



Today's material available at:

https://github.com/julienelachance/ai4all_clustering

Clustering

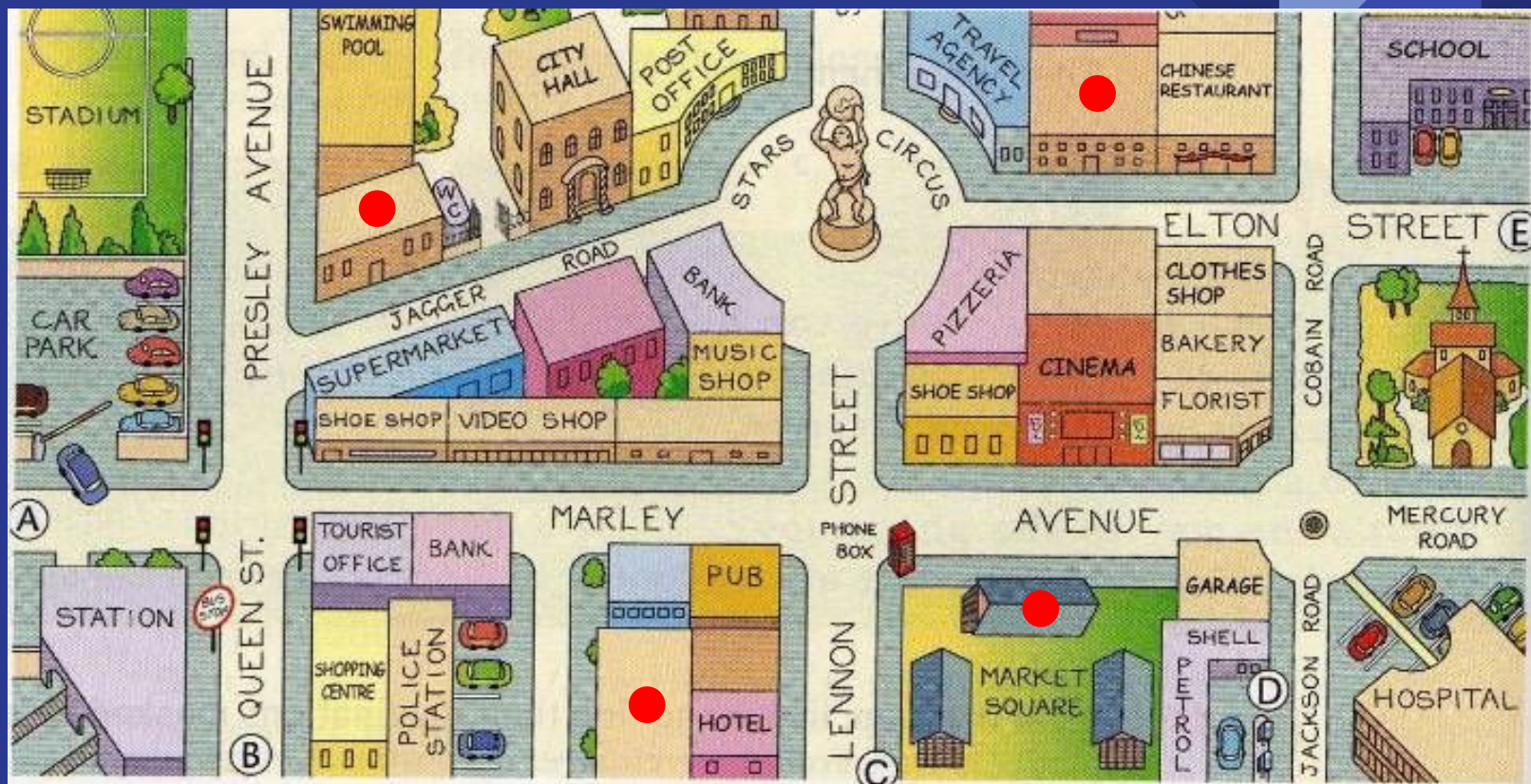
Princeton AI4ALL

By Julie LaChance

Icebreaker

- Everyone receives a slip of paper
- Your task:
Using the information about your person, form dog-walking groups
- Be prepared to justify your answers!
 - What features are important and why?





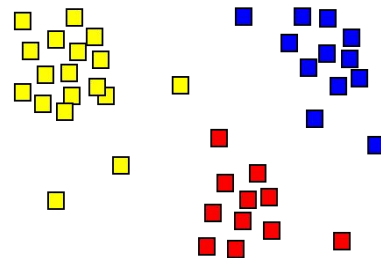
Overview:

- What is clustering?
- Applications of clustering
- Types of clustering
- k-means algorithm
- Recommendation system example



What is clustering?

What is clustering?



- From Wikipedia:

“**Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).”
- Thinking back to our first lecture on machine learning:

This is an important type of ***unsupervised learning***. Why?



