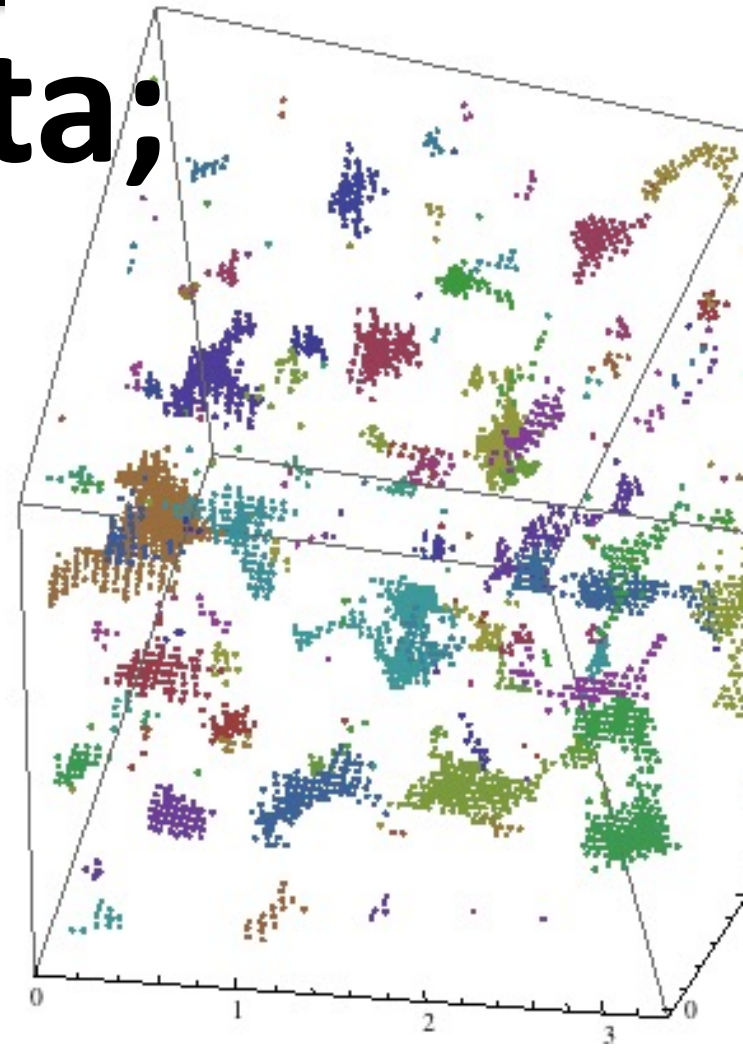
clustering numerical data; document clustering



- Topics from last week
- Clustering context
- Clustering methods

How do I find relevant analysis capabilities in Jupyter notebooks?

1. Has anyone else ever encountered this task?
2. Google search for desired capability

- Review notebooks from this class (posted to BlackBoard)
- Python Standard Library
- https://docs.python.org/3/py-modindex.html
- https://wiki.python.org/moin/UsefulModules
- Python 3 Module of the Week
- https://pypi.org/

Once you've found a library, how to learn more: Activity (part 1 of 2)

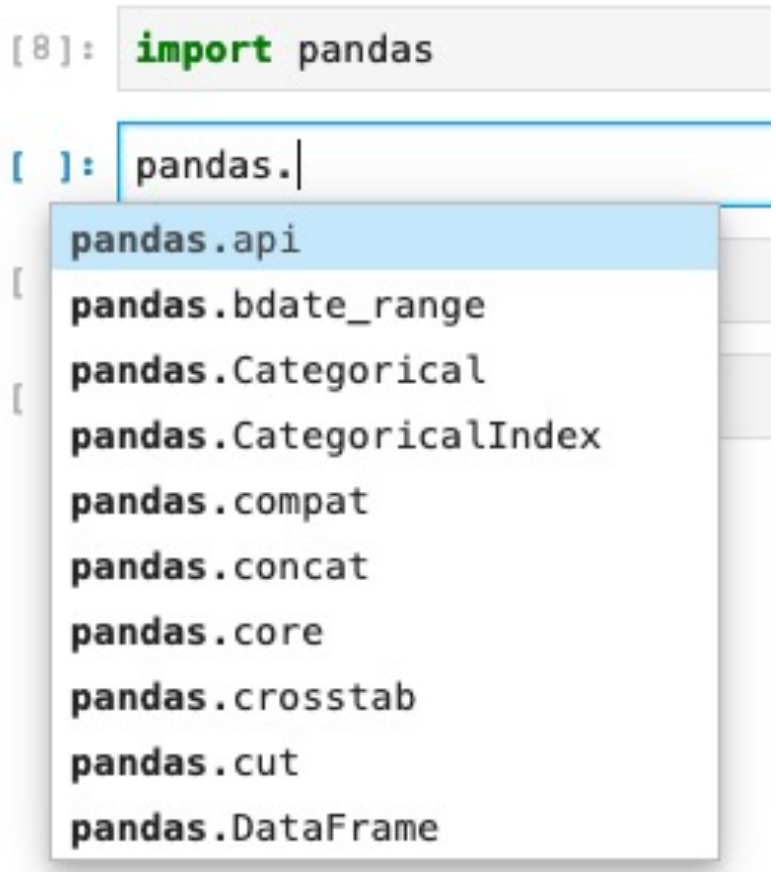
1. Open a Jupyter notebook
2. Type the four commands below into four separate cells

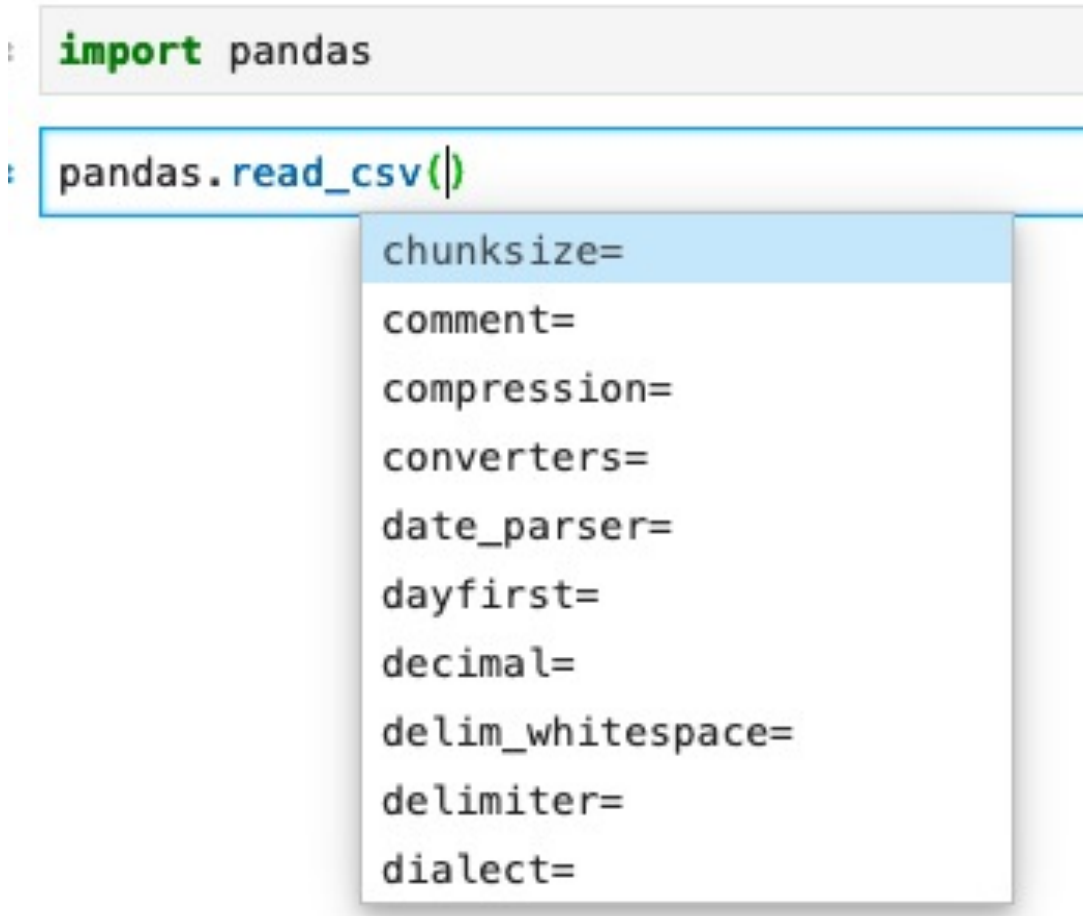
```
import random
dir(random)
help(random)
random?
```

Demo: `dir_help_tab_completion.ipynb`

Activity (2 of 2): Tab completion in Jupyter notebooks

Type the commands below in a code cell





Complex commands do not just appear;

```
dframe = pandas.read_csv("RollingSystemDemand_20180901_0129.csv",
                         index_col=False,
                         skiprows=1,
                         skipfooter=1,
                         engine='python',
                         header=None)

dframe.columns=['VD','time of measurement','value']

dframe.head()
```

I don't memorize these commands. I build them iteratively.
The result is concise; it hides hours of detective work.

Python errors

source

Traceback = sequence of function calls that led to an error.

