generated according to the following recipe:

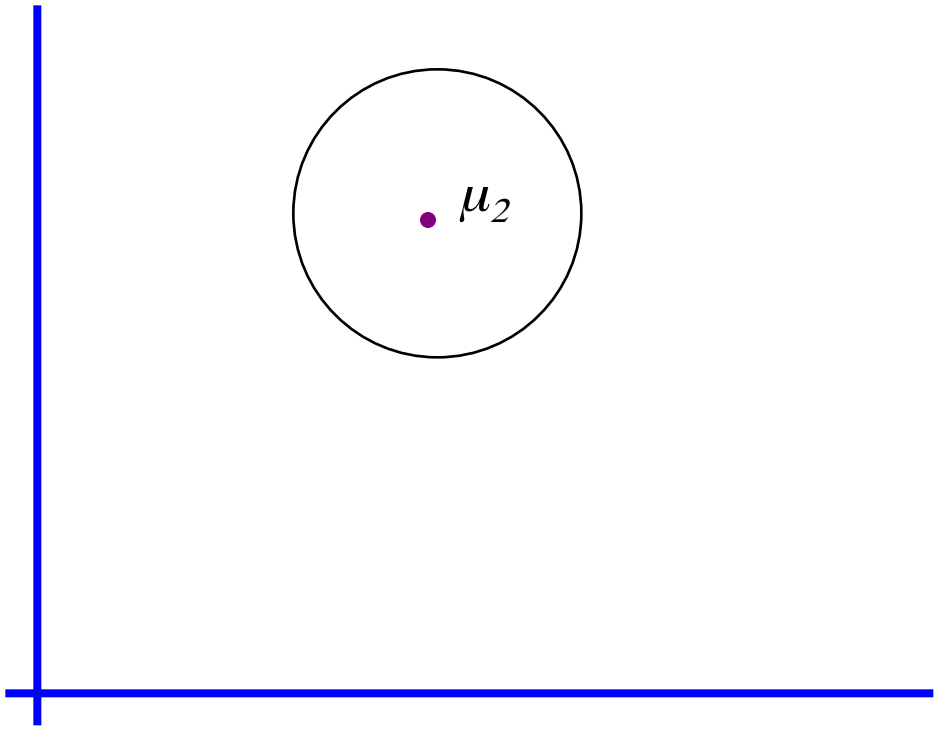


The K-means GMM assumption

- There are k components
- Component *i* has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean m_i and covariance matrix $\sigma^2 \boldsymbol{I}$

Each data point is generated according to the following recipe:

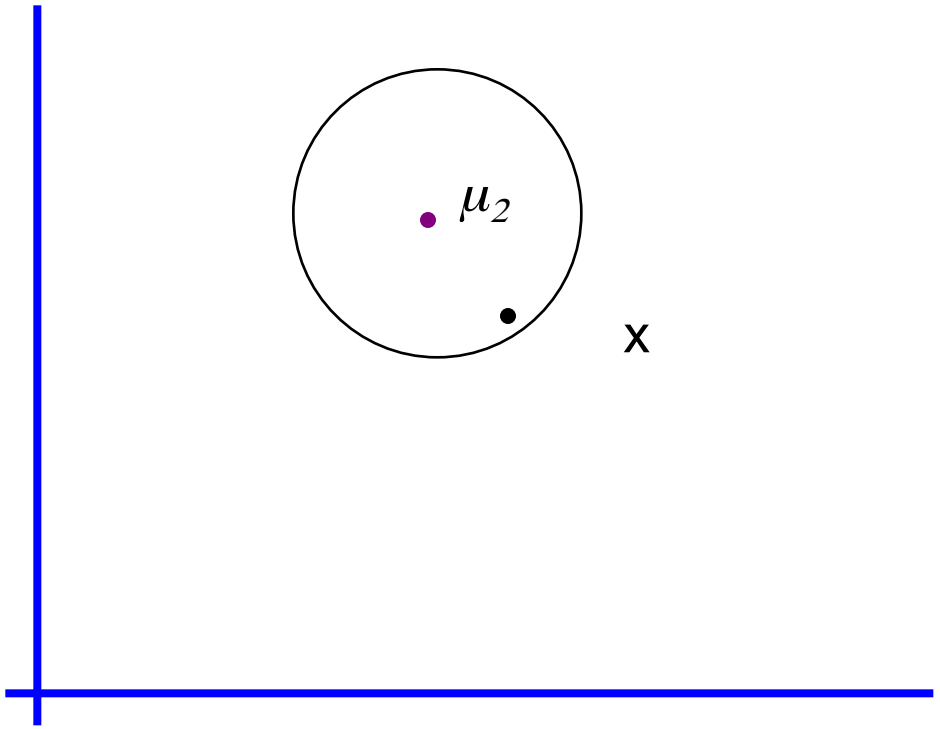
1. Pick a component at random: Choose component i with probability *P(y=i)*