Applications of clustering

Periodic Table of Elements

Periodic Table of the Elements																																																																																																																																																																																																																																																																																																																																																																																																																									
1 IA 1A																		18 VIIIA 8A																																																																																																																																																																																																																																																																																																																																																																																																							
1 H Hydrogen 1.008	2 He Helium 4.003																																																																																																																																																																																																																																																																																																																																																																																																																								
3 Li Lithium 6.941	4 Be Beryllium 9.012																	5 B Boron 10.811	6 C Carbon 12.011	7 N Nitrogen 14.007	8 O Oxygen 15.999	9 F Fluorine 18.998	10 Ne Neon 20.180																																																																																																																																																																																																																																																																																																																																																																																																		
11 Na Sodium 22.990	12 Mg Magnesium 24.305	13 Al Aluminum 26.982	14 Si Silicon 28.086	15 P Phosphorus 30.974	16 S Sulfur 32.065	17 Cl Chlorine 35.453	18 Ar Argon 39.948												19 K Potassium 39.098	20 Ca Calcium 40.078	21 Sc Scandium 44.956	22 Ti Titanium 47.88	23 V Vanadium 50.942	24 Cr Chromium 51.996	25 Mn Manganese 54.938	26 Fe Iron 55.845	27 Co Cobalt 58.933	28 Ni Nickel 58.693	29 Cu Copper 63.546	30 Zn Zinc 65.39	31 Ga Gallium 69.723	32 Ge Germanium 72.61	33 As Arsenic 74.922	34 Se Selenium 78.96	35 Br Bromine 79.904	36 Kr Krypton 83.80																																																																																																																																																																																																																																																																																																																																																																																					
37 Rb Rubidium 85.468	38 Sr Strontium 87.62	39 Y Yttrium 88.906	40 Zr Zirconium 91.224	41 Nb Niobium 92.906	42 Mo Molybdenum 95.94	43 Tc Technetium 98.907	44 Ru Ruthenium 101.07	45 Rh Rhodium 102.905	46 Pd Palladium 106.42	47 Ag Silver 107.868	48 Cd Cadmium 112.411	49 In Indium 114.818	50 Sn Tin 118.71	51 Sb Antimony 121.760	52 Te Tellurium 127.6	53 I Iodine 126.904	54 Xe Xenon 131.29														55 Cs Cesium 132.905	56 Ba Barium 137.327	57-71 Lanthanide Series	72 Hf Hafnium 178.49	73 Ta Tantalum 180.948	74 W Tungsten 183.85	75 Re Rhenium 186.207	76 Os Osmium 190.23	77 Ir Iridium 192.22	78 Pt Platinum 195.08	79 Au Gold 196.967	80 Hg Mercury 200.59	81 Tl Thallium 204.383	82 Pb Lead 207.2	83 Bi Bismuth 208.980	84 Po Polonium [209]	85 At Astatine [210]	86 Rn Radon [222]																																																																																																																																																																																																																																																																																																																																																																									
87 Fr Francium 223.021	88 Ra Radium 226.025	89-103 Actinide Series	104 Rf Rutherfordium [261]	105 Db Dubnium [262]	106 Sg Seaborgium [266]	107 Bh Bohrium [264]	108 Hs Hassium [269]	109 Mt Meitnerium [268]	110 Ds Darmstadtium [271]	111 Rg Roentgenium [272]	112 Cn Copernicium [277]	113 Nh Nihonium [284]	114 Fl Flerovium [289]	115 Mc Moscovium [288]	116 Lv Livermorium [293]	117 Ts Tennessine [294]	118 Og Oganesson [294]														119 Uue Ununennium [289]	120 Uuo Unbinilium [289]	121 Uut Ununtrium [289]	122 Uuq Unquadium [289]	123 Uup Unupennium [289]	124 Uuq Unquadium [289]	125 Uus Unseptennium [289]	126 Uuo Unbinilium [289]	127 Uuh Unheptennium [289]	128 Uuo Unbinilium [289]	129 Uuh Unheptennium [289]	130 Uuo Unbinilium [289]	131 Uuh Unheptennium [289]	132 Uuo Unbinilium [289]	133 Uuh Unheptennium [289]	134 Uuo Unbinilium [289]	135 Uuh Unheptennium [289]	136 Uuo Unbinilium [289]	137 Uuh Unheptennium [289]	138 Uuo Unbinilium [289]	139 Uuh Unheptennium [289]	140 Uuo Unbinilium [289]	141 Uuh Unheptennium [289]	142 Uuo Unbinilium [289]	143 Uuh Unheptennium [289]	144 Uuo Unbinilium [289]	145 Uuh Unheptennium [289]	146 Uuo Unbinilium [289]	147 Uuh Unheptennium [289]	148 Uuo Unbinilium [289]	149 Uuh Unheptennium [289]	150 Uuo Unbinilium [289]	151 Uuh Unheptennium [289]	152 Uuo Unbinilium [289]	153 Uuh Unheptennium [289]	154 Uuo Unbinilium [289]	155 Uuh Unheptennium [289]	156 Uuo Unbinilium [289]	157 Uuh Unheptennium [289]	158 Uuo Unbinilium [289]	159 Uuh Unheptennium [289]	160 Uuo Unbinilium [289]	161 Uuh Unheptennium [289]	162 Uuo Unbinilium [289]	163 Uuh Unheptennium [289]	164 Uuo Unbinilium [289]	165 Uuh Unheptennium [289]	166 Uuo Unbinilium [289]	167 Uuh Unheptennium [289]	168 Uuo Unbinilium [289]	169 Uuh Unheptennium [289]	170 Uuo Unbinilium [289]	171 Uuh Unheptennium [289]	172 Uuo Unbinilium [289]	173 Uuh Unheptennium [289]	174 Uuo Unbinilium [289]	175 Uuh Unheptennium [289]	176 Uuo Unbinilium [289]	177 Uuh Unheptennium [289]	178 Uuo Unbinilium [289]	179 Uuh Unheptennium [289]	180 Uuo Unbinilium [289]	181 Uuh Unheptennium [289]	182 Uuo Unbinilium [289]	183 Uuh Unheptennium [289]	184 Uuo Unbinilium [289]	185 Uuh Unheptennium [289]	186 Uuo Unbinilium [289]	187 Uuh Unheptennium [289]	188 Uuo Unbinilium [289]	189 Uuh Unheptennium [289]	190 Uuo Unbinilium [289]	191 Uuh Unheptennium [289]	192 Uuo Unbinilium [289]	193 Uuh Unheptennium [289]	194 Uuo Unbinilium [289]	195 Uuh Unheptennium [289]	196 Uuo Unbinilium [289]	197 Uuh Unheptennium [289]	198 Uuo Unbinilium [289]	199 Uuh Unheptennium [289]	200 Uuo Unbinilium [289]	201 Uuh Unheptennium [289]	202 Uuo Unbinilium [289]	203 Uuh Unheptennium [289]	204 Uuo Unbinilium [289]	205 Uuh Unheptennium [289]	206 Uuo Unbinilium [289]	207 Uuh Unheptennium [289]	208 Uuo Unbinilium [289]	209 Uuh Unheptennium [289]	210 Uuo Unbinilium [289]	211 Uuh Unheptennium [289]	212 Uuo Unbinilium [289]	213 Uuh Unheptennium [289]	214 Uuo Unbinilium [289]	215 Uuh Unheptennium [289]	216 Uuo Unbinilium [289]	217 Uuh Unheptennium [289]	218 Uuo Unbinilium [289]	219 Uuh Unheptennium [289]	220 Uuo Unbinilium [289]	221 Uuh Unheptennium [289]	222 Uuo Unbinilium [289]	223 Uuh Unheptennium [289]	224 Uuo Unbinilium [289]	225 Uuh Unheptennium [289]	226 Uuo Unbinilium [289]	227 Uuh Unheptennium [289]	228 Uuo Unbinilium [289]	229 Uuh Unheptennium [289]	230 Uuo Unbinilium [289]	231 Uuh Unheptennium [289]	232 Uuo Unbinilium [289]	233 Uuh Unheptennium [289]	234 Uuo Unbinilium [289]	235 Uuh Unheptennium [289]	236 Uuo Unbinilium [289]	237 Uuh Unheptennium [289]	238 Uuo Unbinilium [289]	239 Uuh Unheptennium [289]	240 Uuo Unbinilium [289]	241 Uuh Unheptennium [289]	242 Uuo Unbinilium [289]	243 Uuh Unheptennium [289]	244 Uuo Unbinilium [289]	245 Uuh Unheptennium [289]	246 Uuo Unbinilium [289]	247 Uuh Unheptennium [289]	248 Uuo Unbinilium [289]	249 Uuh Unheptennium [289]	250 Uuo Unbinilium [289]	251 Uuh Unheptennium [289]	252 Uuo Unbinilium [289]	253 Uuh Unheptennium [289]	254 Uuo Unbinilium [289]	255 Uuh Unheptennium [289]	256 Uuo Unbinilium [289]	257 Uuh Unheptennium [289]	258 Uuo Unbinilium [289]	259 Uuh Unheptennium [289]	260 Uuo Unbinilium [289]	261 Uuh Unheptennium [289]	262 Uuo Unbinilium [289]	263 Uuh Unheptennium [289]	264 Uuo Unbinilium [289]	265 Uuh Unheptennium [289]	266 Uuo Unbinilium [289]	267 Uuh Unheptennium [289]	268 Uuo Unbinilium [289]	269 Uuh Unheptennium [289]	270 Uuo Unbinilium [289]	271 Uuh Unheptennium [289]	272 Uuo Unbinilium [289]	273 Uuh Unheptennium [289]	274 Uuo Unbinilium [289]	275 Uuh Unheptennium [289]	276 Uuo Unbinilium [289]	277 Uuh Unheptennium [289]	278 Uuo Unbinilium [289]	279 Uuh Unheptennium [289]	280 Uuo Unbinilium [289]	281 Uuh Unheptennium [289]	282 Uuo Unbinilium [289]	283 Uuh Unheptennium [289]	284 Uuo Unbinilium [289]	285 Uuh Unheptennium [289]	286 Uuo Unbinilium [289]	287 Uuh Unheptennium [289]	288 Uuo Unbinilium [289]	289 Uuh Unheptennium [289]	290 Uuo Unbinilium [289]	291 Uuh Unheptennium [289]	292 Uuo Unbinilium [289]	293 Uuh Unheptennium [289]	294 Uuo Unbinilium [289]	295 Uuh Unheptennium [289]	296 Uuo Unbinilium [289]	297 Uuh Unheptennium [289]	298 Uuo Unbinilium [289]	299 Uuh Unheptennium [289]	300 Uuo Unbinilium [289]	301 Uuh Unheptennium [289]	302 Uuo Unbinilium [289]	303 Uuh Unheptennium [289]	304 Uuo Unbinilium [289]	305 Uuh Unheptennium [289]	306 Uuo Unbinilium [289]	307 Uuh Unheptennium [289]	308 Uuo Unbinilium [289]	309 Uuh Unheptennium [289]	310 Uuo Unbinilium [289]	311 Uuh Unheptennium [289]	312 Uuo Unbinilium [289]	313 Uuh Unheptennium [289]	314 Uuo Unbinilium [289]	315 Uuh Unheptennium [289]	316 Uuo Unbinilium [289]	317 Uuh Unheptennium [289]	318 Uuo Unbinilium [289]	319 Uuh Unheptennium [289]	320 Uuo Unbinilium [289]	321 Uuh Unheptennium [289]	322 Uuo Unbinilium [289]	323 Uuh Unheptennium [289]	324 Uuo Unbinilium [289]	325 Uuh Unheptennium [289]	326 Uuo Unbinilium [289]	327 Uuh Unheptennium [289]	328 Uuo Unbinilium [289]	329 Uuh Unheptennium [289]	330 Uuo Unbinilium [289]	331 Uuh Unheptennium [289]	332 Uuo Unbinilium [289]	333 Uuh Unheptennium [289]	334 Uuo Unbinilium [289]	335 Uuh Unheptennium [289]	336 Uuo Unbinilium [289]	337 Uuh Unheptennium [289]	338 Uuo Unbinilium [289]	339 Uuh Unheptennium [289]	340 Uuo Unbinilium [289]	341 Uuh Unheptennium [289]	342 Uuo Unbinilium [289]	343 Uuh Unheptennium [289]	344 Uuo Unbinilium [289]	345 Uuh Unheptennium [289]	346 Uuo Unbinilium [289]	347 Uuh Unheptennium [289]	348 Uuo Unbinilium [289]	349 Uuh Unheptennium [289]	350 Uuo Unbinilium [289]	351 Uuh Unheptennium [289]	352 Uuo Unbinilium [289]	353 Uuh Unheptennium [289]	354 Uuo Unbinilium [289]	355 Uuh Unheptennium [289]	356 Uuo Unbinilium [289]	357 Uuh Unheptennium [289]	358 Uuo Unbinilium [289]	359 Uuh Unheptennium [289]	360 Uuo Unbinilium [289]	361 Uuh Unheptennium [289]	362 Uuo Unbinilium [289]	363 Uuh Unheptennium [289]	364 Uuo Unbinilium [289]	365 Uuh Unheptennium [289]	366 Uuo Unbinilium [289]	367 Uuh Unheptennium [289]	368 Uuo Unbinilium [289]	369 Uuh Unheptennium [289]	370 Uuo Unbinilium [289]	371 Uuh Unheptennium [289]	372 Uuo Unbinilium [289]	373 Uuh Unheptennium [289]	374 Uuo Unbinilium [289]	375 Uuh Unheptennium [289]	376 Uuo Unbinilium [289]	377 Uuh Unheptennium [289]	378 Uuo Unbinilium [289]	379 Uuh Unheptennium [289]	380 Uuo Unbinilium [289]	381 Uuh Unheptennium [289]	382 Uuo Unbinilium [289]	383 Uuh Unheptennium [289]	384 Uuo Unbinilium [289]	385 Uuh Unheptennium [289]	386 Uuo Unbinilium [289]	387 Uuh Unheptennium [289]	388 Uuo Unbinilium [289]	389 Uuh Unheptennium [289]	390 Uuo Unbinilium [289]	391 Uuh Unheptennium [289]	392 Uuo Unbinilium [289]	393 Uuh Unheptennium [289]	394 Uuo Unbinilium [289]	395 Uuh Unheptennium [289]	396 Uuo Unbinilium [289]	397 Uuh Unheptennium [289]	398 Uuo Unbinilium [289]	399 Uuh Unheptennium [289]	400 Uuo Unbinilium [289]	401 Uuh Unheptennium [289]	402 Uuo Unbinilium [289]	403 Uuh Unheptennium [289]	404 Uuo Unbinilium [289]	405 Uuh Unheptennium [289]	406 Uuo Unbinilium [289]	407 Uuh Unheptennium [289]	408 Uuo Unbinilium [289]	409 Uuh Unheptennium [289]	410 Uuo Unbinilium [289]	411 Uuh Unheptennium [289]	412 Uuo Unbinilium [289]	413 Uuh Unheptennium [289]	414 Uuo Unbinilium [289]	415 Uuh Unheptennium [289]	416 Uuo Unbinilium [289]	417 Uuh Unheptennium [289]	418 Uuo Unbinilium [289]	419 Uuh Unheptennium [289]	420 Uuo Unbinilium [289]	421 Uuh Unheptennium [289]	422 Uuo Unbinilium [289]	423 Uuh Unheptennium [289]	424 Uuo Unbinilium [289]	425 Uuh Unheptennium [289]	426 Uuo Unbinilium [289]	427 Uuh Unheptennium [289]	428 Uuo Unbinilium [289]	429 Uuh Unheptennium [289]	430 Uuo Unbinilium [289]	431 Uuh Unheptennium [289]	432 Uuo Unbinilium [289]	433 Uuh Unheptennium [289]	434 Uuo Unbinilium [289]	435 Uuh Unheptennium [289]	436 Uuo Unbinilium [289]	437 Uuh Unheptennium [289]	438 Uuo Unbinilium [289]	439 Uuh Unheptennium [289]	440 Uuo Unbinilium [289]	441 Uuh Unheptennium [289]	442 Uuo Unbinilium [289]	443 Uuh Unheptennium [289]	444 Uuo Unbinilium [289]	445 Uuh Unheptennium [289]	446 Uuo Unbinilium [289]	447 Uuh Unheptennium [289]	448 Uuo Unbinilium [289]	449 Uuh Unheptennium [289]	450 Uuo Unbinilium [289]	451 Uuh Unheptennium [289]	452 Uuo Unbinilium [289]	453 Uuh Unheptennium [289]	454 Uuo Unbinilium [289]	455 Uuh Unheptennium [289]	456 Uuo Unbinilium [289]	457 Uuh Unheptennium [289]	458 Uuo Unbinilium [289]	459 Uuh Unheptennium [289]	460 Uuo Unbinilium [289]	461 Uuh Unheptennium [289]	462 Uuo Unbinilium [289]	463 Uuh Unheptennium [289]	464 Uuo Unbinilium [289]	465 Uuh Unheptennium [289]	466 Uuo Unbinilium [289]	467 Uuh Unheptennium [289]	468 Uuo Unbinilium [289]	469 Uuh Unheptennium [289]	470 Uuo Unbinilium [289]	471 Uuh Unheptennium [289]	472 Uuo Unbinilium [289]	473 Uuh Unheptennium [289]	474 Uuo Unbinilium [289]	475 Uuh Unheptennium [289]	476 Uuo Unbinilium [289]	477 Uuh Unheptennium [289]	478 Uuo Unbinilium [289]	479 Uuh Unheptennium [289]	480 Uuo Unbinilium [289]	481 Uuh Unheptennium [289]	482 Uuo Unbinilium [289]	483 Uuh Unheptennium [289]	484 Uuo Unbinilium [289]	485 Uuh Unheptennium [289]	486 Uuo Unbinilium [289]	487 Uuh Unheptennium [289]	488 Uuo Unbinilium [289]	489 Uuh Unheptennium [289]	490 Uuo Unbinilium [289]	491 Uuh Unheptennium [289]	492 Uuo Unbinilium [289]	493 Uuh Unheptennium [289]	494 Uuo Unbinilium [289]	495 Uuh Unheptennium [289]	496 Uuo Unbinilium [289]	497 Uuh Unheptennium [289]