Demo: `debugging_Python.ipynb`

Outcomes for this evening

By the end of today's class, you should be able to do the following:

- Load function from .py file in a .ipynb notebook
- Explain difference between agglomerative and divisive clustering
- Use k-Means to identify subsets of data
- Identify common "stop words"
- Write simple regular expressions

Load function from .py file in a .ipynb notebook

Demo: `use_a_function_in_a_notebook.ipynb`

- ~~Topics from last week~~
- Clustering context
- Clustering methods

Linear regression: fit data for multiple purposes

- [Week 7] Measure Correlation of two variables
- [Week 9] Fitting Trends

- Prediction
  - Extrapolation = independent variable (x) being evaluated is outside the range of values you already have information on
  - Interpolation = value you are evaluating is within the range of values you already know

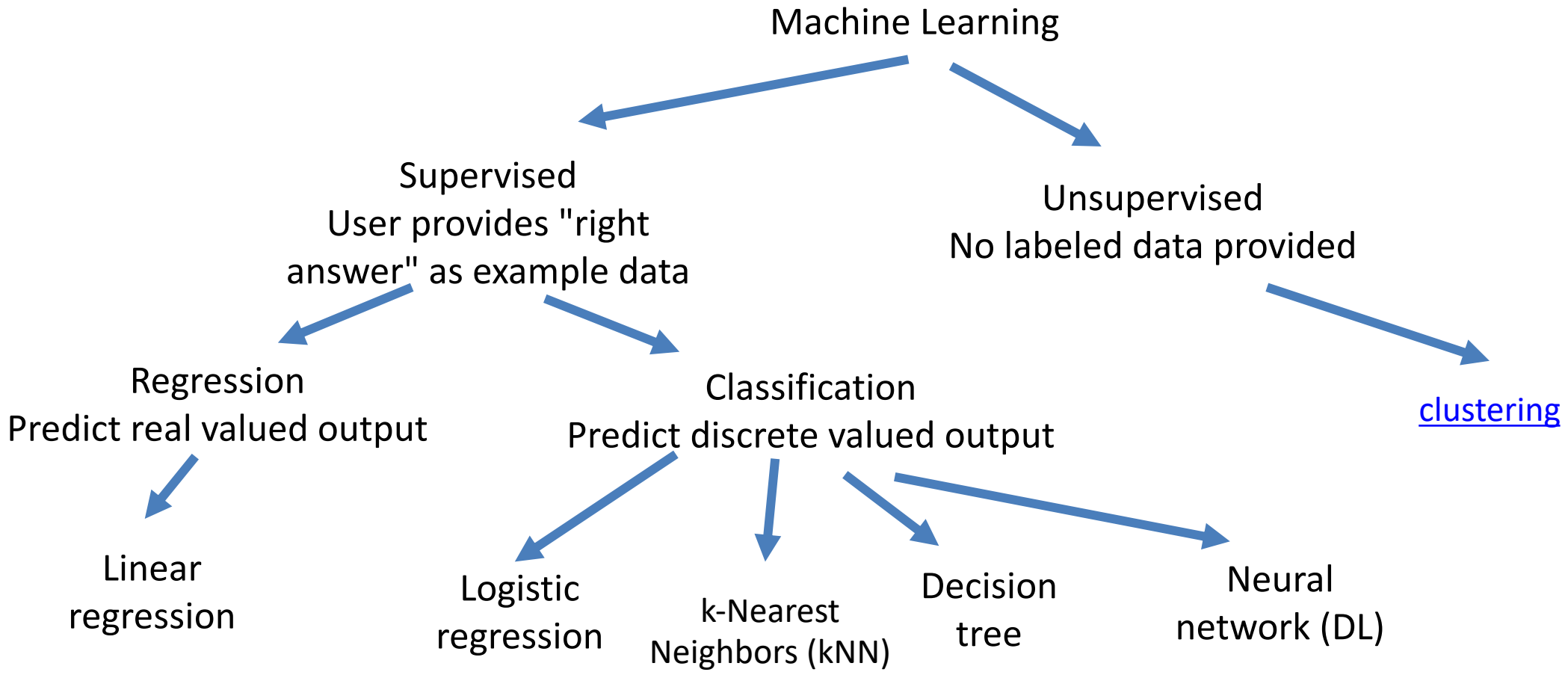
Linear regression: fit data for multiple purposes

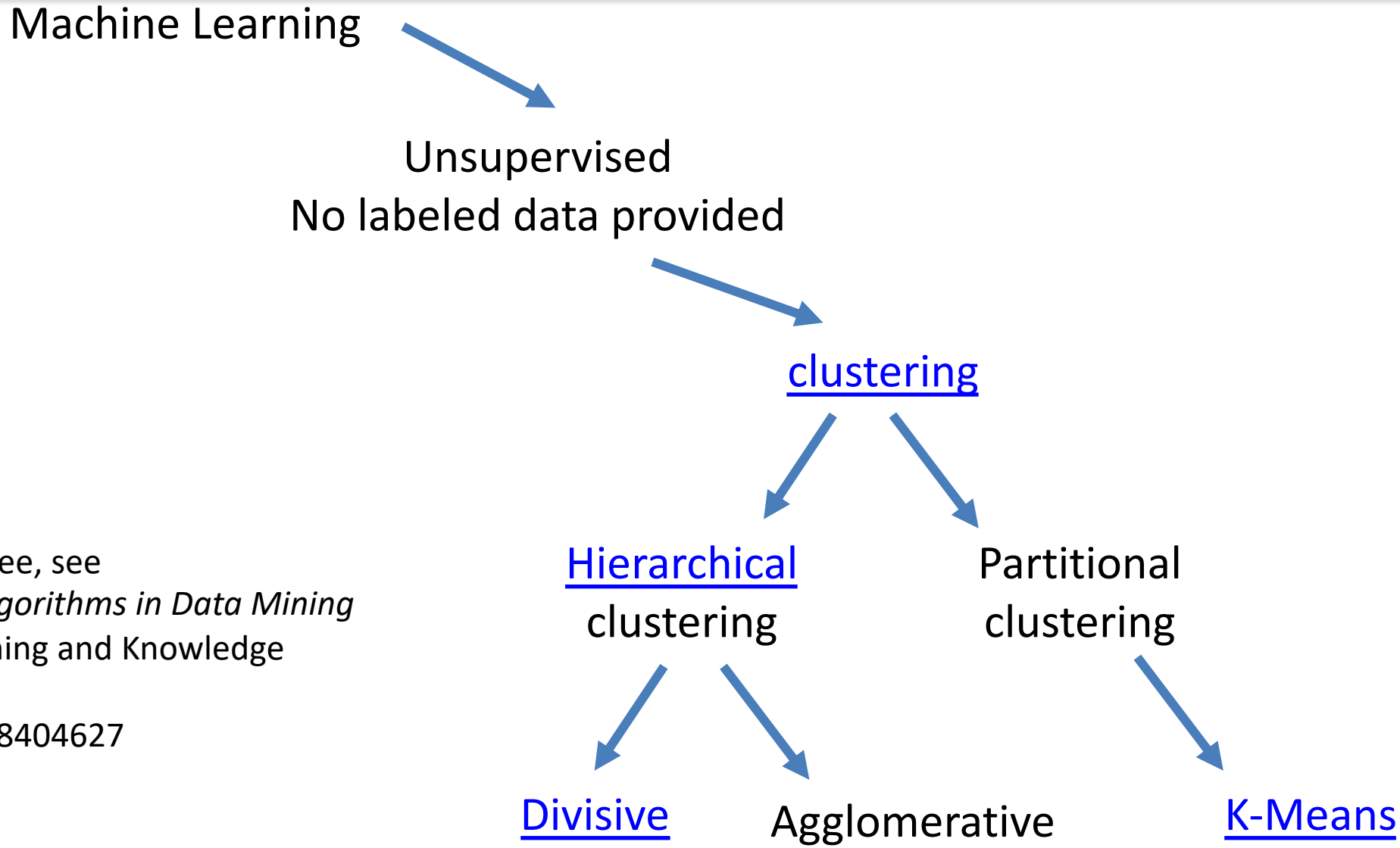
- [Week 7] Measure Correlation of two variables
- [Week 9] Fitting Trends

- Prediction
  - Extrapolation = independent variable (x) being evaluated is outside the range of values you already have information on
  - Interpolation = value you are evaluating is within the range of values you already know

--> Supervised machine learning

Map of Machine Learning topics





For a better overview tree, see
Techniques of Cluster Algorithms in Data Mining
October 2002; Data Mining and Knowledge
Discovery 6(4):303-360
DOI: 10.1023/A:1016308404627

Application for clustering: Search engine supplies similar topics

http://yippy.com/ is a metasearch site that groups results into "nearby" topics

Activity: in your web browser, go to yippy.com and search for data science

Application for clustering: Customer or Market segmentation

- important for optimization of marketing strategies.

- Better targeting of consumers through behavioral analysis of purchases, timing, which channel gets accessed, customer attributes

Application for clustering: Detecting anomalies

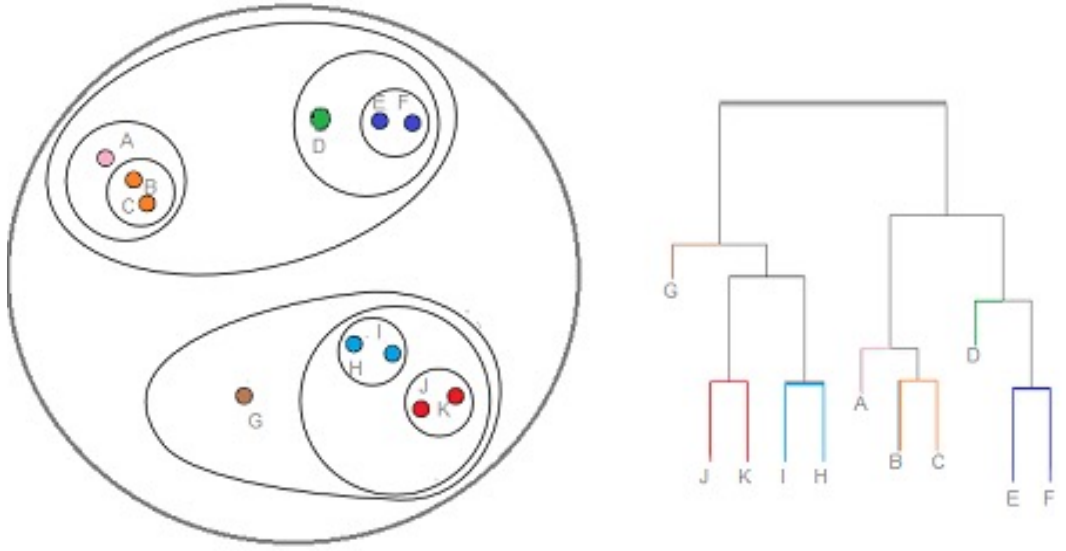
- Group valid activity to enable outlier detection

– monitoring if a tracked data point switches between groups over time can be used to detect meaningful changes in the data.