The K-means GMM assumption

- There are k components
- Component *i* has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean m_i and covariance matrix $\sigma^2 \boldsymbol{I}$

Each data point is generated according to the following recipe:

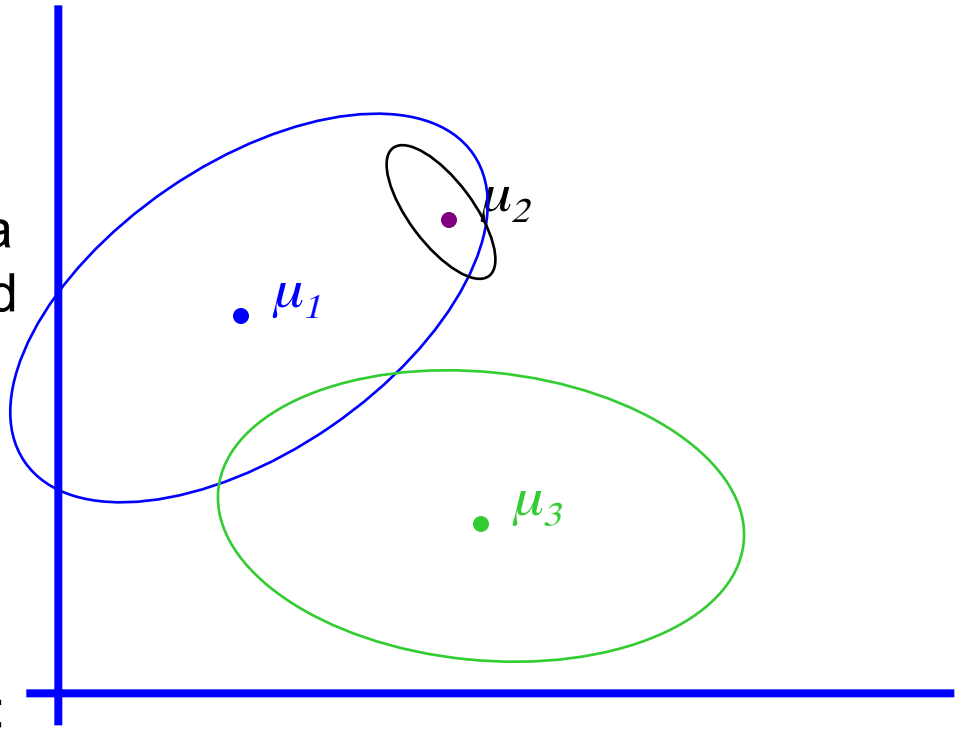


1. Pick a component at random: Choose component i with probability *P(y=i)*
2. Datapoint ~ $\mathrm{N}(\mu_i, \sigma^2 \boldsymbol{I})$

The General GMM assumption

- There are k components
- Component *i* has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean m_i and covariance matrix Σ_i

Each data point is generated according to the following recipe:



1. Pick a component at random: Choose component i with probability *P(y=i)*
2. Datapoint ~ $\mathrm{N}(\mu_i, \Sigma_i)$

K-means vs GMM

- K-Means
  - http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
- GMM
  - http://www.socr.ucla.edu/applets.dir/mixtureem.html