Atomic Number

Symbol

Name

Atomic Mass

Alkali Metal

Alkaline Earth

Transition Metal

Basic Metal

Semimetal

Nonmetal

Halogen

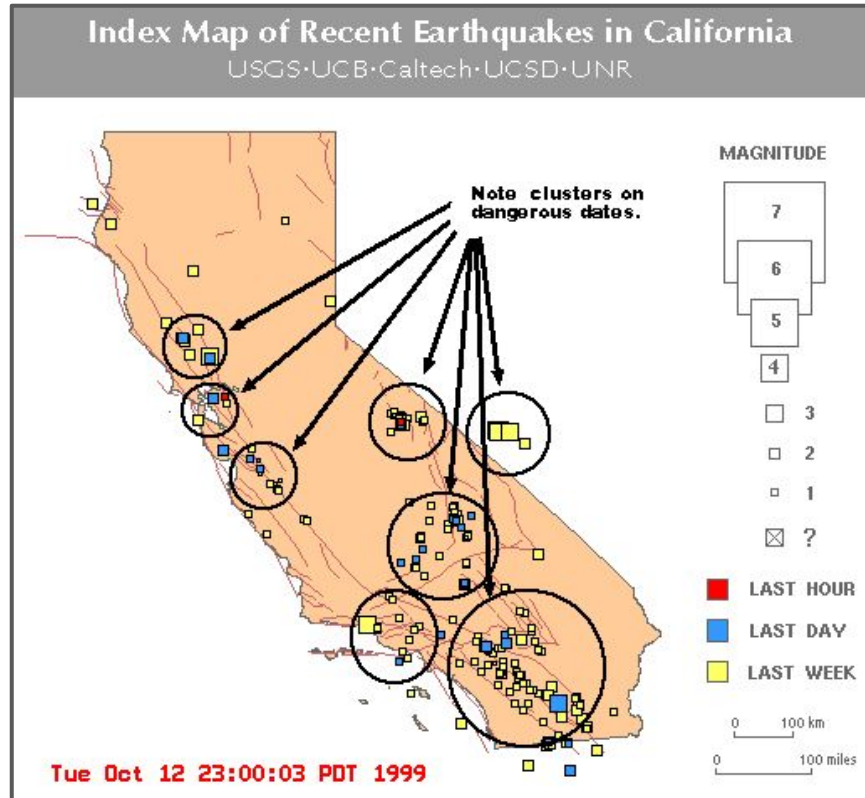
Noble Gas

Lanthanide

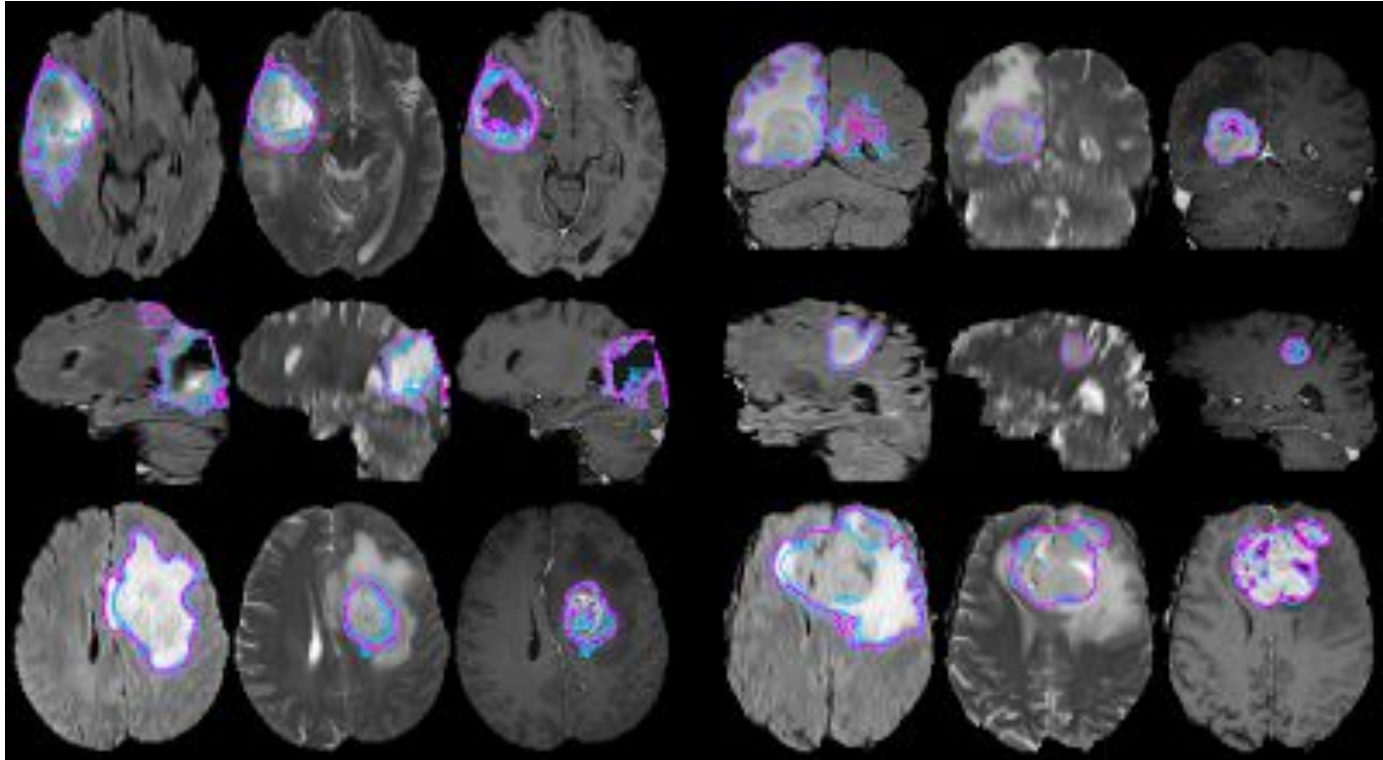
Actinide

© 2013 Todd Helmenstein
sciencenotes.org

Earthquakes and Seismology



Medical Imaging



Recommendation Systems: Amazon

Hello, Joshua B. Gordon. We have [recommendations](#) for you. ([Not Joshua?](#))

New: Kindle with Special Offers - \$114

Joshua's Amazon.com

Today's Deals

Gifts & Wish Lists

Gift Cards

Your Digital Items

Your Account

Help

Shop All Departments

Search All Departments

GO

Cart

Wish List

Your Amazon.com

Your Browsing History

Recommended For You

Rate These Items

Improve Your Recommendations

Your Profile

Learn More

Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to [see all recommendations.](#)

Page 1 of 35

[Introduction to Algorithms \(Hardcover\)](#) by Thomas H. Cormen
★★★★☆ (27) \$53.81
[Fix this recommendation](#)

[Programming Python \(Paperback\)](#) by Mark Lutz
★★★★☆ (9) \$39.23
[Fix this recommendation](#)

[Mining the Social Web: Analysis, Algorithms, and Visualization \(Paperback\)](#) by Matthew A. Richardson
★★★★☆ (13) \$26.12
[Fix this recommendation](#)

[Learning Python: Powerful Object-Oriented Programming \(Paperback\)](#) by Mark Lutz
★★★★☆ (27) \$33.86
[Fix this recommendation](#)

Coming Soon for You

[Madden NFL 12](#)
\$59.99
[Fix this recommendation](#)

See more recommended future releases

Tap into Your Friends

BETA

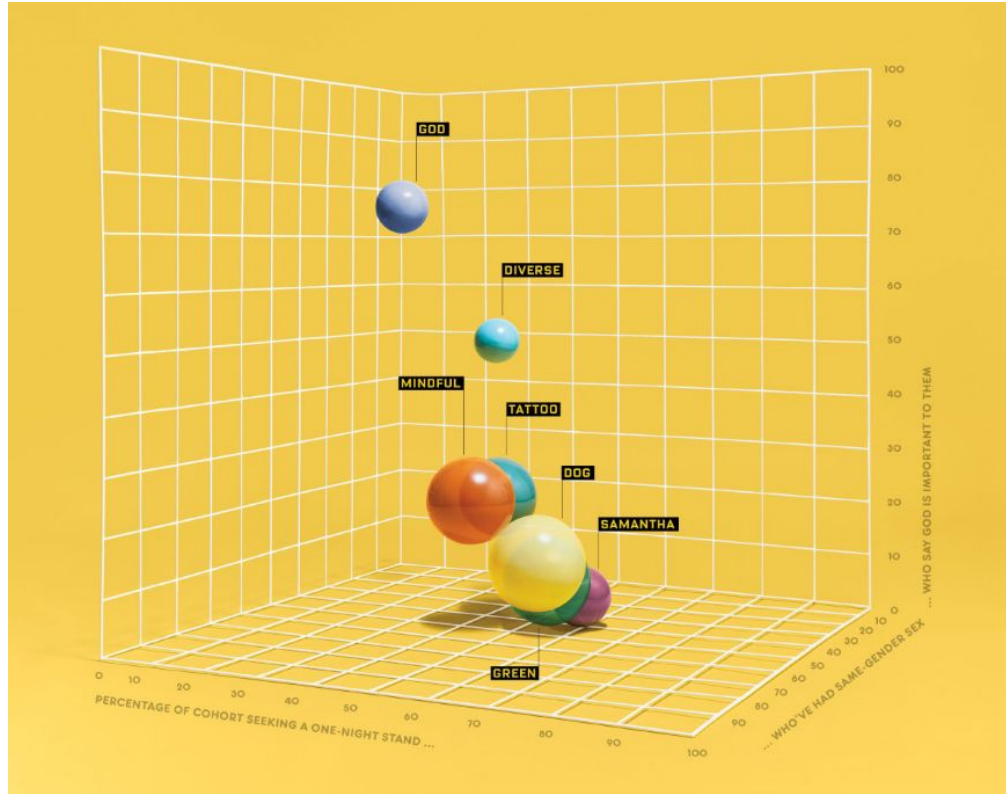
Connect to Facebook to get Amazon recommendations for you and discover your friends' Favorites and Likes

[Learn more and Connect](#)

Recommendation Systems: Amazon

- Amazon doesn't know what it's like to read a book, or what you feel like when you read a particular book
- Amazon *does* know that people who bought a certain book also bought other books
- Patterns in the data can be used to make recommendations
- If you've built up a long purchase history you'll often see pretty sophisticated recommendations

And even more nefarious purposes...



WIRED

KEVIN POULSEN SCIENCE 01.21.14 06:30 AM

HOW A MATH GENIUS HACKED OKCUPID TO FIND TRUE LOVE

Discussion:

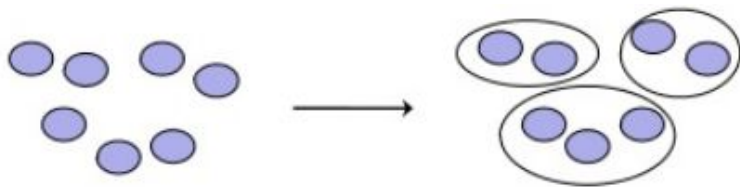
Can you come up with other applications of clustering?

Types of clustering

Two main types of clustering:

- **Hard clustering:**

Each object is in exactly one class.



- **Soft clustering:**

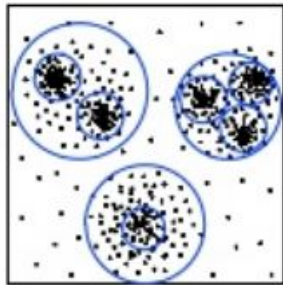
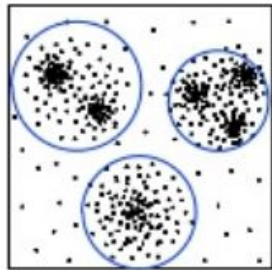
Objects are assigned a degree to which they are in each class.