- ~~Topics from last week~~
- ~~Clustering context~~
- Clustering methods

Clustering approaches

- Partitional algorithms: Construct various partitions and then evaluate them by some criterion
- Hierarchical algorithms: Create a hierarchical decomposition of the set of objects using some criterion



Clustering approaches

- Partitional algorithms: Construct various partitions and then evaluate them by some criterion
- *Hierarchical algorithms*: Create a hierarchical decomposition of the set of objects using some criterion
  - **Bottom-up (**Hierarchical **Agglomerative** Clustering): starting with **isolated points**, merge two items at a time into a new cluster by calculating a dissimilarity between each merged pair and the other samples.
  - Top-down (Hierarchical Divisive Clustering): Starting with **a single set**, split into two distinct parts according to some degree of similarity.

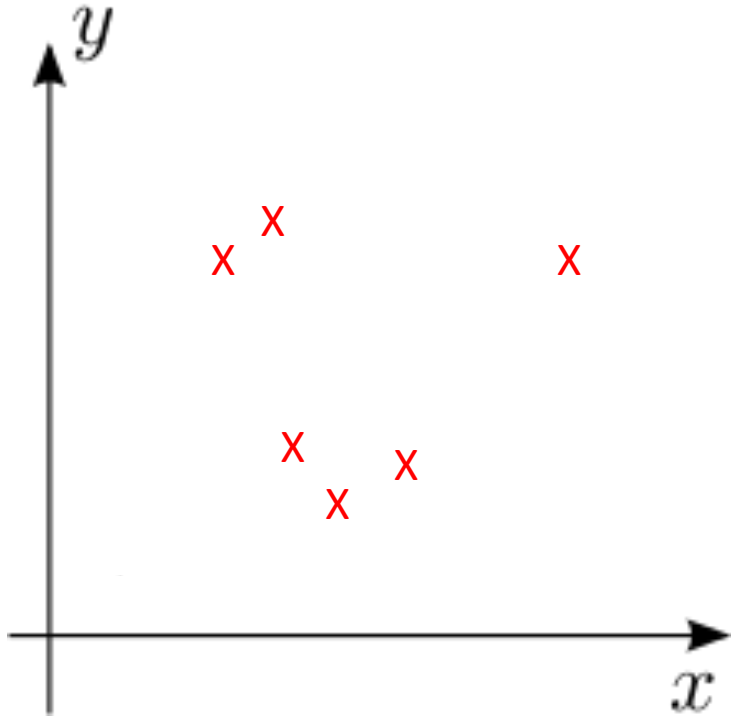
Clustering approaches

- *Partitional algorithms*: Construct various partitions and then evaluate them by some criterion

Activity: you already know how to cluster

- Given a scatter plot and a guess for k, identify which points should be associated with k groups

Need a volunteer who is not in Data 602
and who has not previously heard of k-means



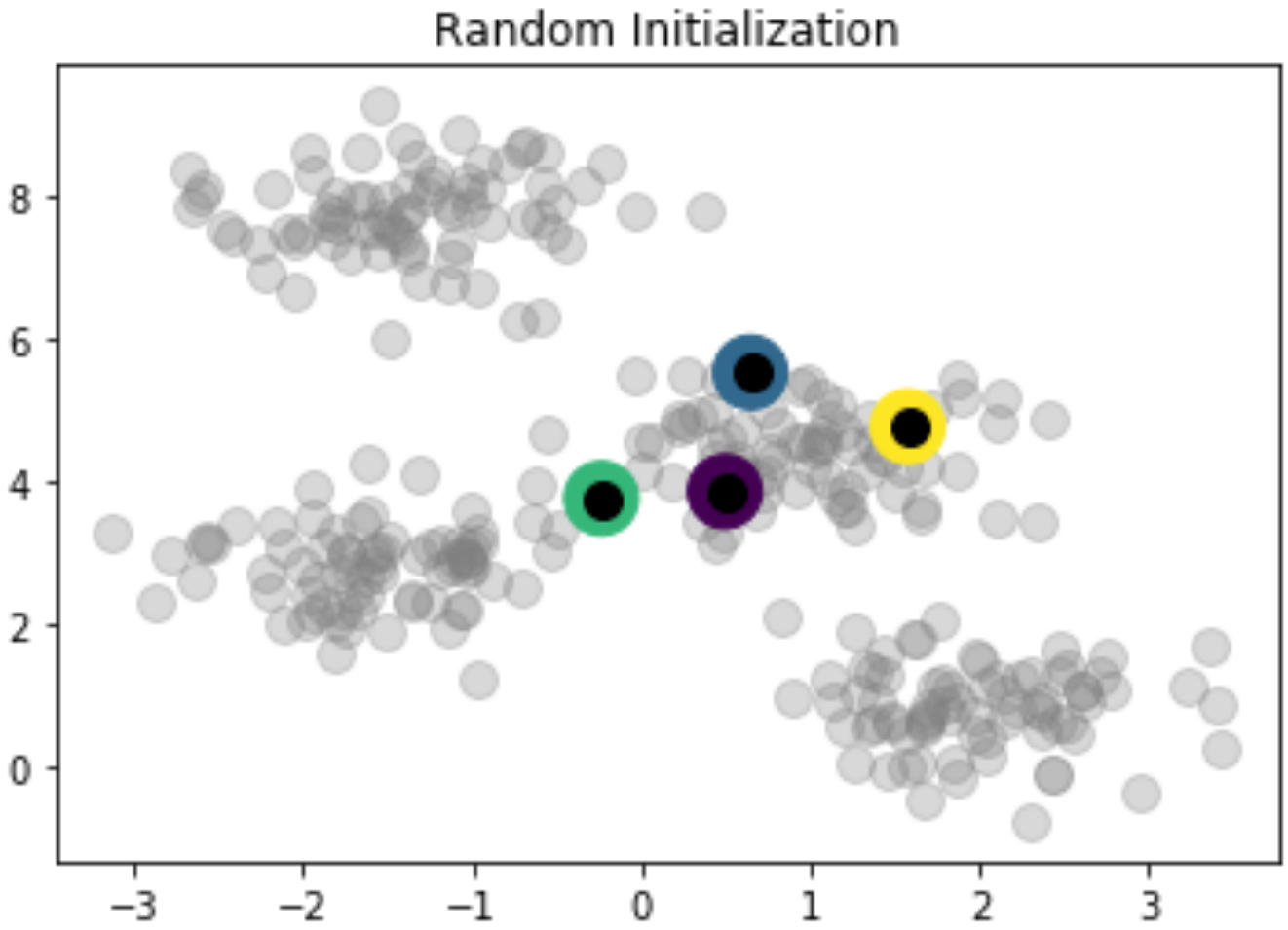
A formalized approach: K-means algorithm

1. Guess number of clusters, *k*
2. Guess location of cluster centers
3. Loop following until converged
   1. *Expectation step*: assign points to the nearest cluster center using a distance measurement
   2. *Maximization step*: set the cluster centers to the mean

Example distance measurement:

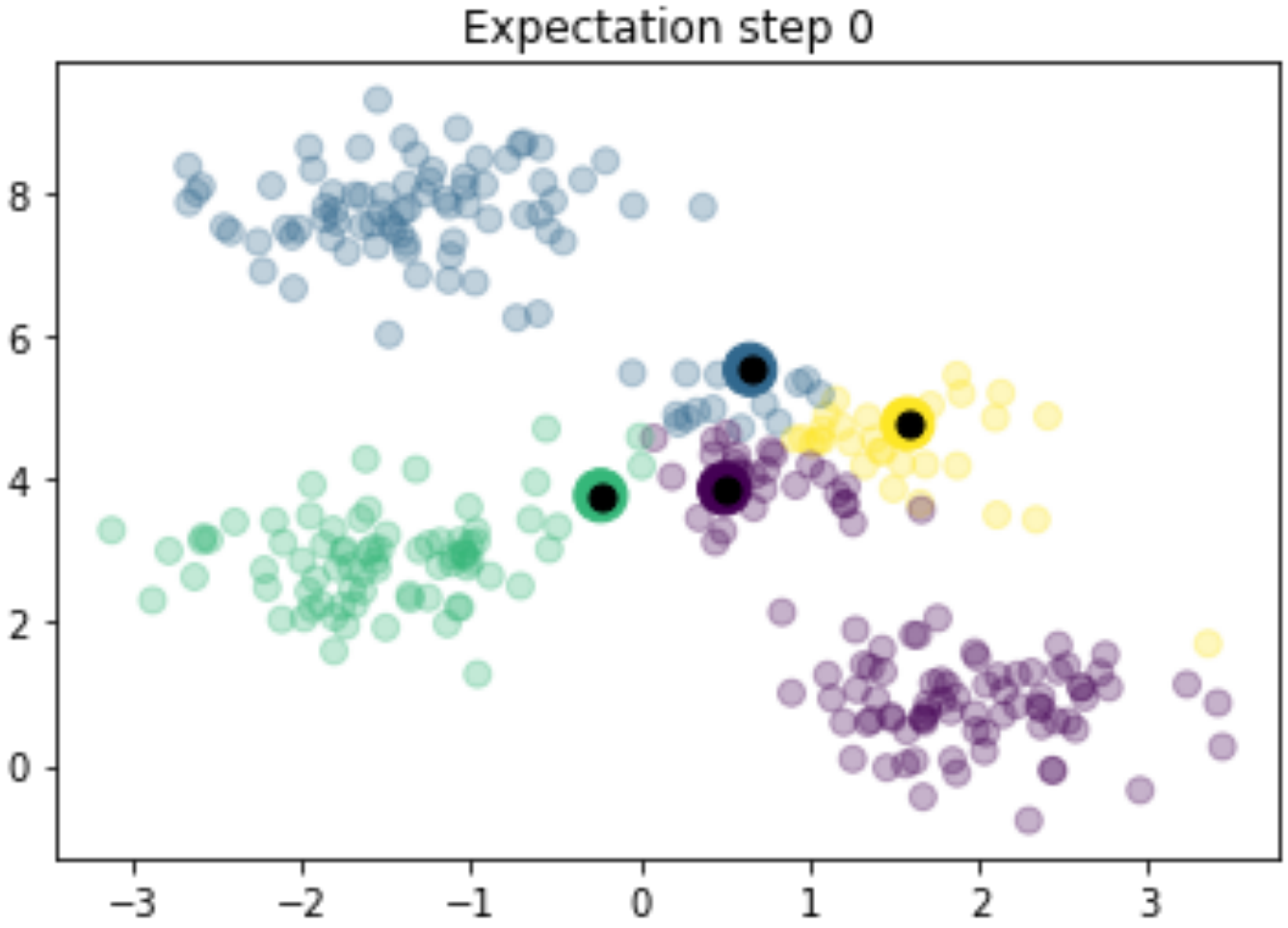
- Residual Sum of Squares: minimize sum of squared distances of each point to its centroid