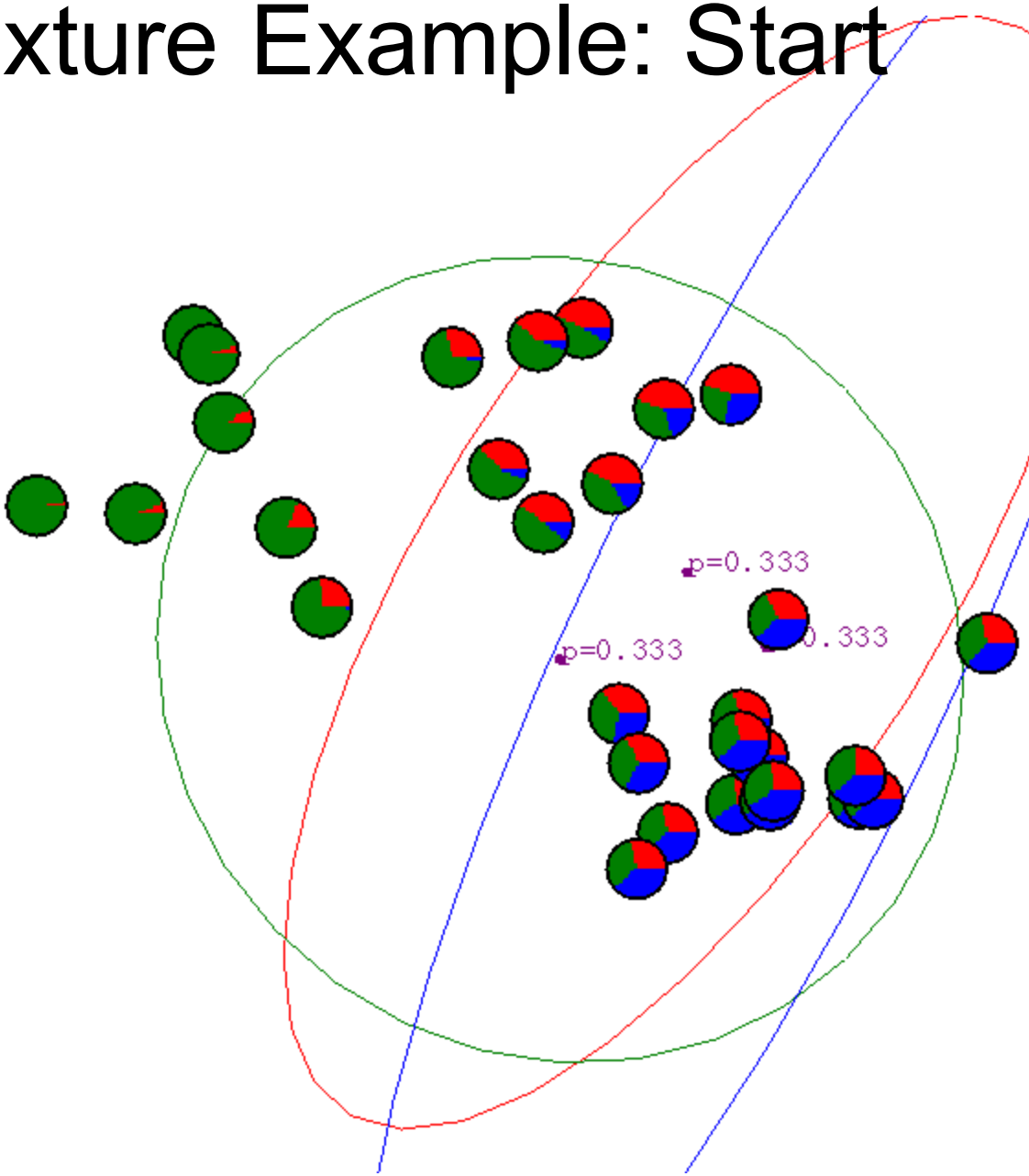
EM

- Expectation Maximization [Dempster '77]
- Often looks like "soft" K-means
- Extremely general
- Extremely useful algorithm
  - Essentially THE goto algorithm for unsupervised learning
- Plan
  - EM for learning GMM parameters
  - EM for general unsupervised learning problems