- (Think of this as probability of being in a class)



Flat vs. Hierarchical clustering

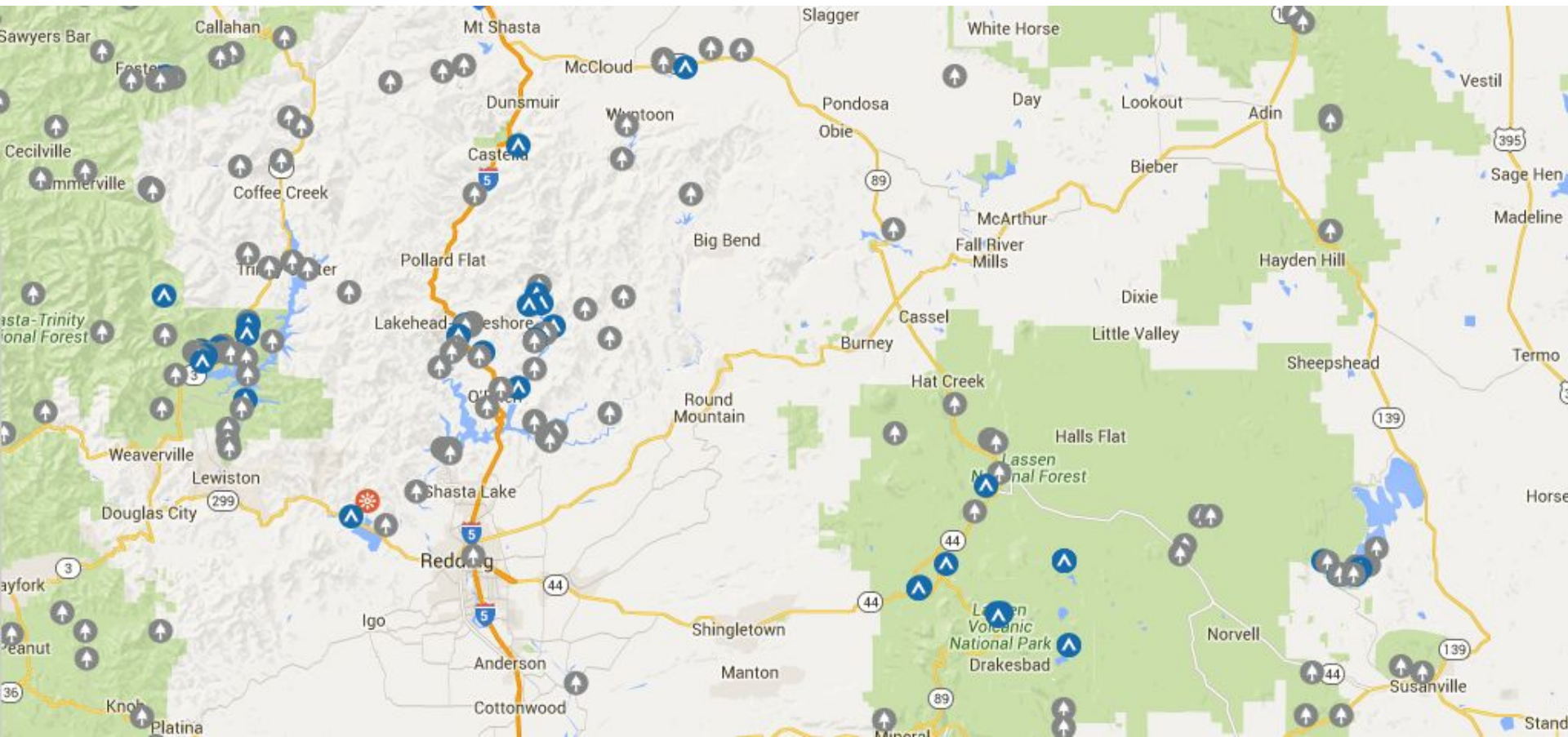
- Flat = distinct clusters
- Hierarchical = clusters within clusters



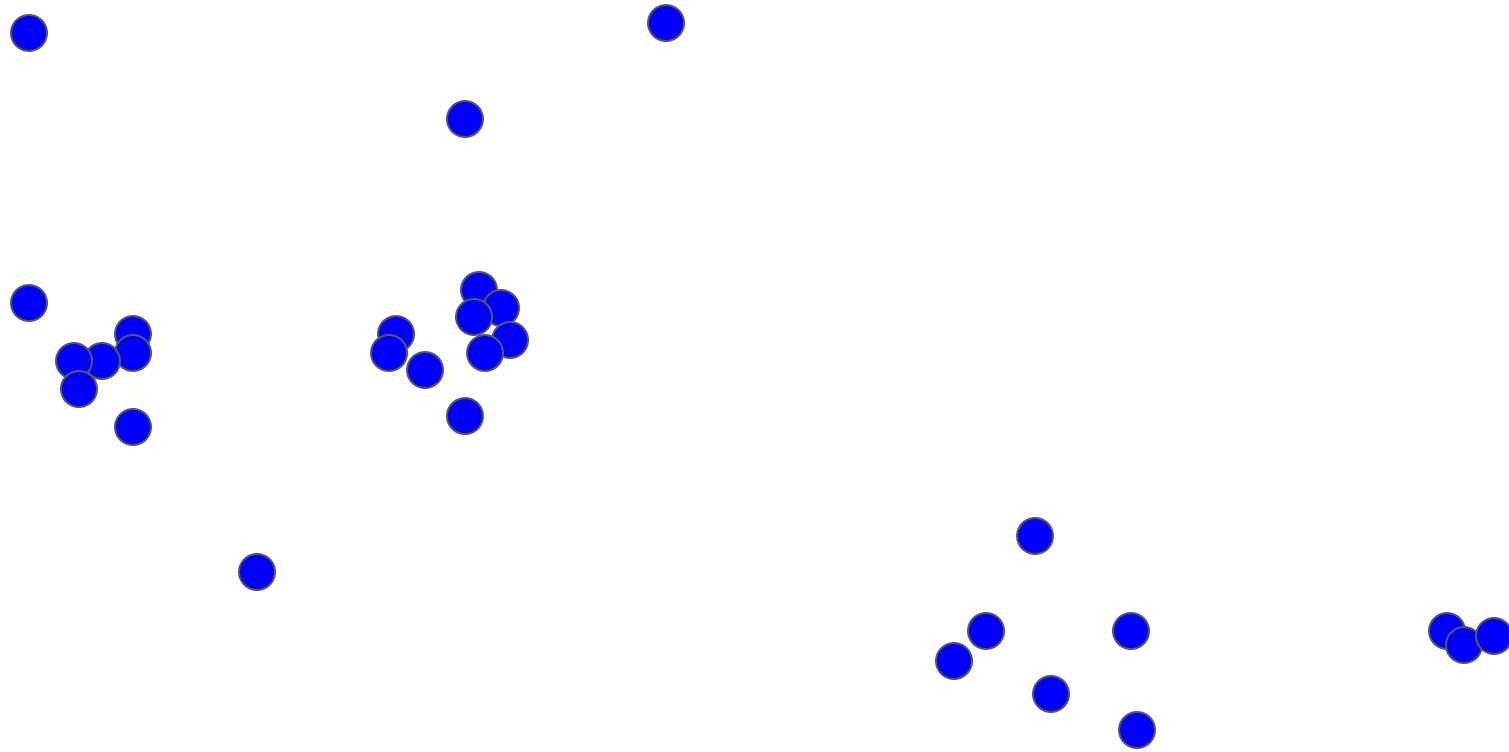
More later on types of clustering *algorithms*.
But first, an example...



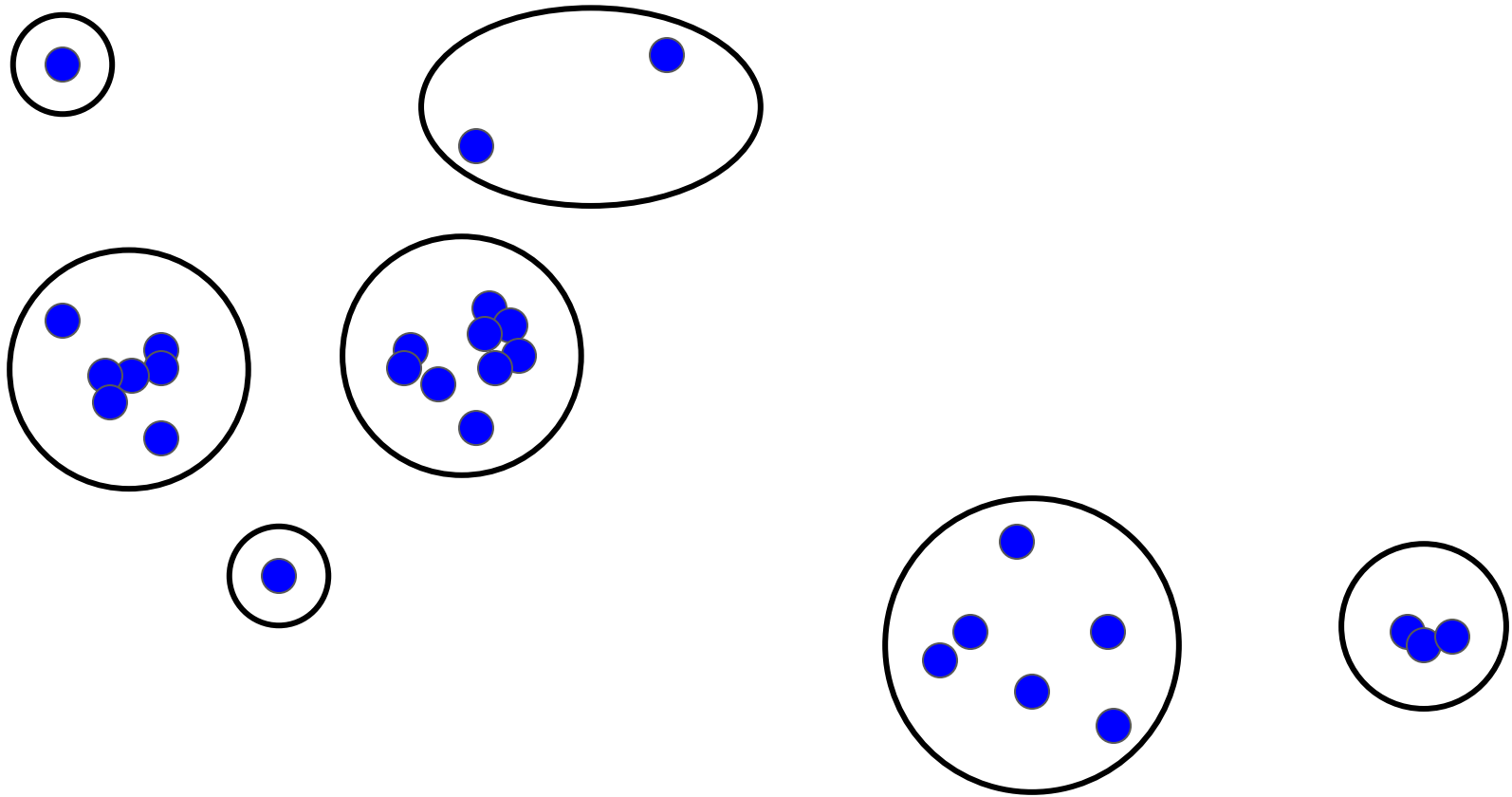
Campgrounds



Campgrounds



Campgrounds



Breakout Session:

Try out the demo at

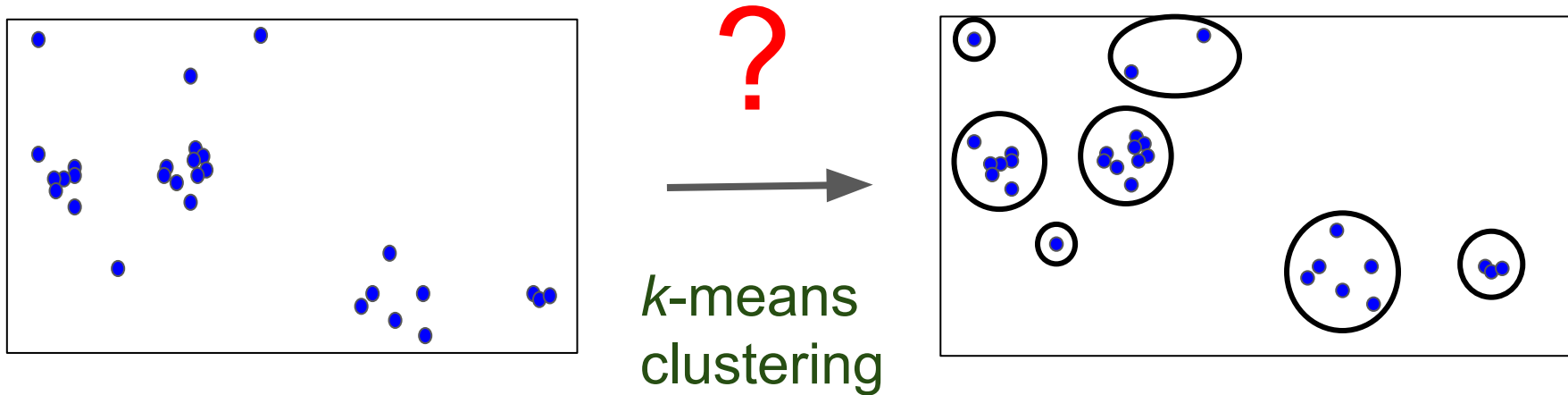
<http://mlehman.github.io/kmeans-javascript/>

Discuss with your table: how does the algorithm work?