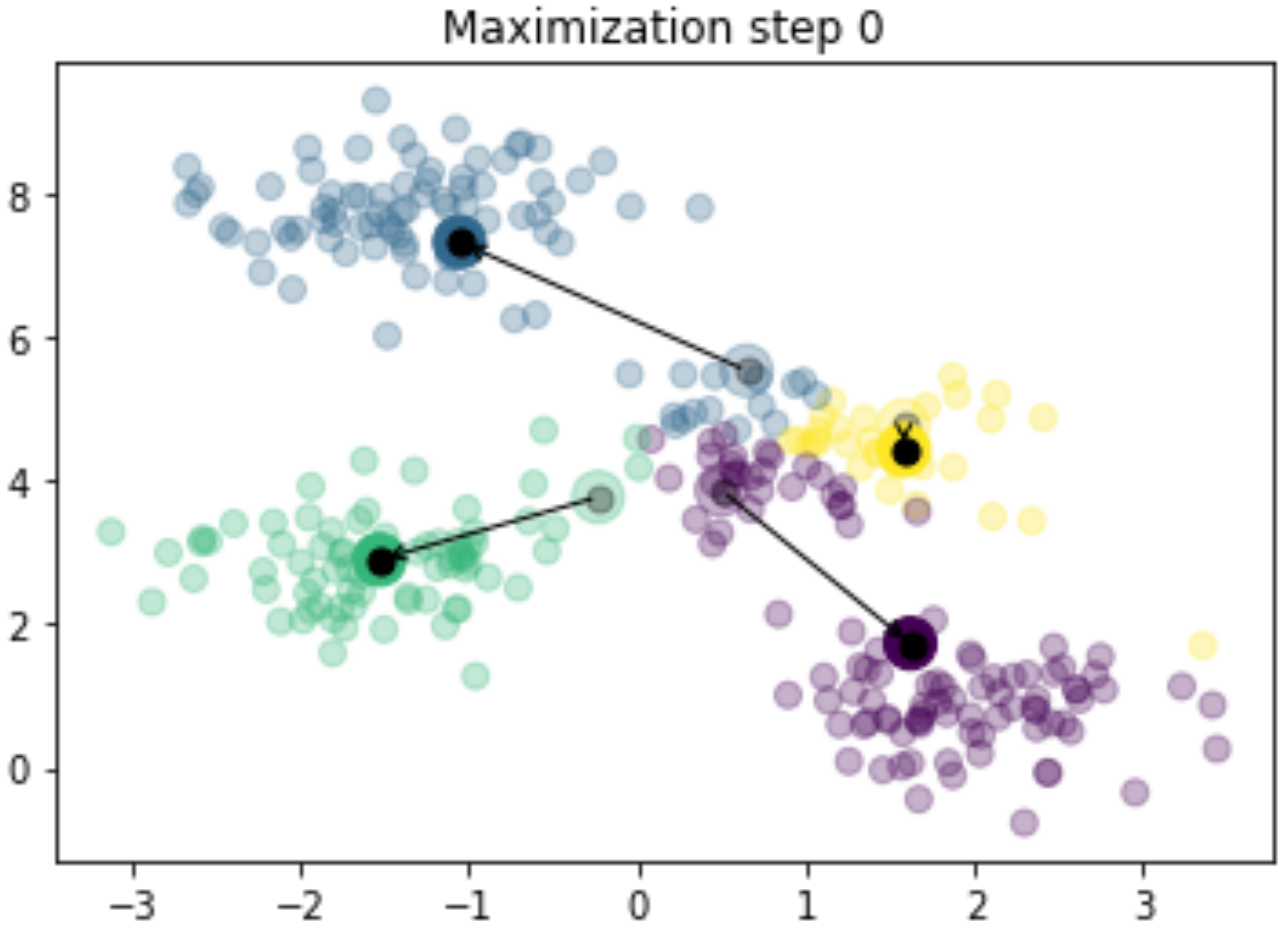
Demo: `jakevdp_kmeans_iterations.ipynb`



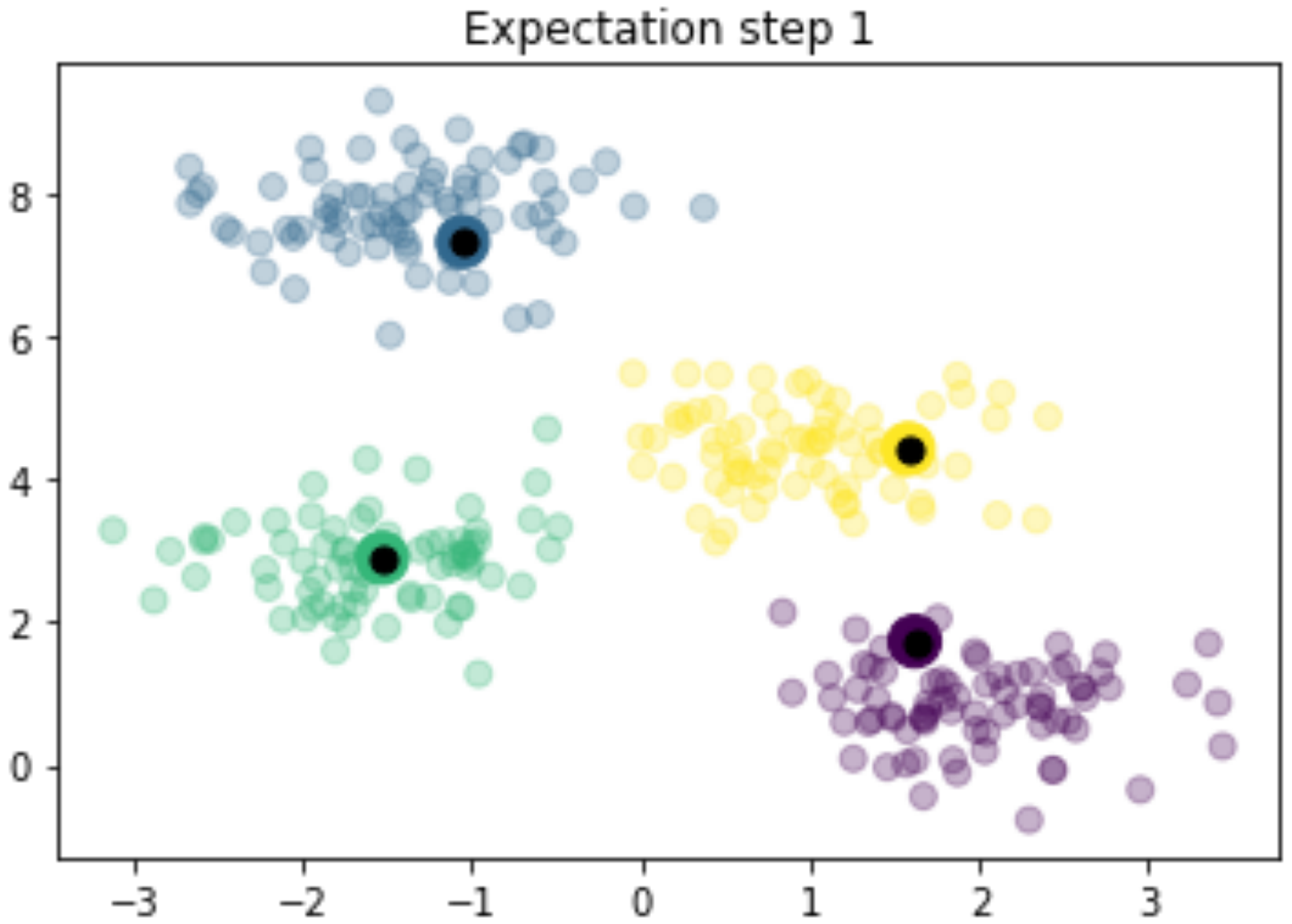

```
[[ 0.49671415  3.8617357 ]
 [ 0.64768854  5.52302986]
 [-0.23415337  3.76586304]
 [ 1.57921282  4.76743473]]
```