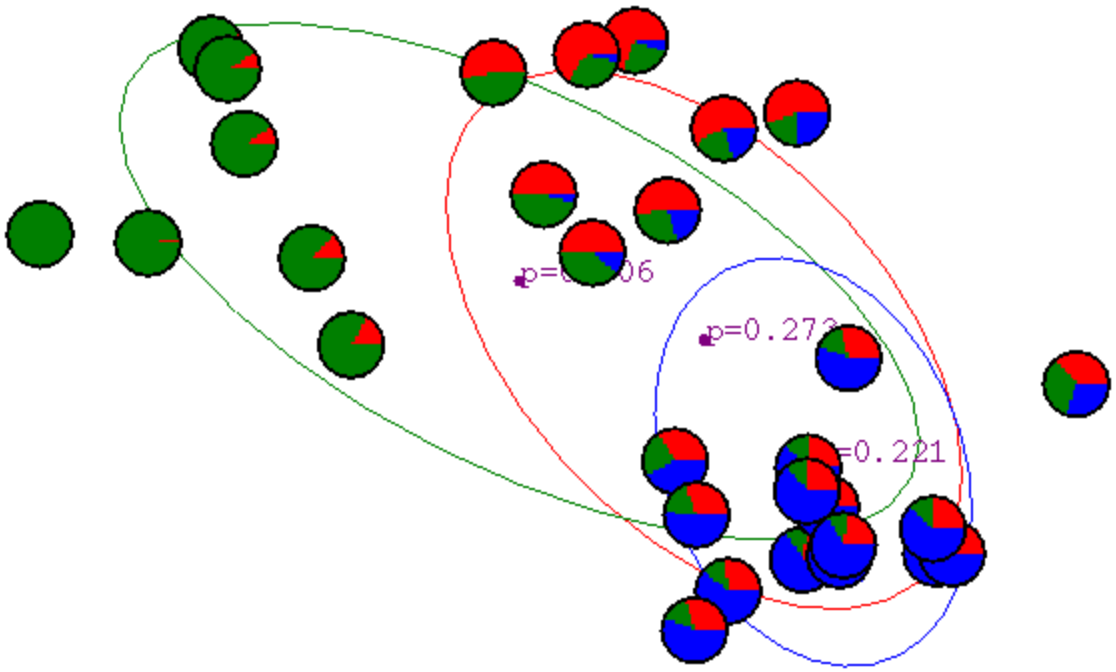
EM for Learning GMMs

- Simple Update Rules
  - E-Step: estimate $P(z_i = j \mid x_i)$
  - M-Step: maximize full likelihood weighted by posterior