(Note: this demo has a bug! Can't use more than 6 clusters)

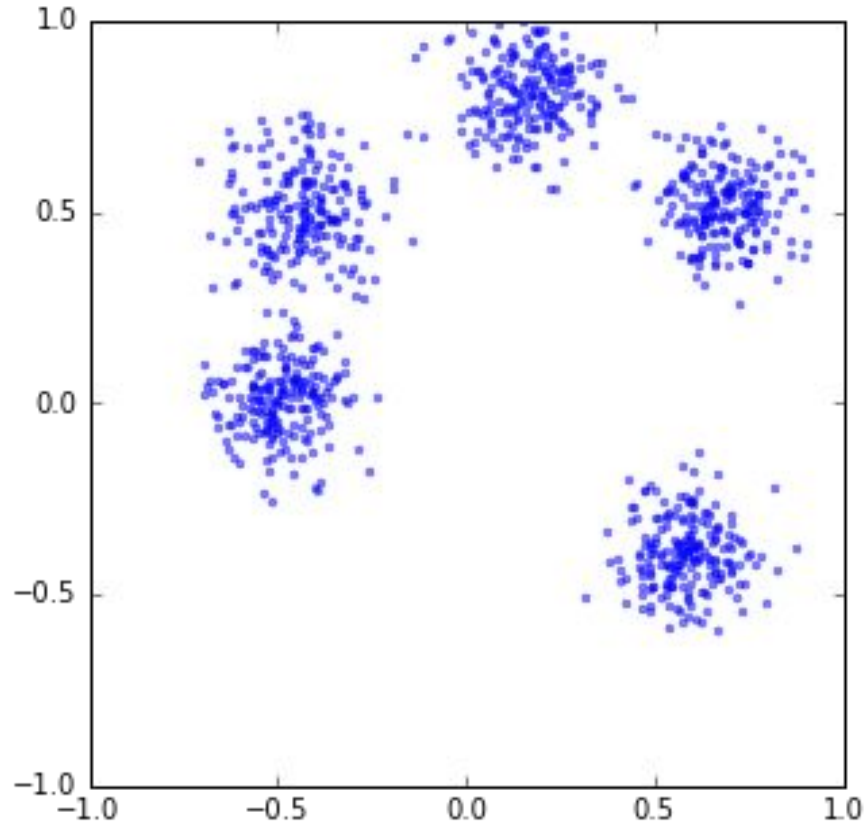
k-means clustering

How can we automatically cluster campgrounds?

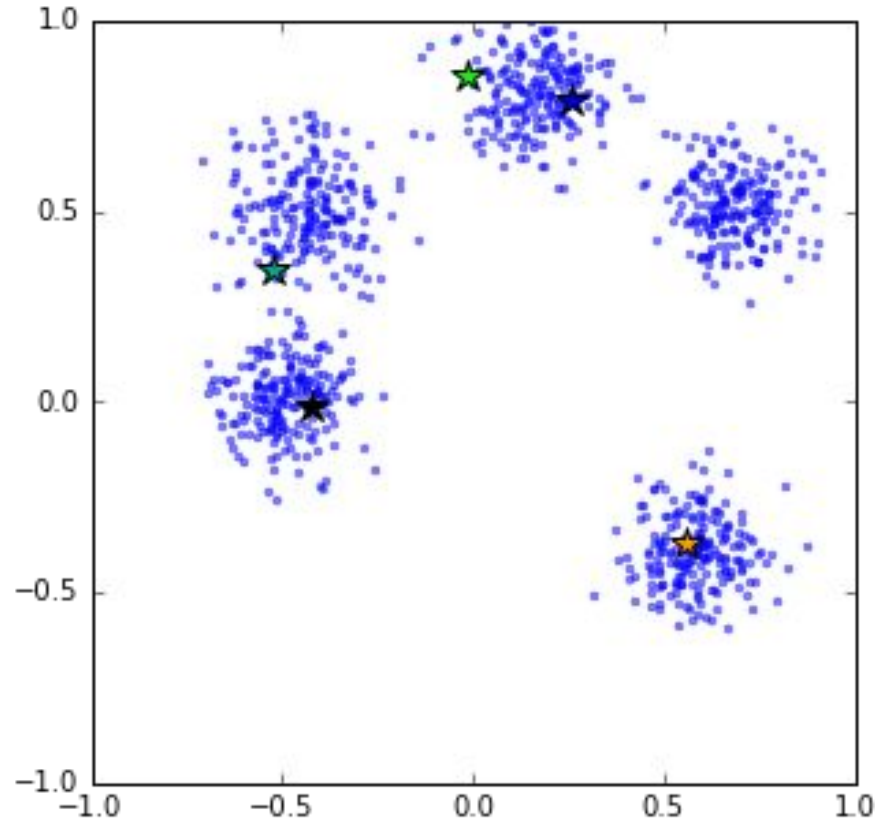


Goal: Assign each of the n points to one out of k clusters (defined by a cluster centroid).

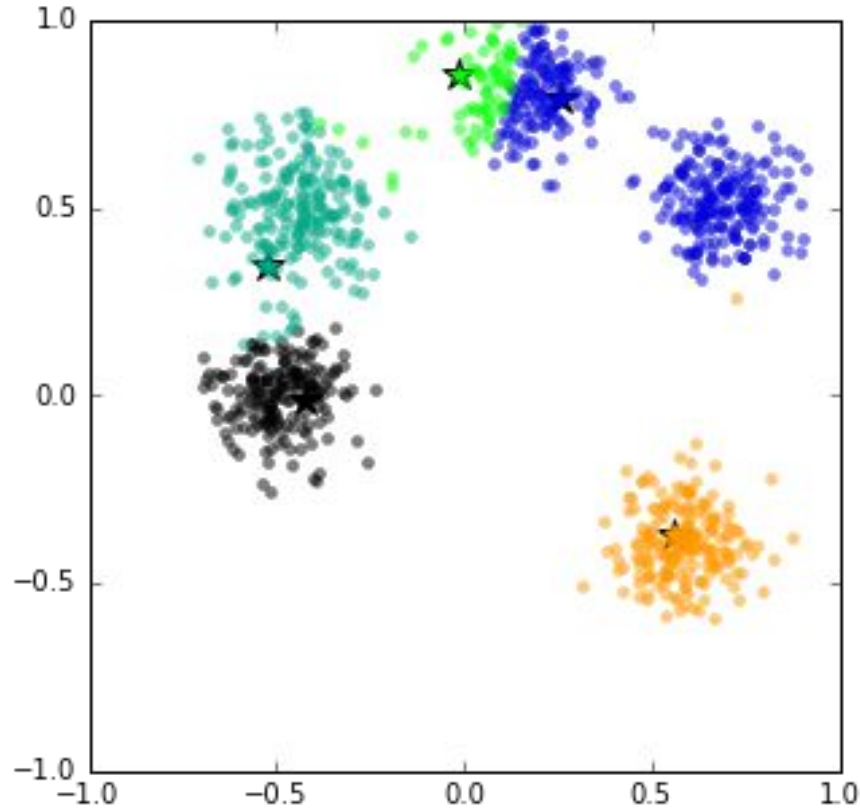
Find 5 clusters for the 1000 points below!



Step 0: Pick k random points as the cluster centroids

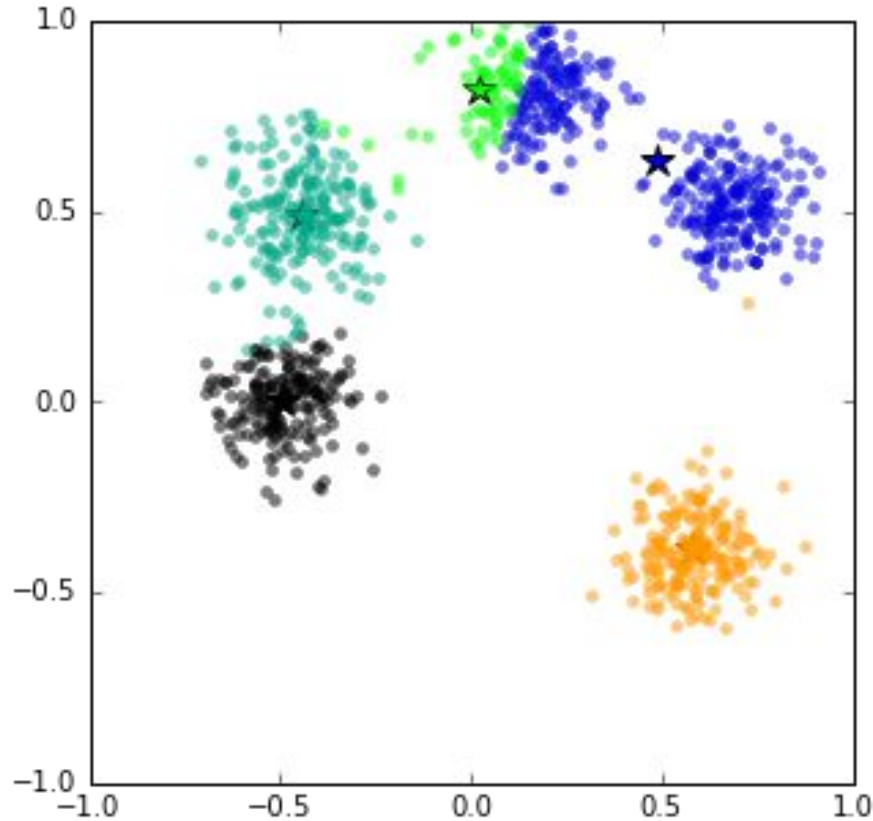


Iteration 1



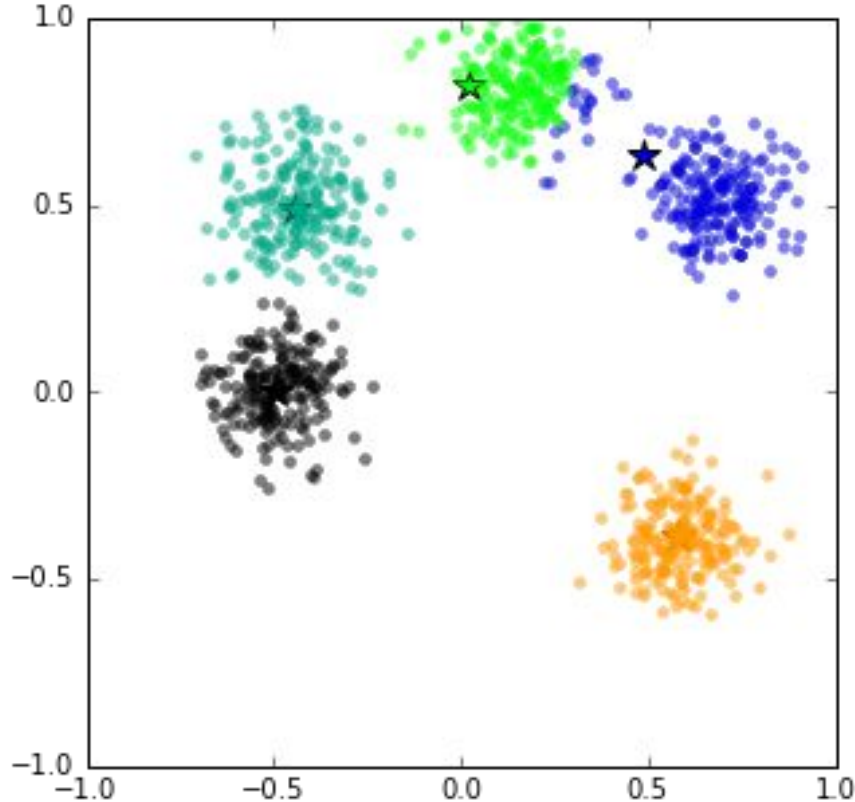
Step 1: Assign each point to closest cluster centroid