Identify which points are associated with centroids

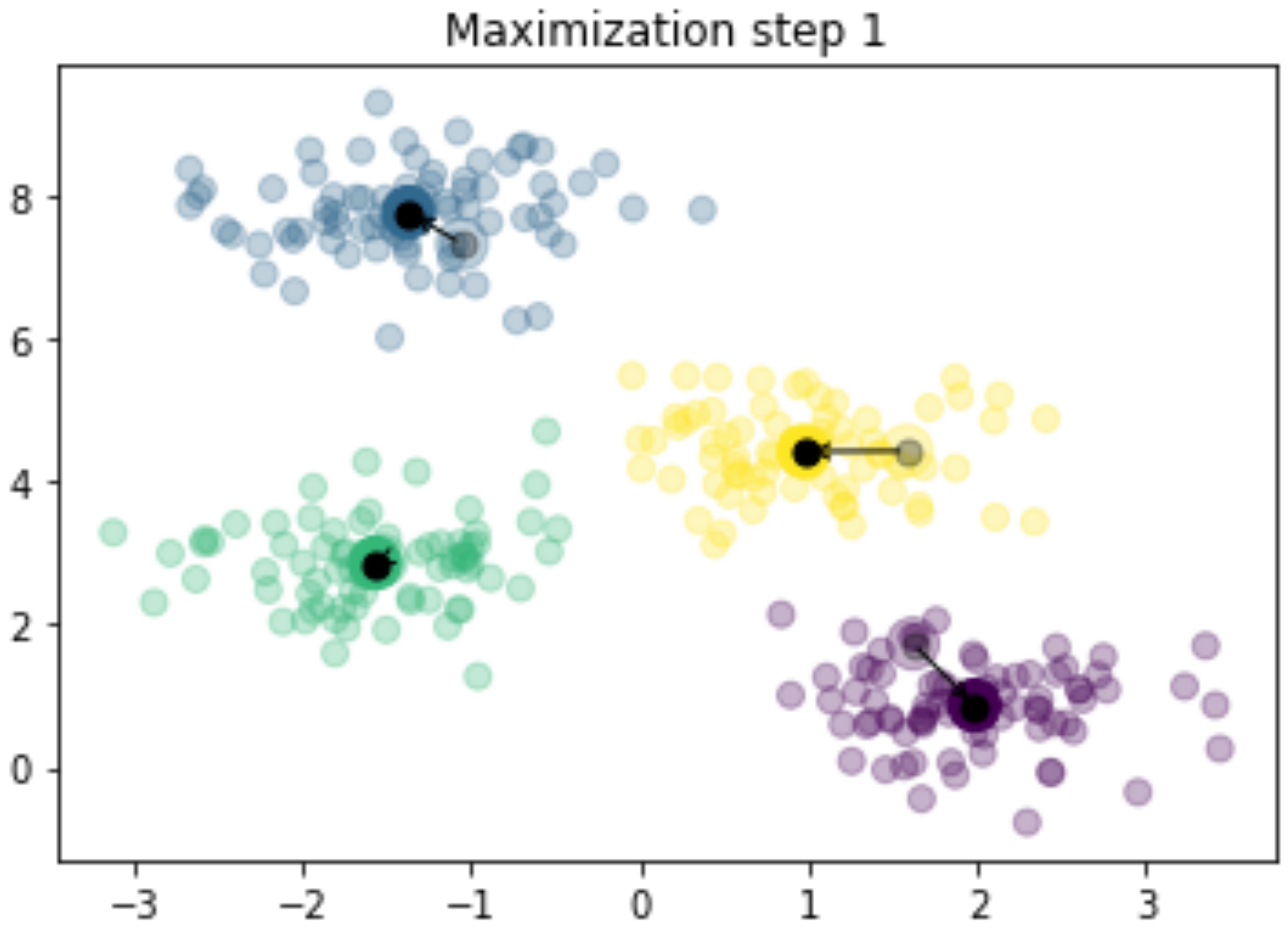




[[1.61788404 1.73021643]
 [-1.05544832 7.31289   ]
 [-1.53049092 2.89441674]
 [1.58028607 4.4210127]]

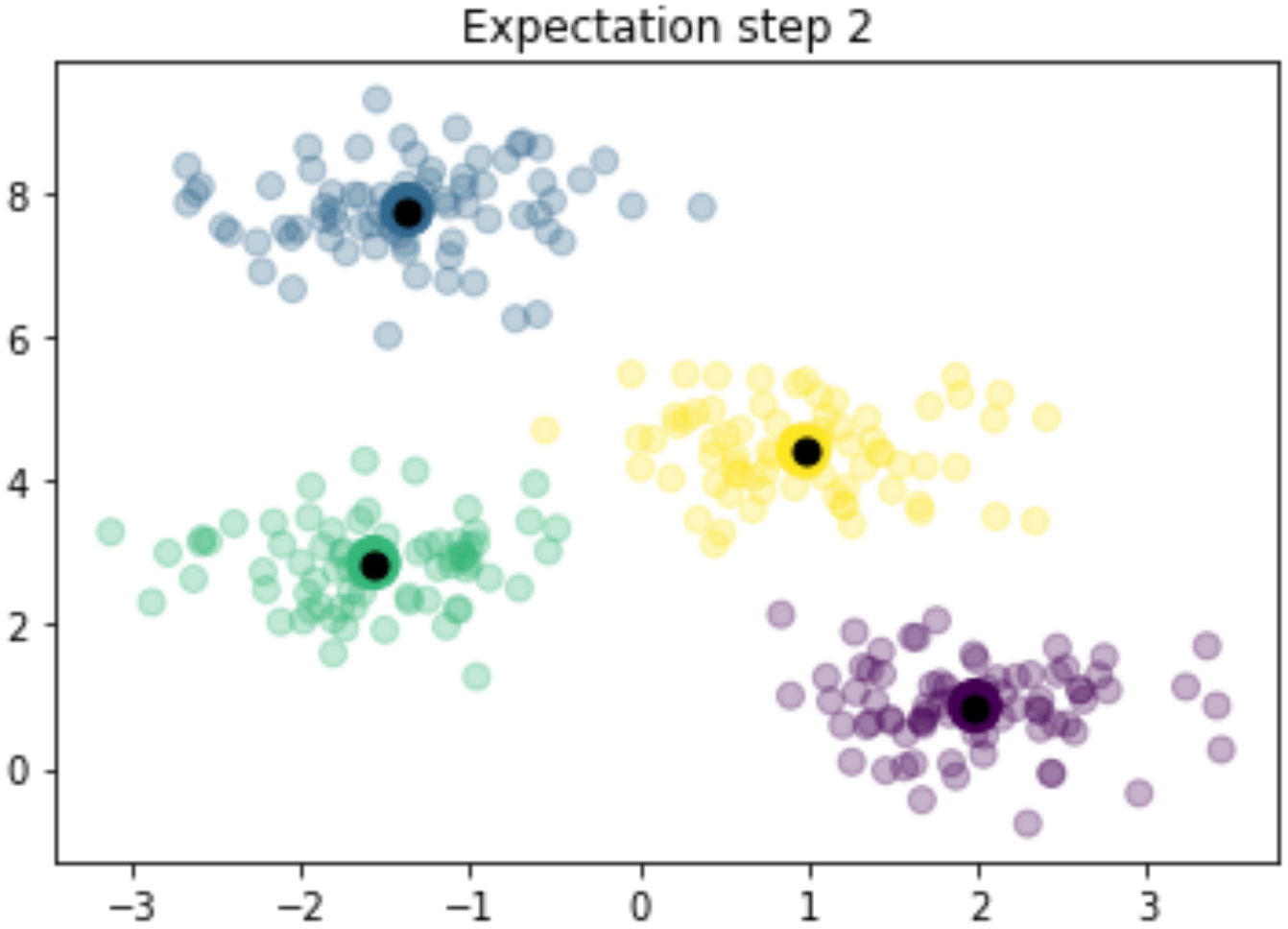


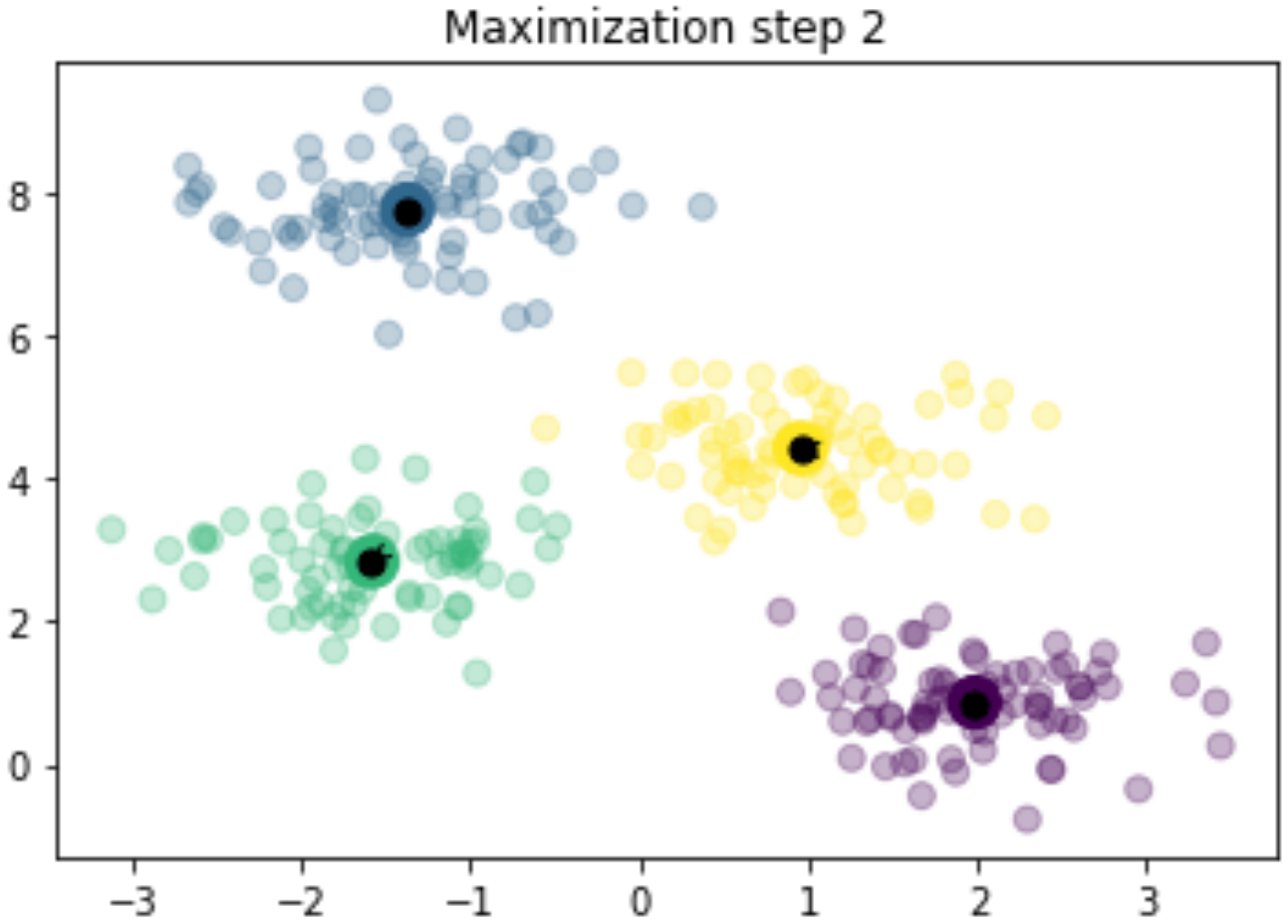
Identify which points are associated with centroids