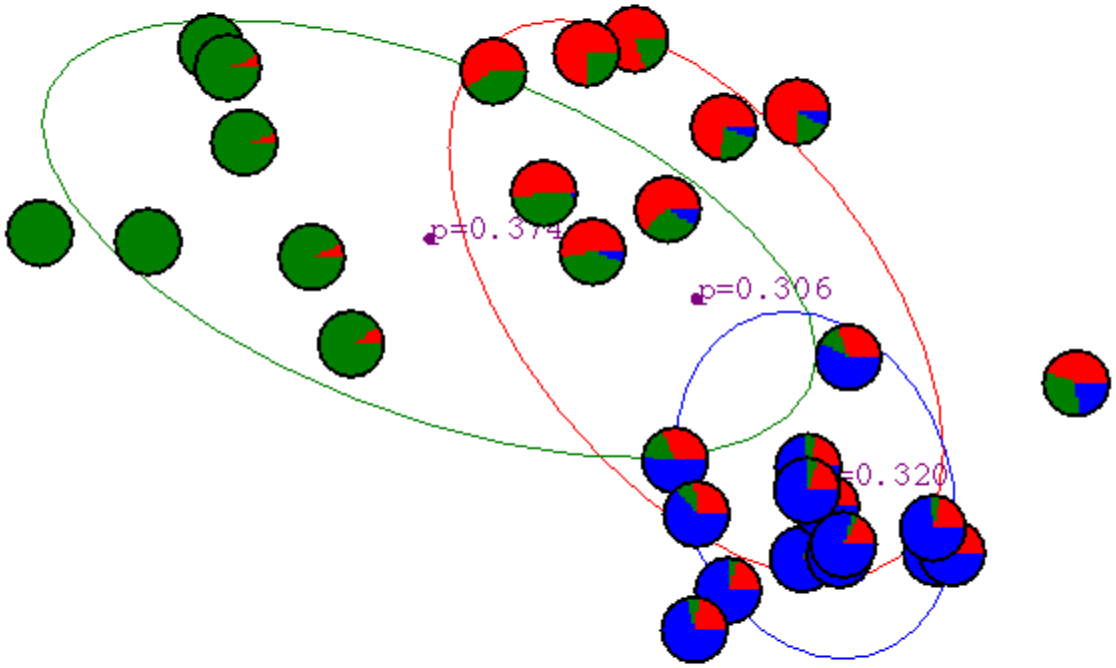
Gaussian Mixture Example: Start



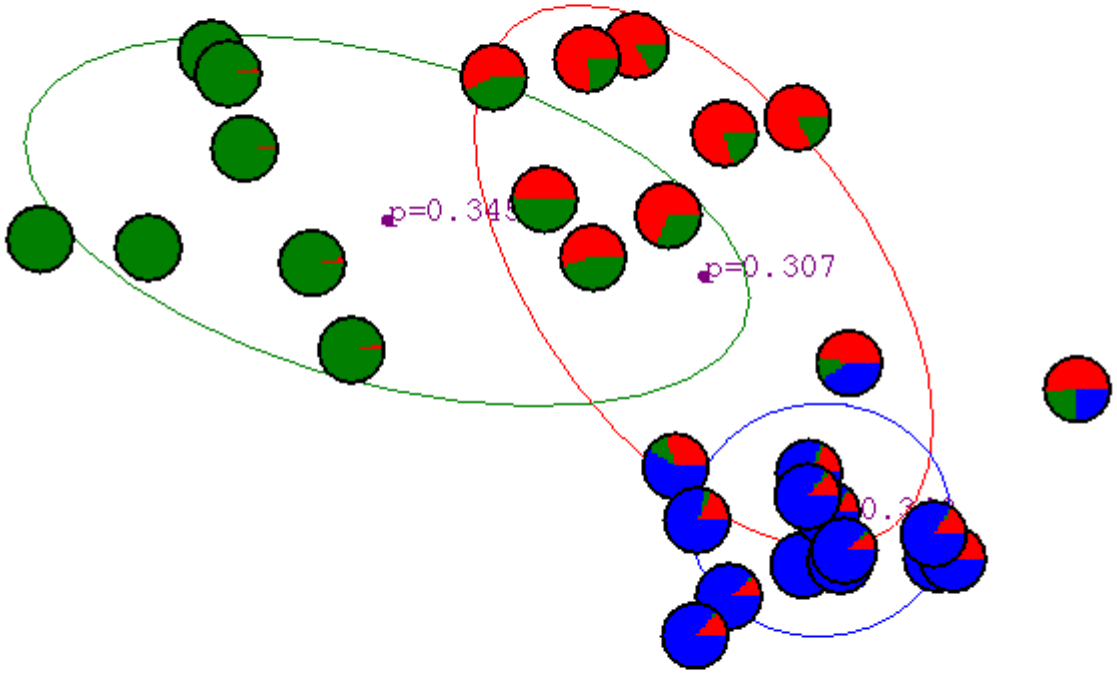
After 1st iteration