Iteration 1 continued



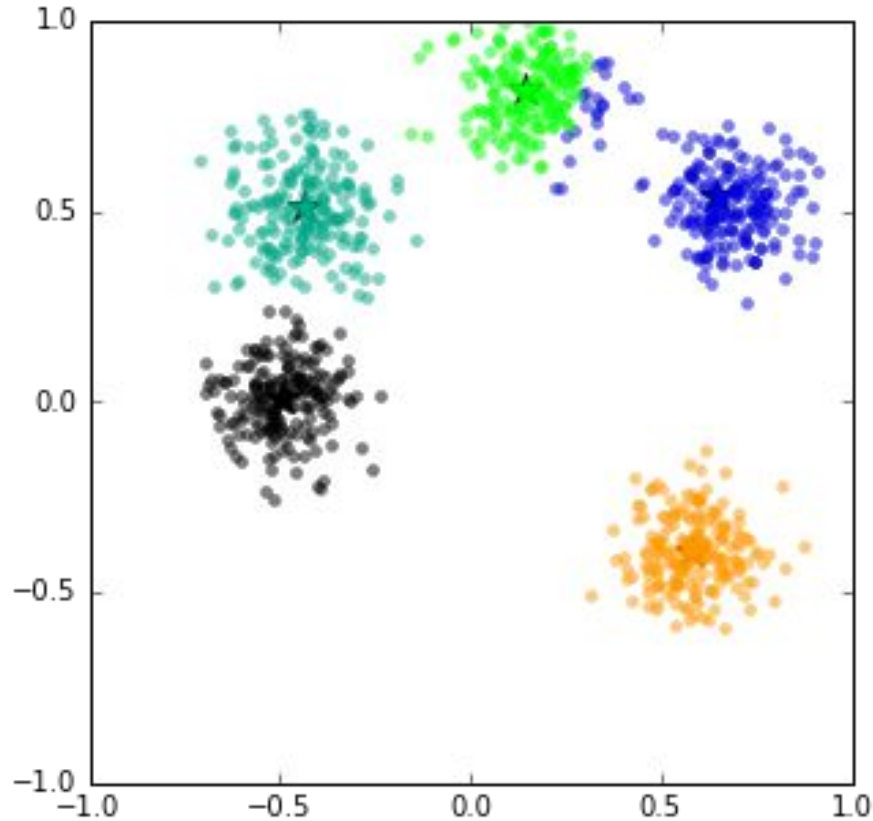
Step 2: Find cluster centroids as center of mass of elements

Iteration 2



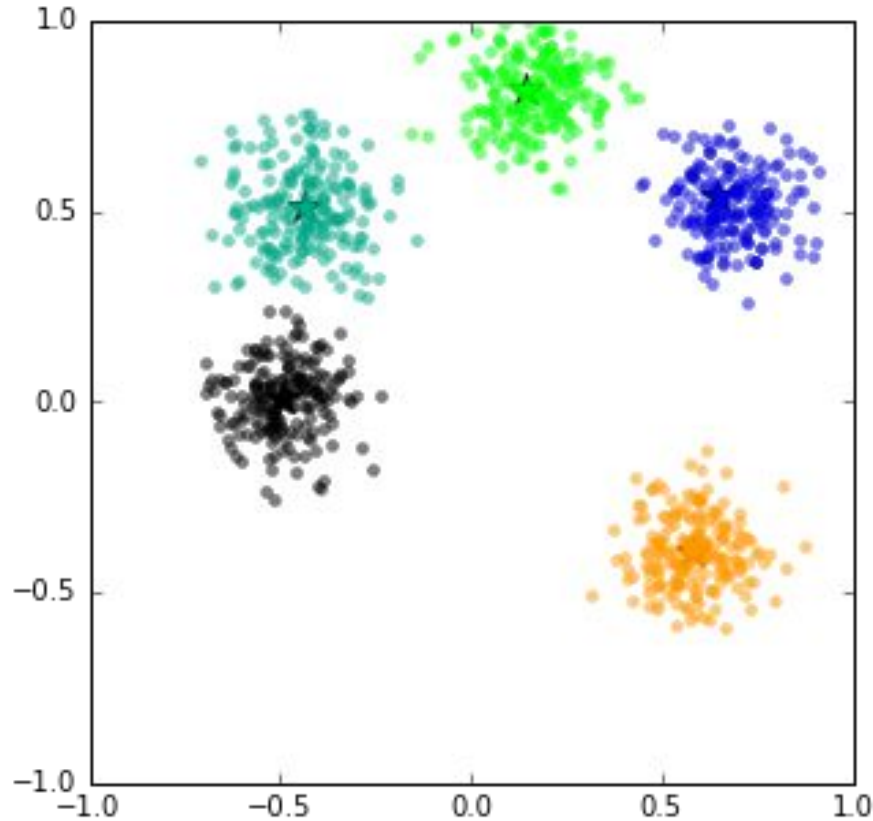
Step 1: Assign each point to closest cluster centroid

Iteration 2 continued



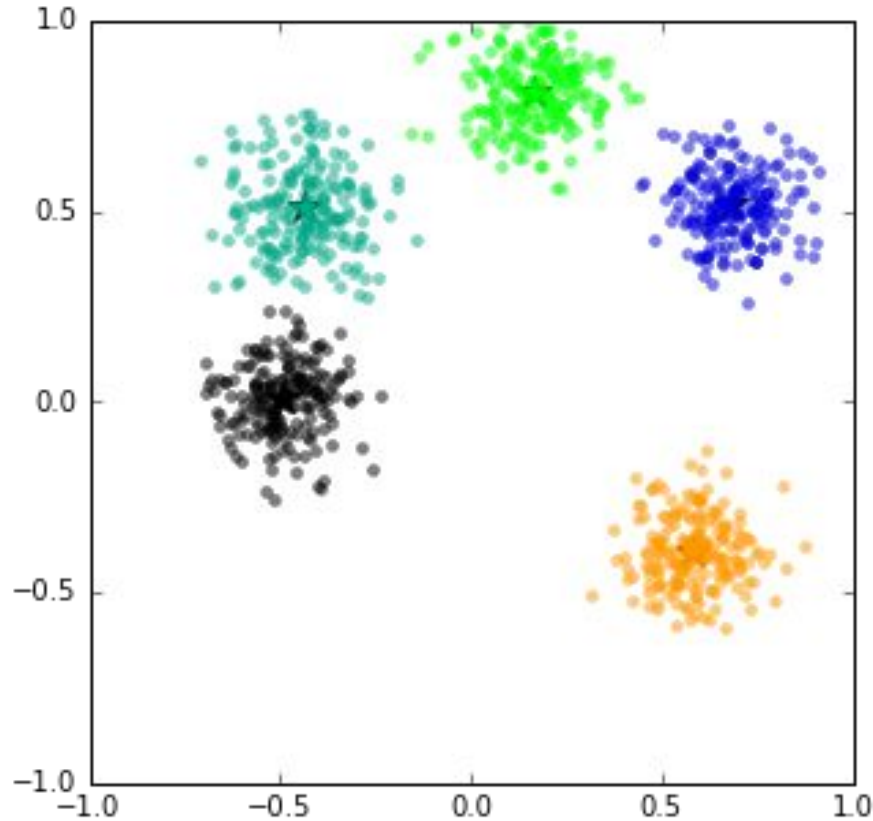
Step 2: Find cluster centroids as center of mass of elements

Iteration 3



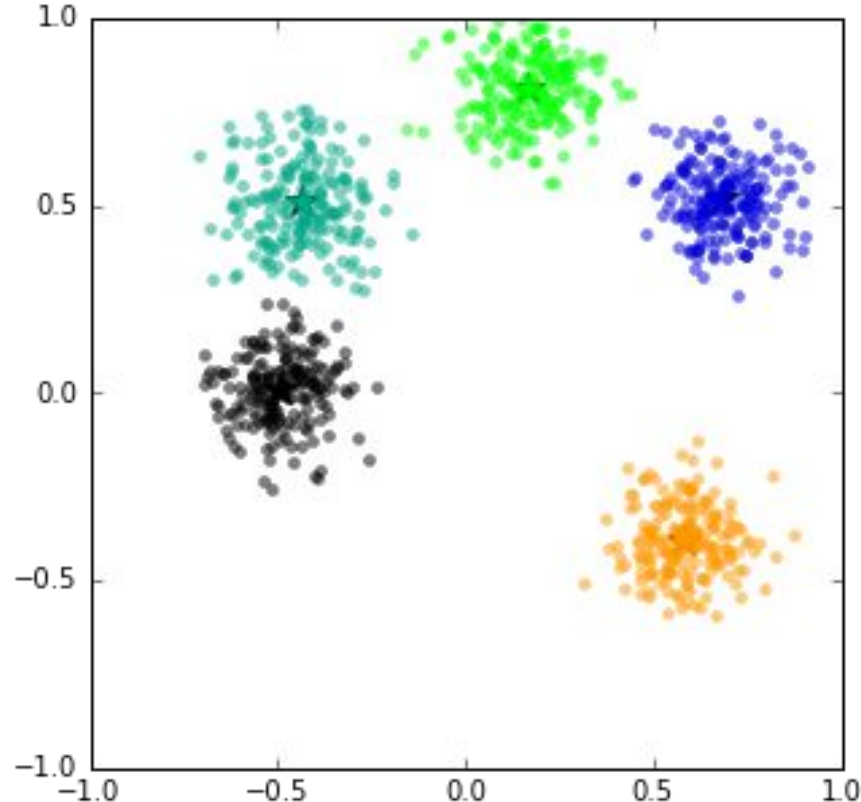
Step 1: Assign each point to closest cluster centroid

Iteration 3 continued



Step 2: Find cluster centroids as center of mass of elements

Iteration 4



Step 1: Assign each point to closest cluster centroid

No change in cluster assignment.
TERMINATE!

k-means overview

- Choose the value of k
 - This is how many cluster centroids we're finding
- Choose k points in the set
 - These are the initial centroid locations
- For each point not selected, assign to its nearest centroid
 - Now all points have an initial cluster assignment
- Until “happy” do:
 - 1. Recompute centroids of clusters
 - 2. Reassign all points to closest centroid (forms new clusters)



Discussion:

Is this “hard” or “soft” clustering?

Is this “flat” or “hierarchical” clustering?

Discussion:

When would k-means fail?

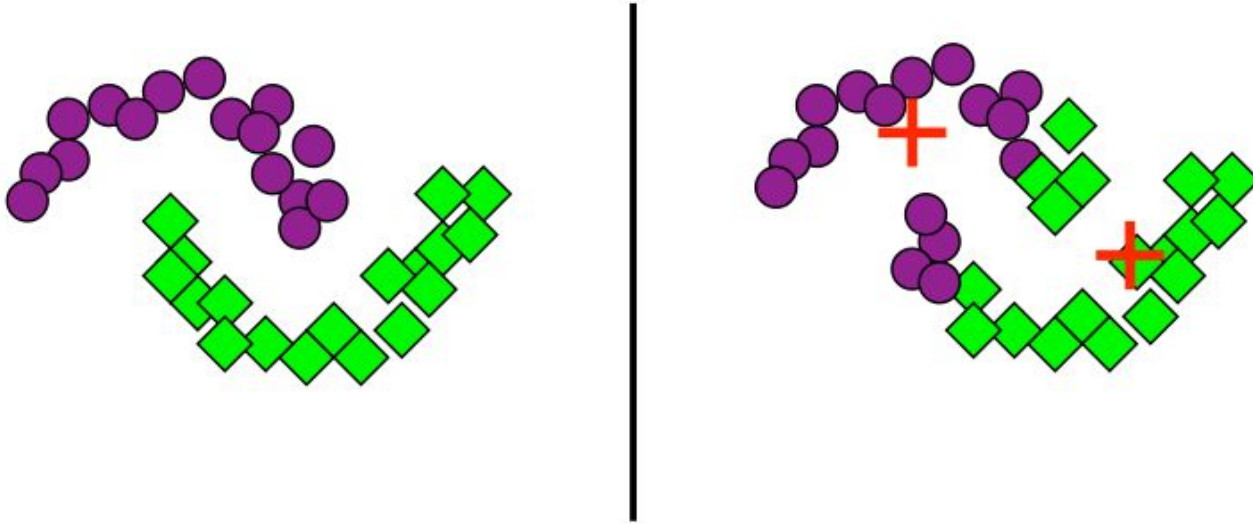
When would it succeed?

What are some drawbacks?

What are some advantages?