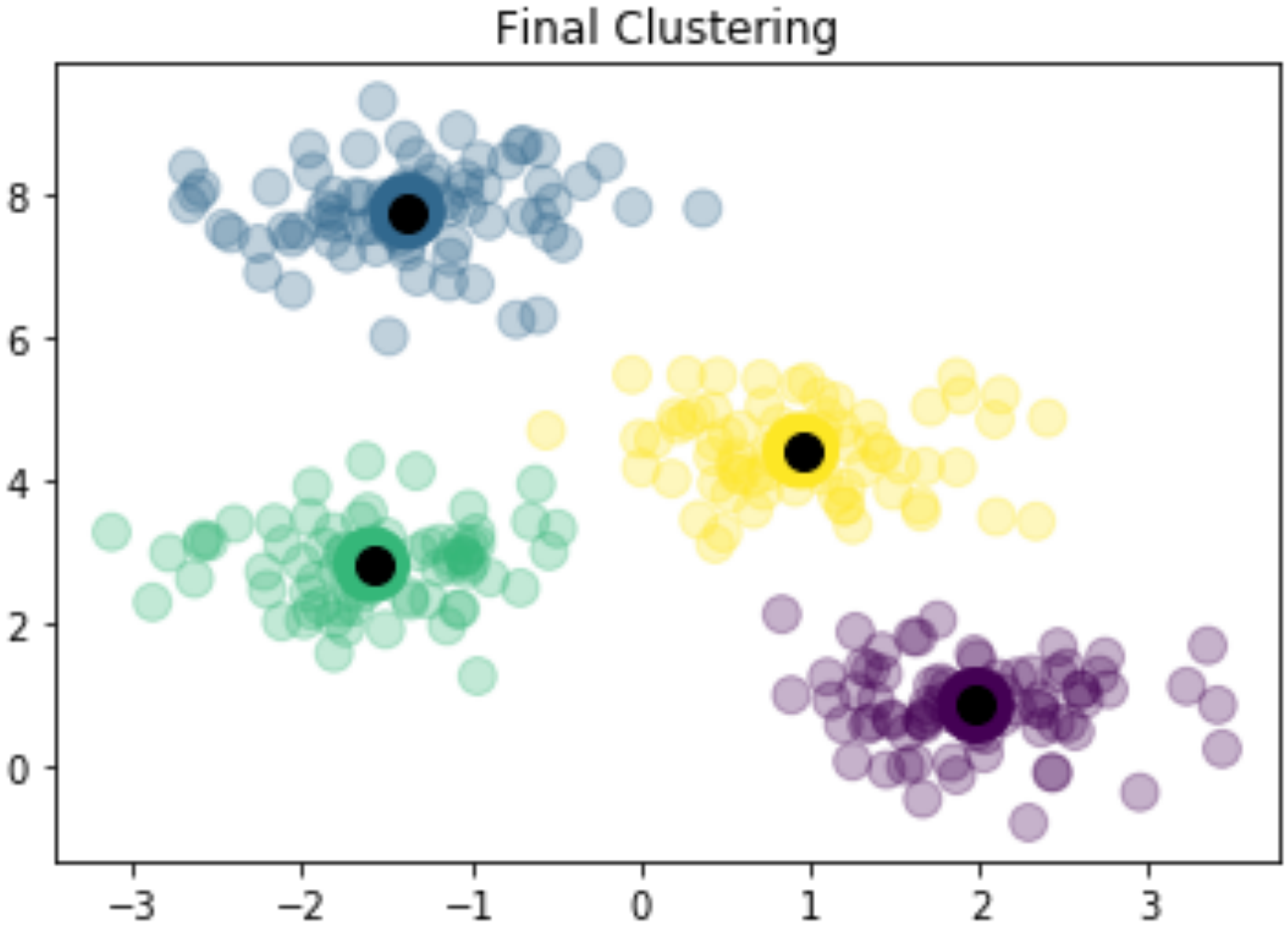
[[1.98258281 0.86771314]
[-1.37324398 7.75368871]
[-1.57084703 2.85535402]
[0.97007666 4.41532732]]

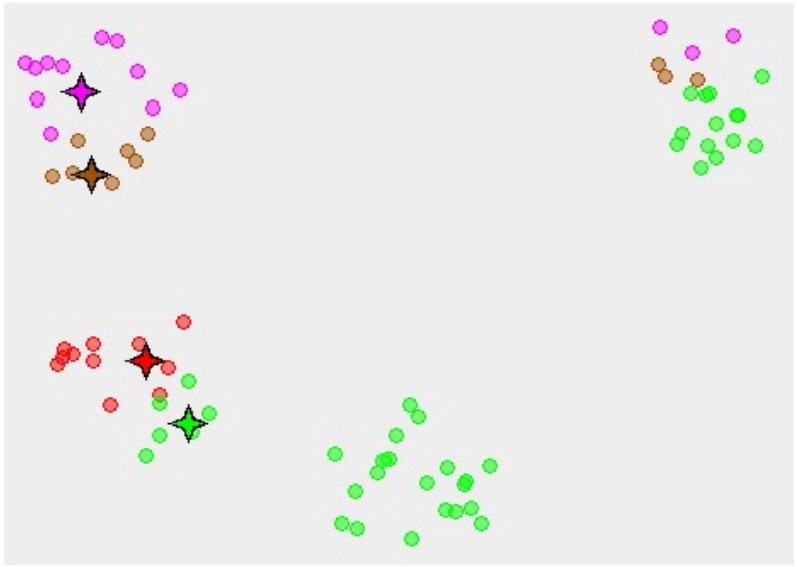
Identify which points are associated with centroids



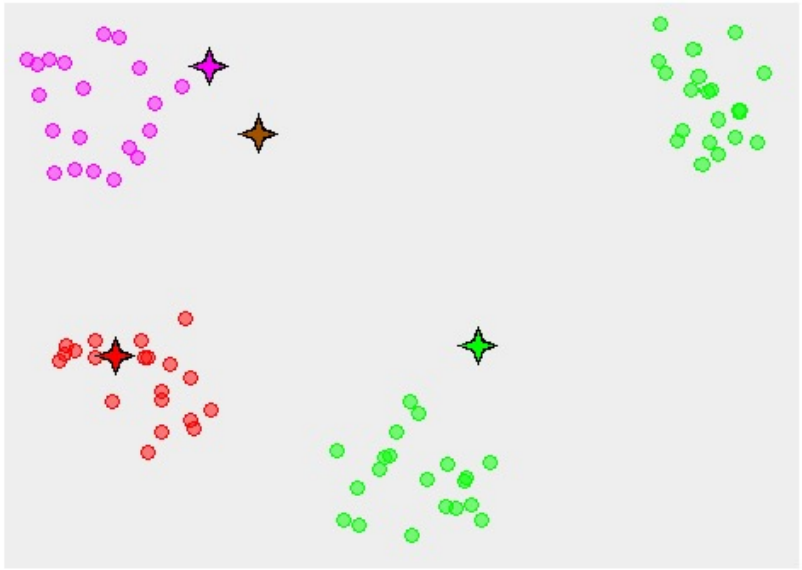


[[1.98258281 0.86771314]
[-1.37324398 7.75368871]
[-1.58438467 2.83081263]
[0.94973532 4.41906906]]