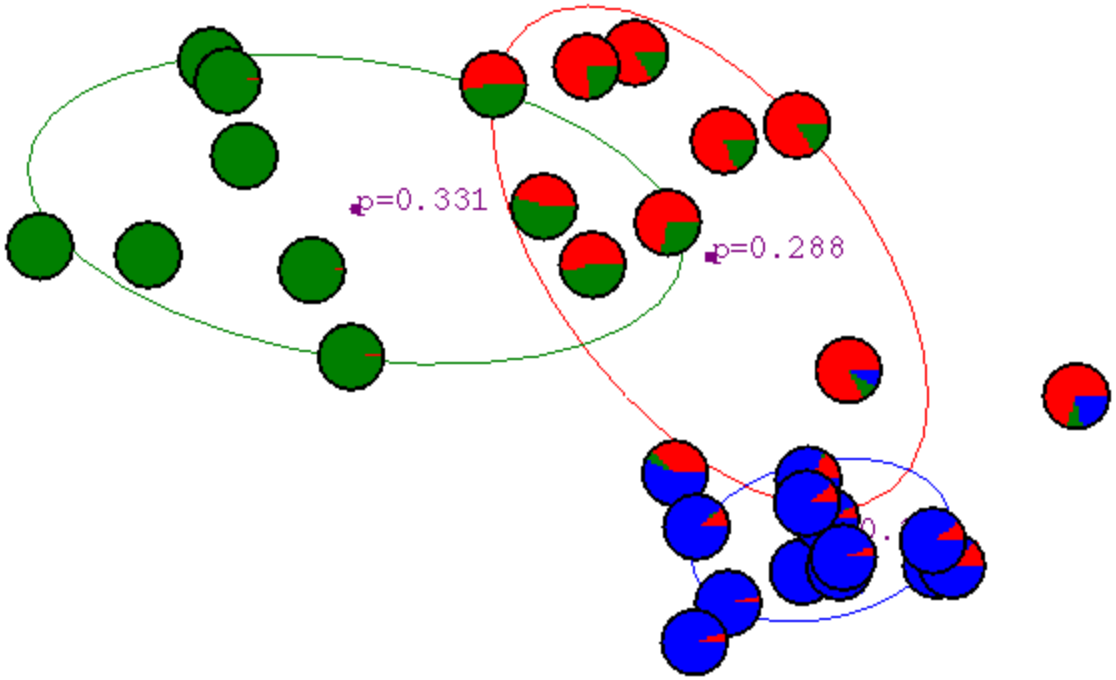
After 2nd iteration



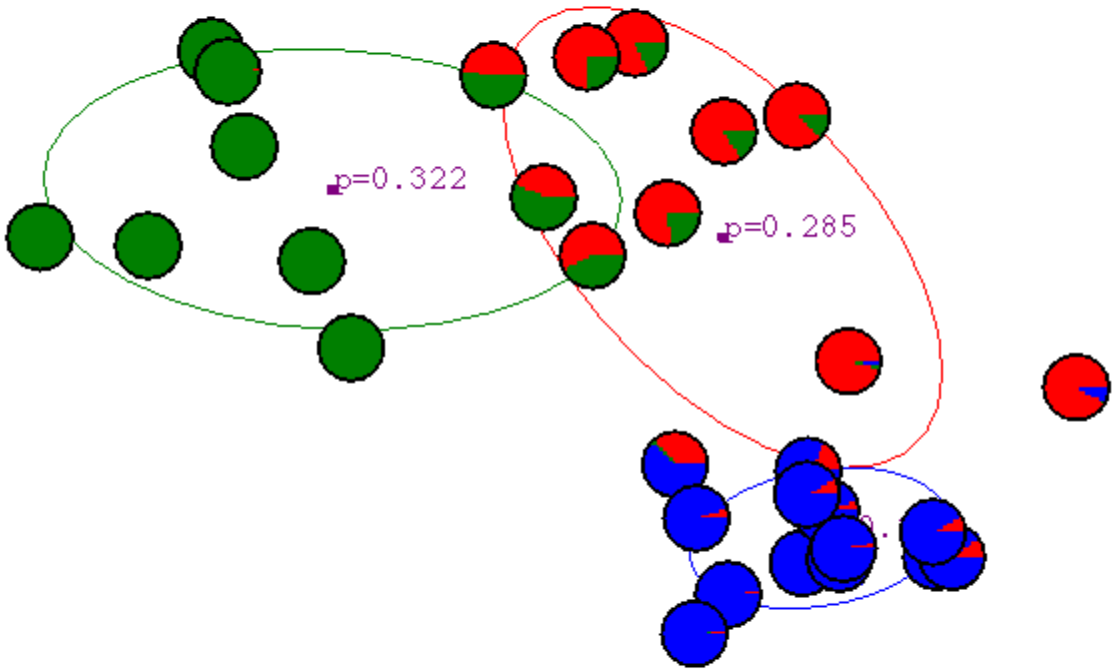
After 3rd iteration



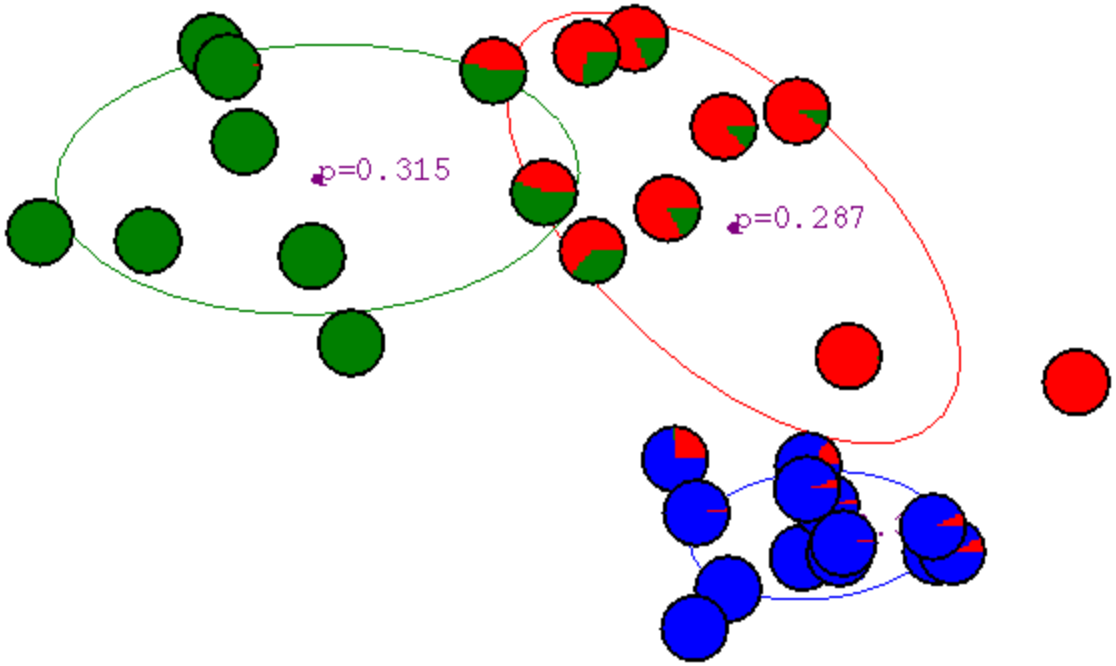