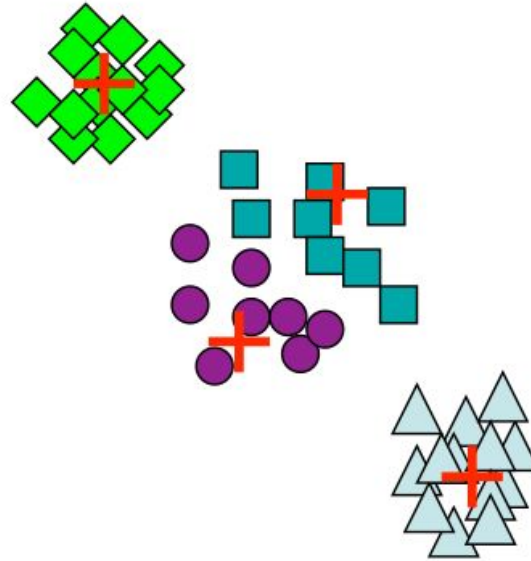
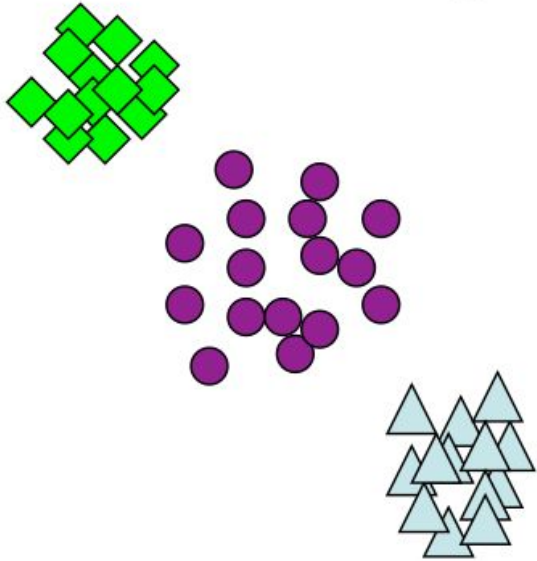
Weaknesses:

Non-globular clusters



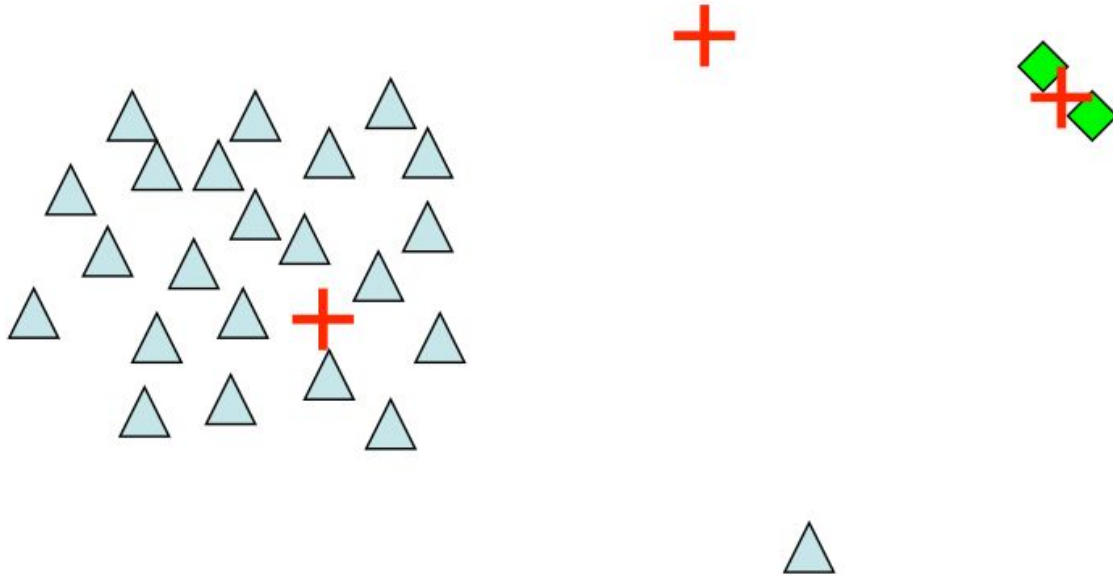
Weaknesses:

Wrong number of clusters



Weaknesses:

Outliers and empty clusters



Algorithm design: Distance metrics and features

- How do we determine what elements are “close”?
 - Spatial distance, similarity of campsite attributes,...
- “Features” are attributes that we use to mathematically compute closeness.
 - Height, age, hometown
- Weights for features
 - Are all features equally important or are some more important than others?
 - Are the features on different scales?



Types of features

- **Categorical features** take a fixed set of values. The values cannot be ordered.
 - For example, the type of phone (android, iphone, windows phone).
- **Ordinal features** take a fixed set of values. The values can be ordered.
 - For example, the ranks of football teams in the PAC-12 conference.
- **Continuous features** can take any real value.
 - For example, the distance of Princeton University from your hometown.



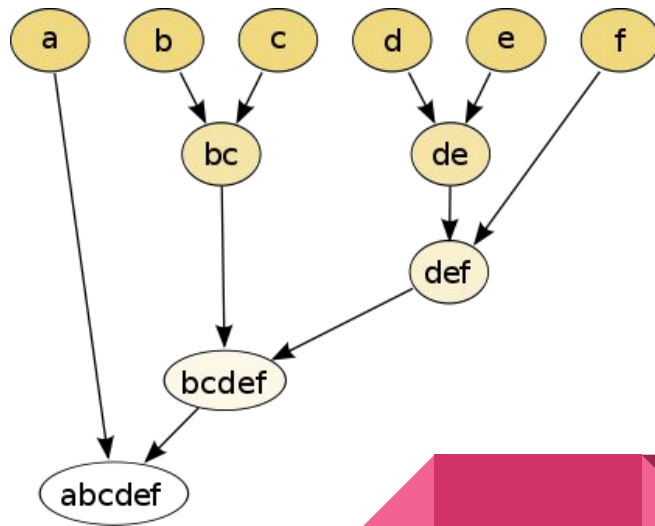
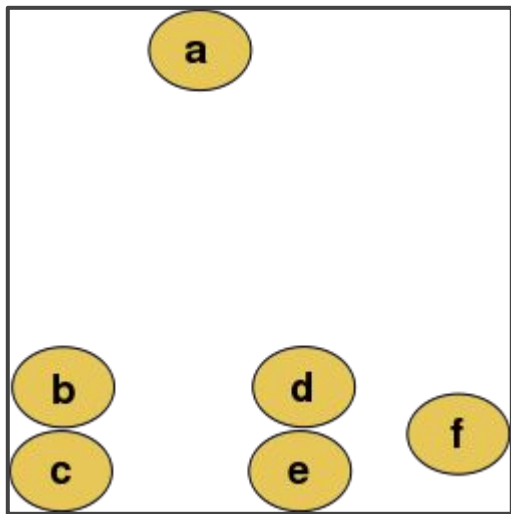
Main types of clustering algorithms

- Connectivity-based (hierarchical) clustering
- Centroid-based clustering
- Distribution-based clustering
- Density-based clustering



Connectivity-based (hierarchical) clustering

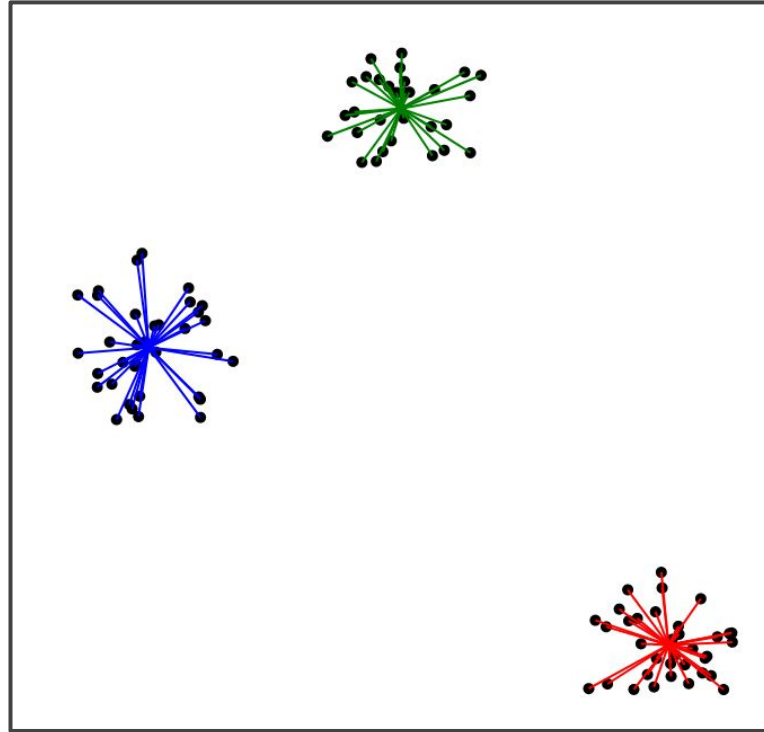
- A hierarchy of clusters based on distance:



- Typically, form a “dendrogram”

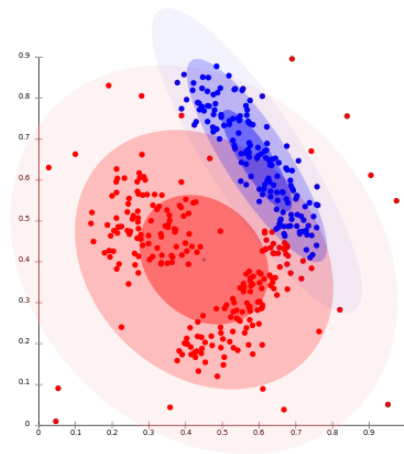
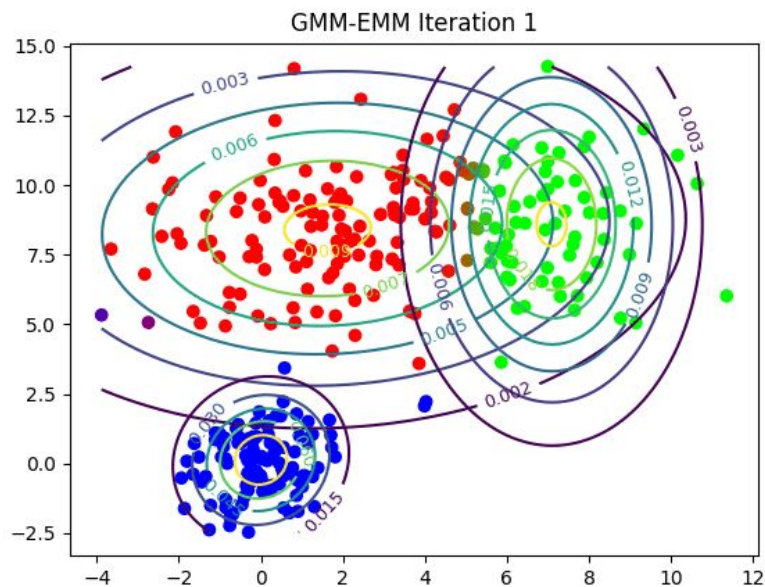
Centroid-based clustering

- k-means!