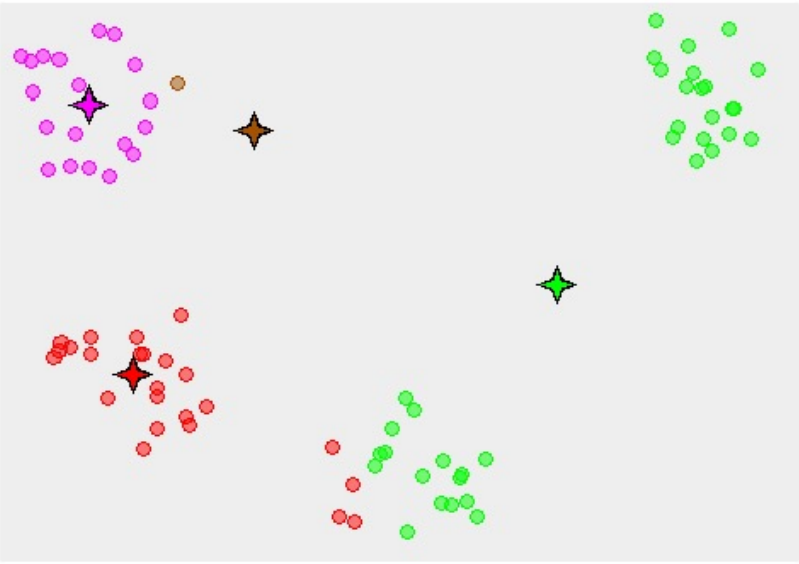

Final Clustering
8
6
4
2
0
−3
−2
−1
0
1
2
3



Guess #1



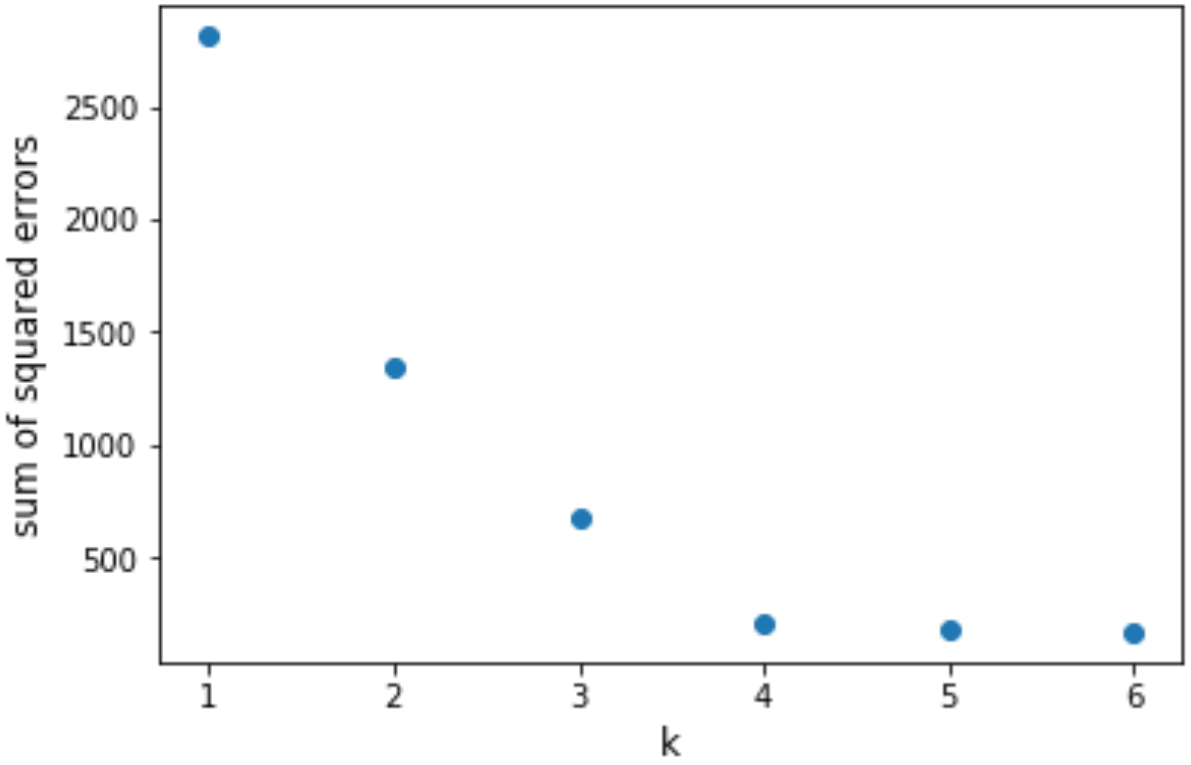
Guess #2



Guess #3

What *k* is best? Find the elbow

- Compare within-cluster sum of squares for multiple *k* values



Caveat: this is a visual estimation method; fancier approaches are available

Demo: `jakevdp_kmeans_iterations_elbow.ipynb`

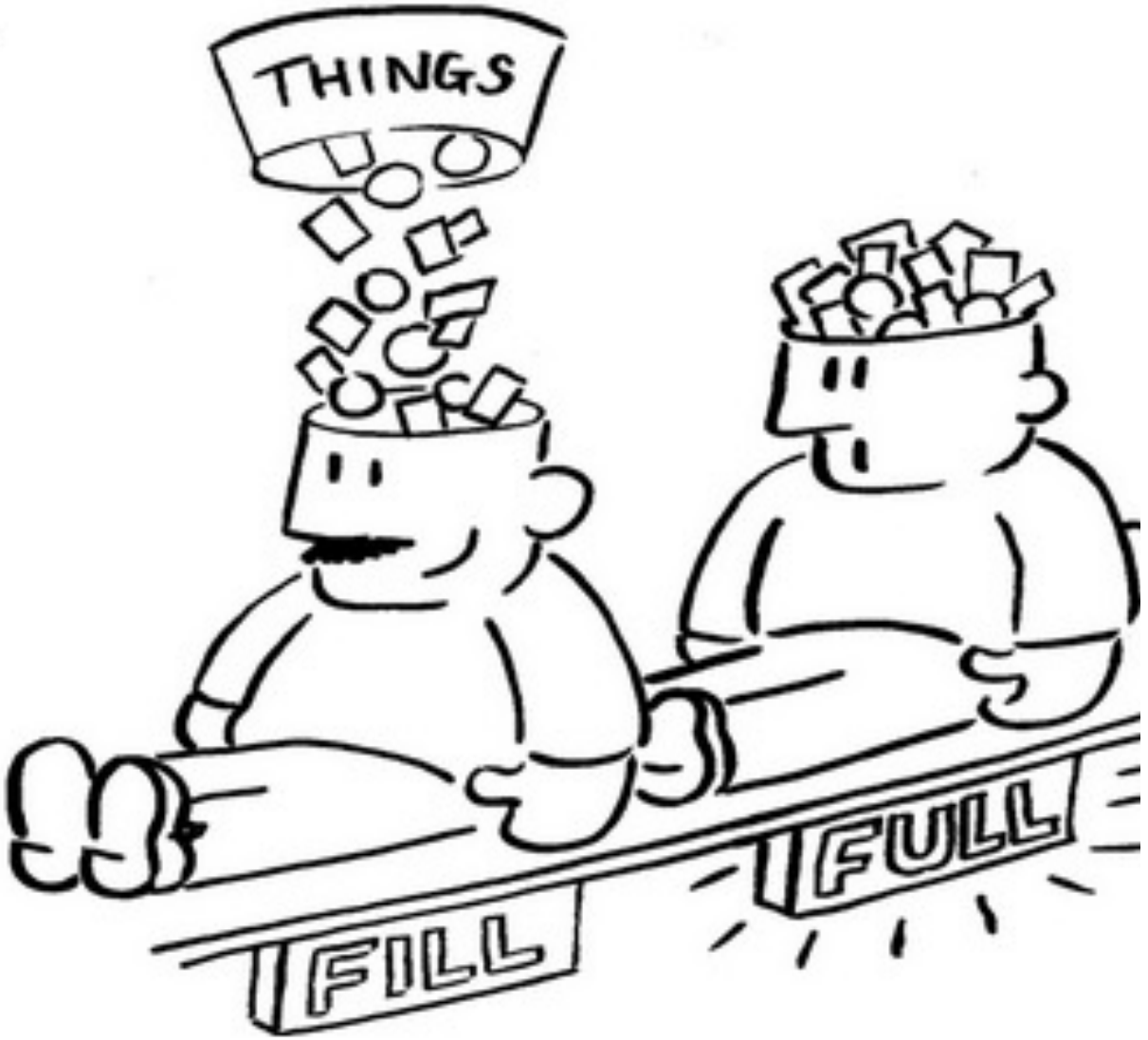
Given *k* clusters, how to identify labels?

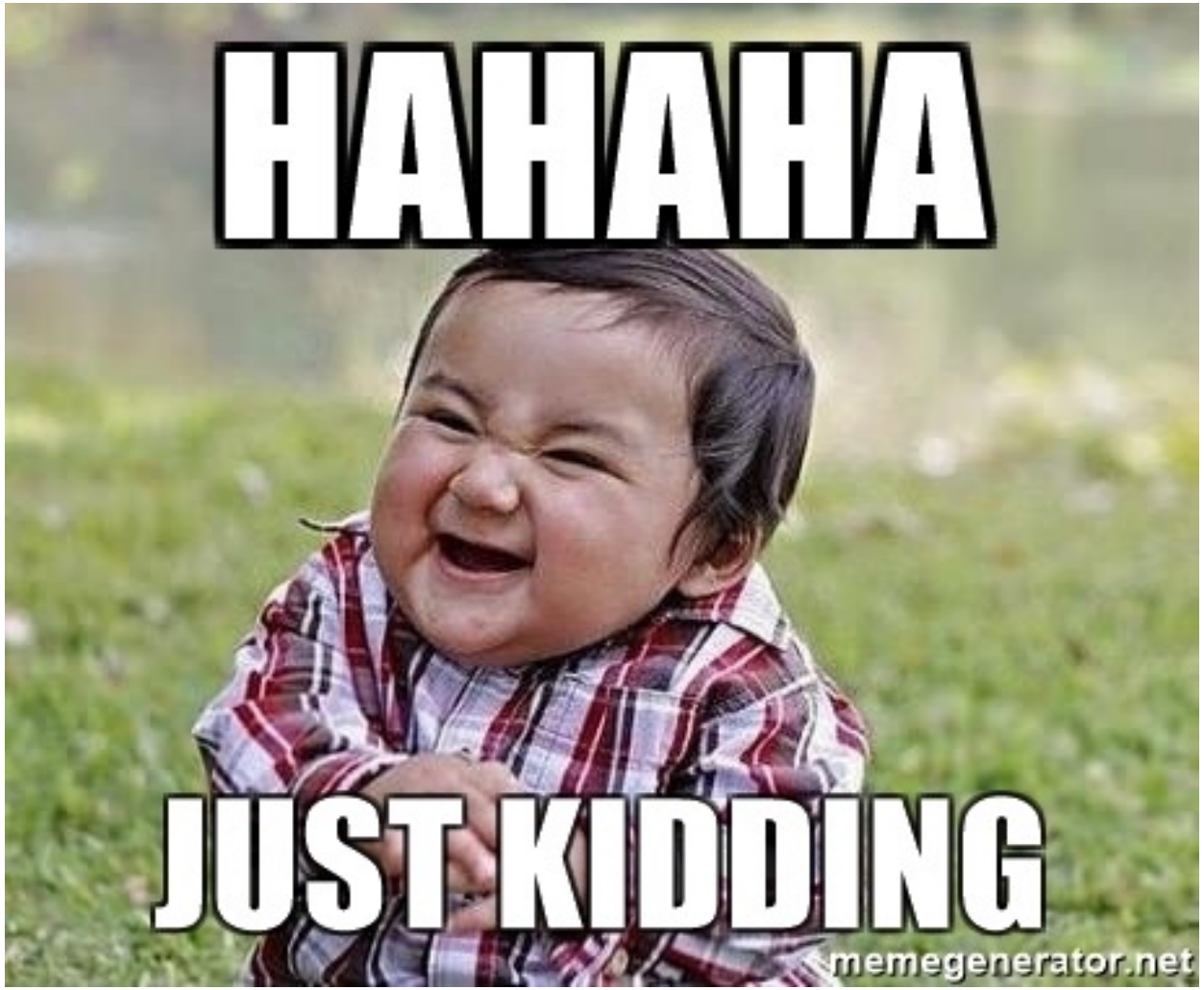
Options:

- summarize the distribution of attribute values
- learn a decision tree from the clusters; then the relevant attributes are identified by the decision tree model

Figuring out labels for data is challenging

All done




HAHAHA
JUST KIDDING
memegenerator.net

What if the data isn't numeric?

Suppose you have 1000 documents

- Examples: web pages, PDFs, Word documents
- Don't have time to manually read each

- Can documents be clustered? By what attributes?
  - Metrics like word count, average word length probably aren't useful

- ~~Clustering context~~
- ~~Clustering methods~~
- Text Analysis
- Finding patterns in Text
- Cleaning text
- Clustering Documents
- Homework

Written text is unstructured data

- structured data typically has defined data types and is organized into patterns, making search and analysis easy
- Unstructured data lacks data types and patterns

- (There's also semi-structured data, ie JSON and XML.)

But what about grammatical rules? Spelling?

- Natural language (ie English, Spanish, Hindi, Mandarian) each have rules for construction.
- Rules have exceptions.

- Spelling has variations.
- Spelling can be incorrect.

Mapping human communication to machine understandable content is the domain of Natural Language Processing

A quick example

- *Claim*: sentences end with a period.

--> to split a document into sentences, simply split the string on periods.

Example:

"This is a fun description of nothing. Another example of sound is provided by my radio. The rug is red."

How many sentences are present in the example?

A quick example of the complexity

- *Claim*: sentences end with a period.

--> to split a document into sentences, simply split the string on periods.