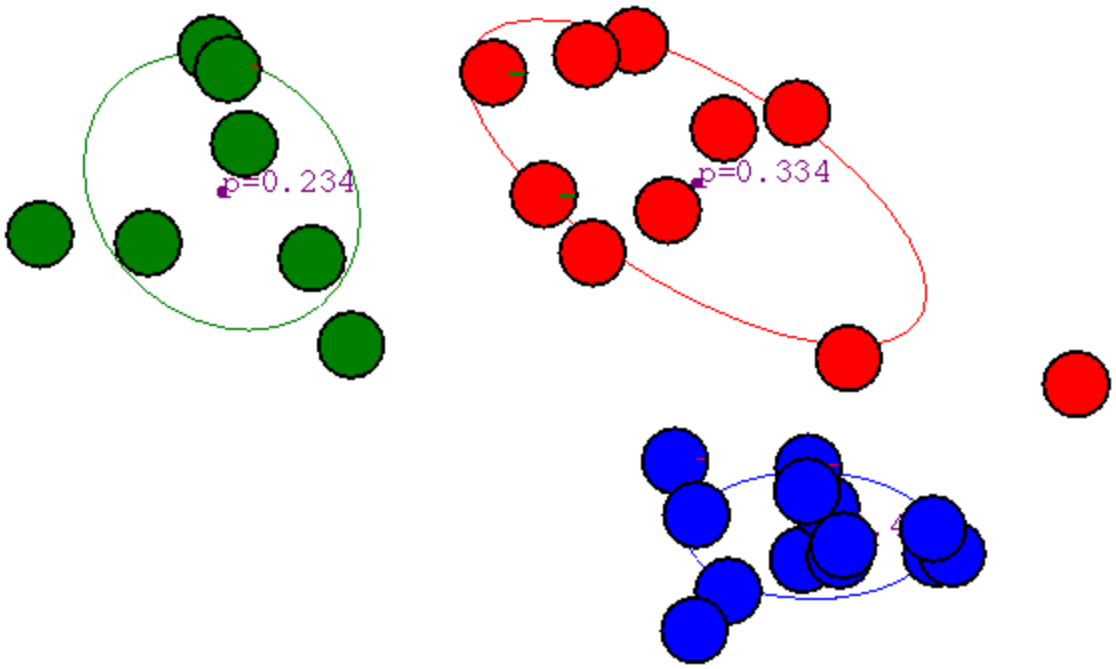
After 4th iteration



After 5th iteration



After 6th iteration



After 20th iteration