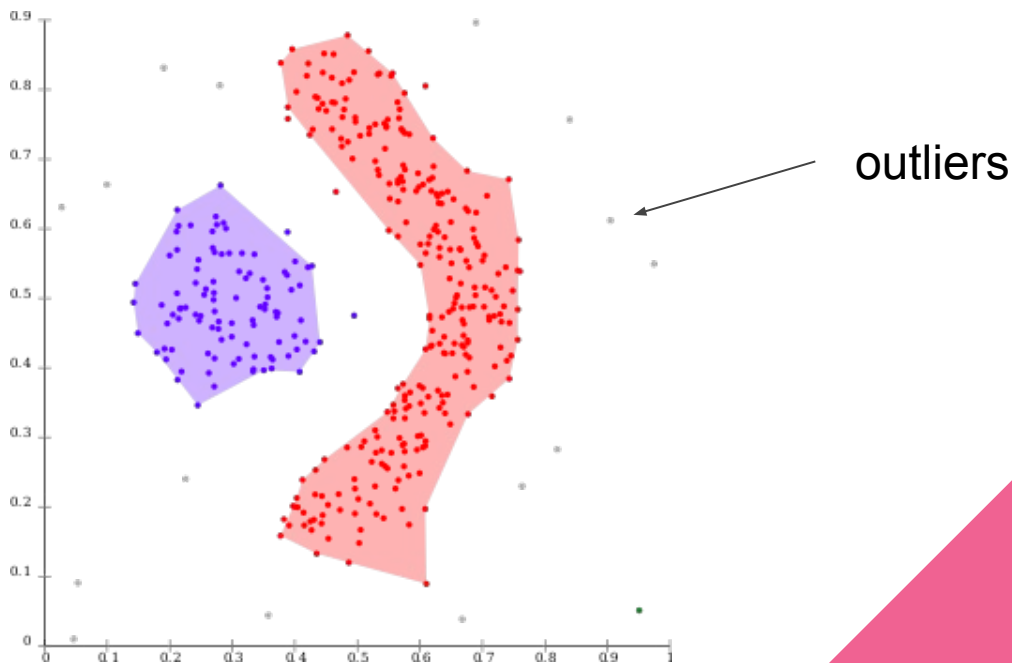
Distribution-based clustering

- Cluster according to how likely it is that points lie in a certain distribution



Density-based clustering

- Group together points that are closely spaced, and marks farther-out points as outliers



Recommendation system example

Dataset: Movie critics

Critic	Star Wars	Raiders of the Lost Arc	Casablanca	Singin' in the Rain
Sam	****	****	*	**
Sandy	*****	****	**	*
Matt	**	**	****	***
Julia	**	*	***	****
Sarah	*****	?	?	**

- Could an algorithm use this data to recommend movies?
- How would you do it?

Dataset: Movie critics

Critic	Star Wars	Raiders of the Lost Arc	Casablanca	Singin' in the Rain
Sam	****	****	*	**
Sandy	*****	****	**	*
Matt	**	**	****	***
Julia	**	*	***	****
Sarah	*****	?	?	**

- Could an algorithm use this data to recommend movies?
- How would you do it?

Critics with similar taste

■ Preference space

Star Wars

5

4

3

2

1

Julia

Matt

Sam
Sandy

1

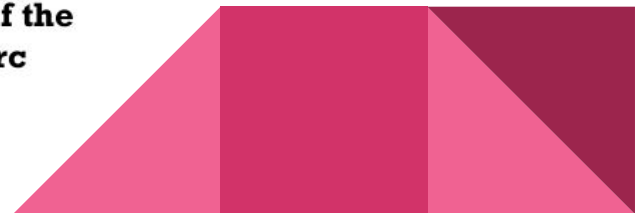
2

3

4

5

**Raiders of the
Lost Arc**



Measuring distance

Star Wars

5

4

3

2

1

Julia

Matt

Sam
Sandy

1

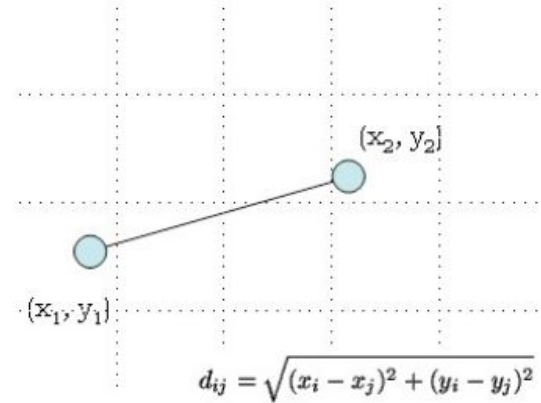
2

3

4

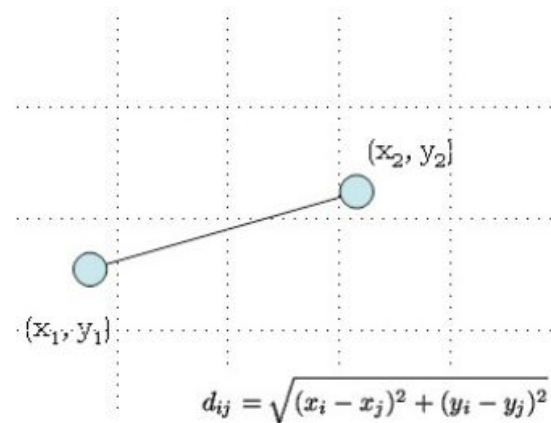
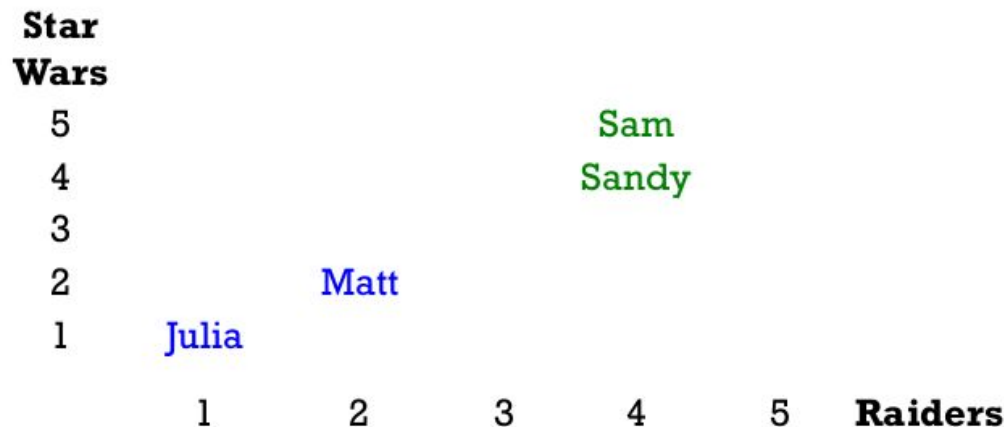
5

Raiders



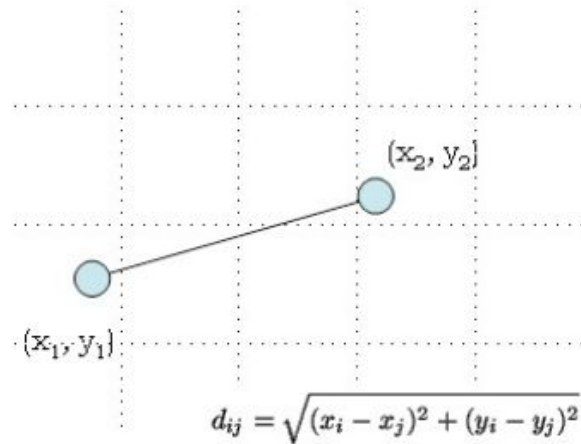
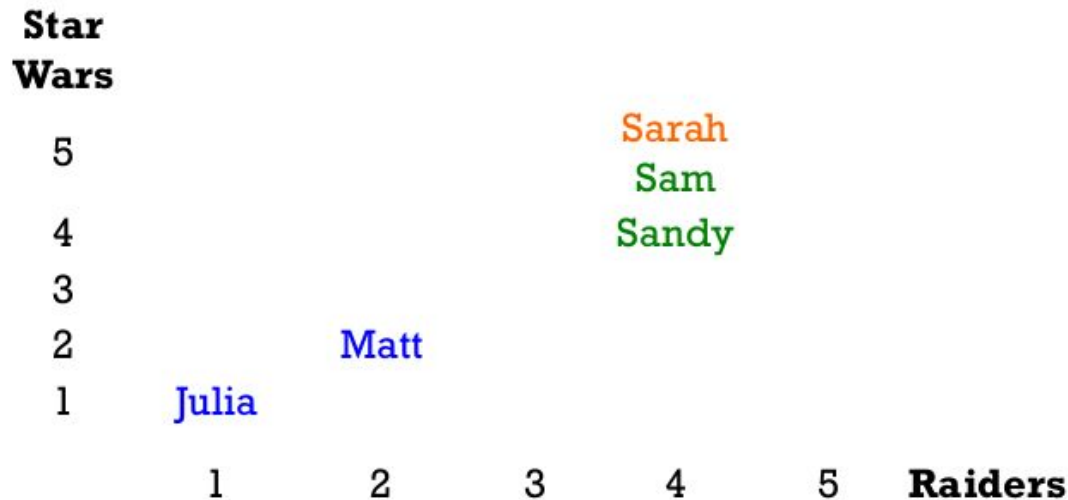
- Measure similarity between points using a measure of distance

Finding critics with similar taste



- People who liked Star Wars are close in preference space to those who liked Raiders

Making a recommendation



- Sarah hasn't seen Raiders, but gave Star Wars five stars.
- Chances are she'll like Raiders too!

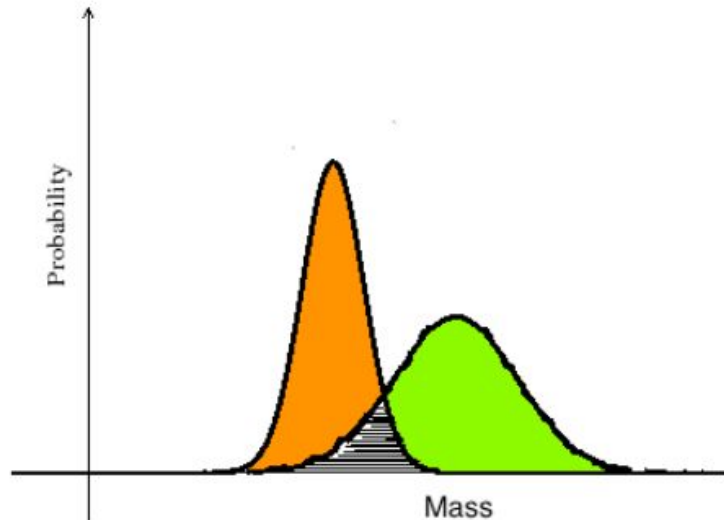
Features

- We used features to compare critics
- Feature: a data attribute used to make a comparison
- Quantify attributes of an object (size, weight, color, shape, density) in a way a computer can understand
- Quality is important



Apples vs. Oranges

- A good feature discriminates between classes
- Think: how well does a feature help us tell two things apart?
- Is mass a good feature? By itself?
- What about in conjunction with another feature like color?



Features to compare movies

Feature	Star Wars	Raiders of the Lost Arc	Casablanca	Singin' in the Rain
...				
...				
...				
...				
...				

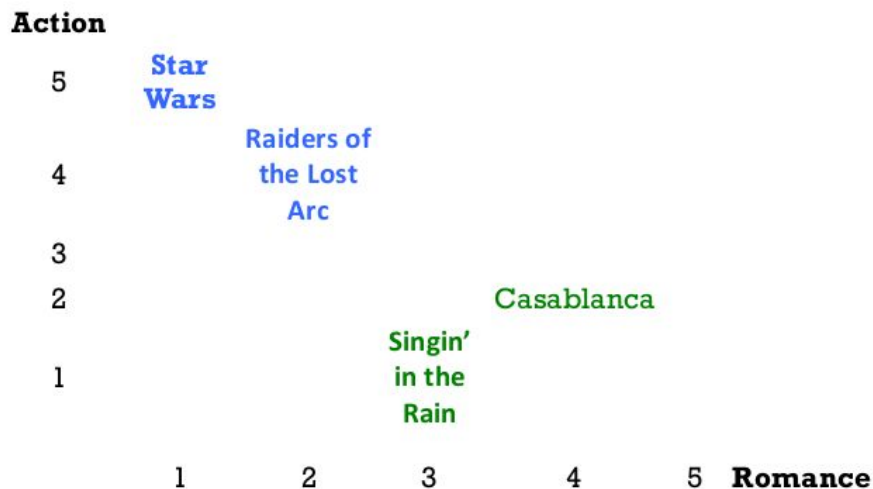
- Can you suggest some features to compare movies?

Features to compare movies

Feature	Star Wars	Raiders of the Lost Arc	Casablanca	Singin' in the Rain
Action (1 to 5)	5	4	2	1
Romance (1 to 5)	1	2	4	3
Length (min)	121	115	102	103
Harrison Ford	Y	Y	N	N
Year	1977	1981	1942	1952

- What type of features are these?

Comparing movies in feature space



- Here we can “eyeball” the clusters, but what if we had more features? (i.e. higher dimensional data)?



Exercise time:

Jupyter notebook available at

https://github.com/juliennelachance/ai4all_clustering