Another example:

"My doctor is Dr. Stark who knows stuff, e.g., medicine, is smart. I apply the K.I.S.S. principle: Keep It Simple, Stupid."

The naïve approach isn't sufficient.


Natural Language Processing:
Text, Audio, Images

- ~~Clustering context~~
- ~~Clustering methods~~
- ~~Text Analysis~~
- Finding patterns in Text
- Cleaning text
- Clustering Documents
- Homework

Activity: Find a string in a text document

- Go to https://library.umbc.edu/
- Search the page using "ctrl + f"
- Search for the string "loan"

Find a string in a text document

- Go to https://library.umbc.edu/
- Search the page using "ctrl + f"
- Search for the string "loan"

(rhetorical questions)

- What if we needed to find every email address in a document?
- What if we needed to find every phone number in a document?

Patterns in text

We are familiar with email addresses: aok@umbc.edu

Phone numbers look something like 410-455-2232

Describing Patterns in text

We are familiar with email addresses: aok@umbc.edu

- A sequence of letters or numbers, then @, then more letters and numbers, a period, and two or more letters.

Phone numbers look something like 410-455-2232

- Three numbers, a space or dash, three numbers, a space or dash, then four more numbers.

Regular Expressions

Symbolic way to represent digits: \d

Then phone number: \d\d\d-\d\d\d-\d\d\d\d\

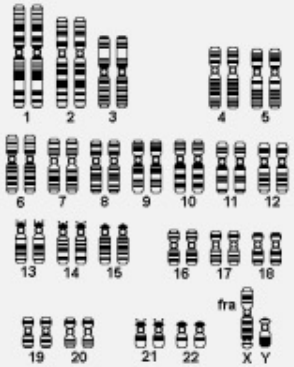
- Another way to represent digits: [0-9]

Demo: `introduction_to_regular_expressions.ipynb`

RegEx apply where ever text exists

Ex. [genomics]

- Fragile X syndrome is a common cause of mental retardation.
- A human's genome is a string.
- It contains triplet repeats of CGG or AGG, bracketed by GCG at the beginning and CTG at the end.
- Number of repeats is variable and is correlated to syndrome.



pattern `GCG(CGG|AGG)*CTG`

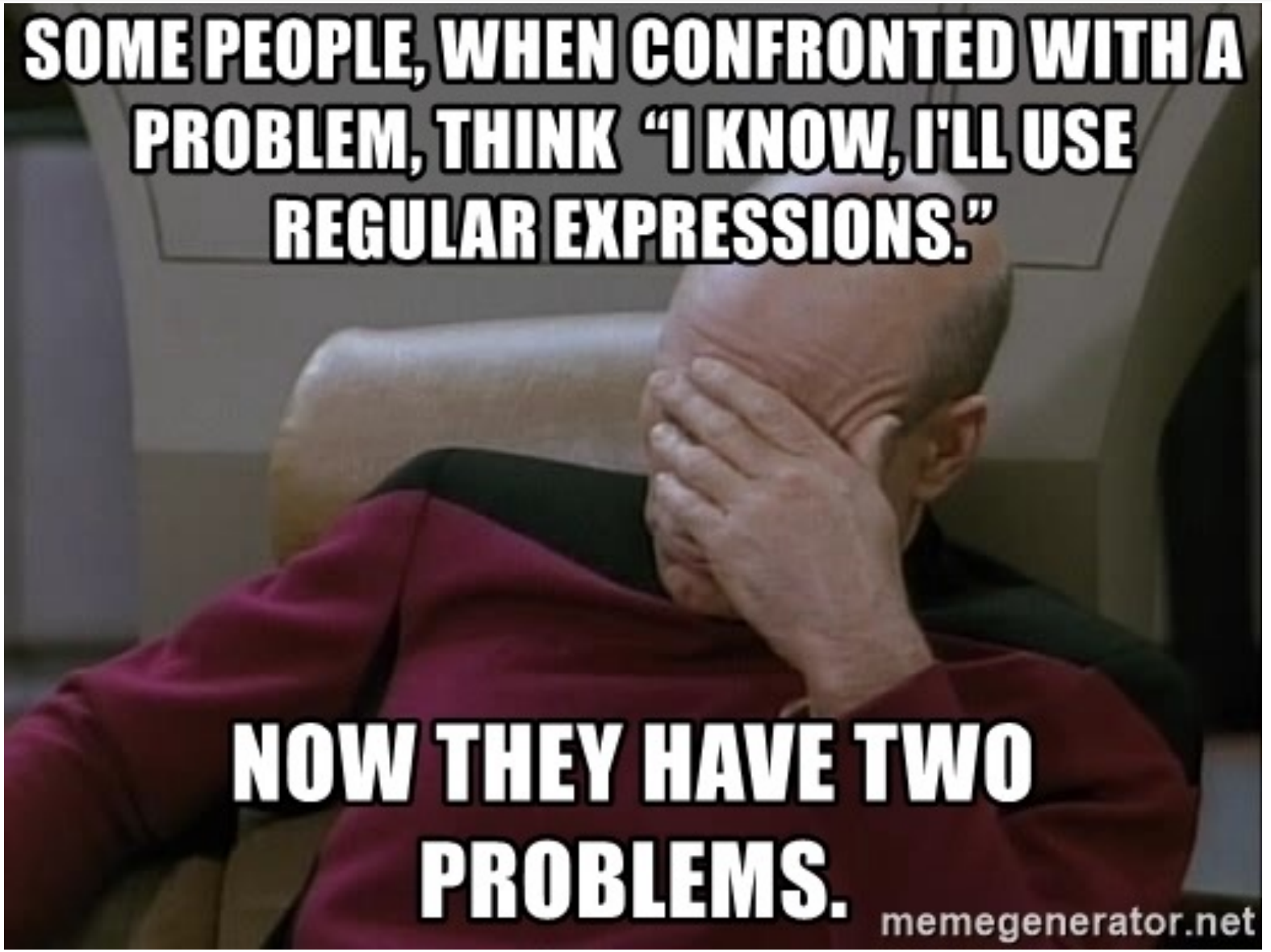
text `GCGGCGTGTGTGCGAGAGAGTGGGTTTAAAGCTGGCGCGGAGGCGGCTGGCGCGGAGGCTG`

Regular Expression references

- https://regexper.com/
- https://regexr.com/
- https://github.com/ziishaned/learn-regex
- https://qntm.org/files/re/re.html
- https://www.regexpal.com/
- https://www.debuggex.com/
- https://www.regular-expressions.info/
- https://stackoverflow.com/tags/regex/info
- https://regex101.com/

Cheatsheets:

- https://ryanstutorials.net/regular-expressions-tutorial/regular-expressions-cheat-sheet.php
- http://www.cbs.dtu.dk/courses/27610/regular-expressions-cheat-sheet-v2.pdf
- https://www.cheatography.com/davechild/cheat-sheets/regular-expressions/



https://en.wikiquote.org/wiki/Jamie_Zawinski

When RegEx is insufficient, use a parser

RegEx applies to Regular Languages

HTML is not a Regular Language

See also

- https://stackoverflow.com/questions/6751105/why-its-not-possible-to-use-regex-to-parse-html-xml-a-formal-explanation-in-la
- https://blog.codinghorror.com/parsing-beyond-regex/
- https://stackoverflow.com/questions/11905506/regular-expression-vs-string-parsing
- https://blog.honeybadger.io/replacing-regular-expressions-with-parsers/

Activity: Agglomerative and Divisive Clustering

Divisive clustering of students by number of letters in first name:

- Starting with **a single group**, split into two distinct parts according to length of name

End state: a set of groups of students, where members of each group have the same number of letters in their first name

- ~~Clustering context~~
- ~~Clustering methods~~
- ~~Text Analysis~~
- ~~Finding patterns in Text~~
- Cleaning text
- Clustering Documents
- Homework

Now that we know how to find complex strings in text,

How would we identify which words are relevant in a document?

Activity: which words in this text are irrelevant?

I go to the store. A car is parked. Many cars are parked or moving. Some are blue. Some are tan. They have windows. In the store, there are items for sale. These include such things as soap, detergent, magazines, and lettuce. You can enhance your life with these products. Soap can be used for bathing, be it in a bathtub or in a shower. Apply the soap to your body and rinse. Detergent is used to wash clothes. Place your dirty clothes into a washing machine and add some detergent as directed on the box. Select the appropriate settings on your washing machine and you should be ready to begin. Magazines are stapled reading material made with glossy paper, and they cover a wide variety of topics, ranging from news and politics to business and stock market information. Some magazines are concerned with more recreational topics, like sports card collecting or different kinds of hairstyles. Lettuce is a vegetable. It is usually green and leafy, and is the main ingredient of salads. You may have an appliance at home that can quickly shred lettuce for use in salads. Lettuce is also used as an optional item for hamburgers and deli sandwiches. Some people even eat lettuce by itself. I have not done this. So you can purchase many types of things at stores.

Activity: which words in this text are irrelevant?

I go to the store. A car is parked. Many cars are parked or moving. Some are blue. Some are tan. They have windows. In the store, there are items for sale. These include such things as soap, detergent, magazines, and lettuce. You can enhance your life with these products. Soap can be used for bathing, be it in a bathtub or in a shower. Apply the soap to your body and rinse. Detergent is used to wash clothes. Place your dirty clothes into a washing machine and add some detergent as directed on the box. Select the appropriate settings on your washing machine and you should be ready to begin. Magazines are stapled reading material made with glossy paper, and they cover a wide variety of topics, ranging from news and politics to business and stock market information. Some magazines are concerned with more recreational topics, like sports card collecting or different kinds of hairstyles. Lettuce is a vegetable. It is usually green and leafy, and is the main ingredient of salads. You may have an appliance at home that can quickly shred lettuce for use in salads. Lettuce is also used as an optional item for hamburgers and deli sandwiches. Some people even eat lettuce by itself. I have not done this. So you can purchase many types of things at stores.

short words are less useful?

That naïve heuristic has exceptions:

- Cat
- Dog
- Gun
- Ben

Demo: `nltk_for_text_processing.ipynb`

Stop words

There are many lists of stop words available, i.e.

- https://gist.github.com/sebleier/554280
- http://xpo6.com/list-of-english-stop-words/

the specific list of stop words isn't priority

We can clean documents! Now what?

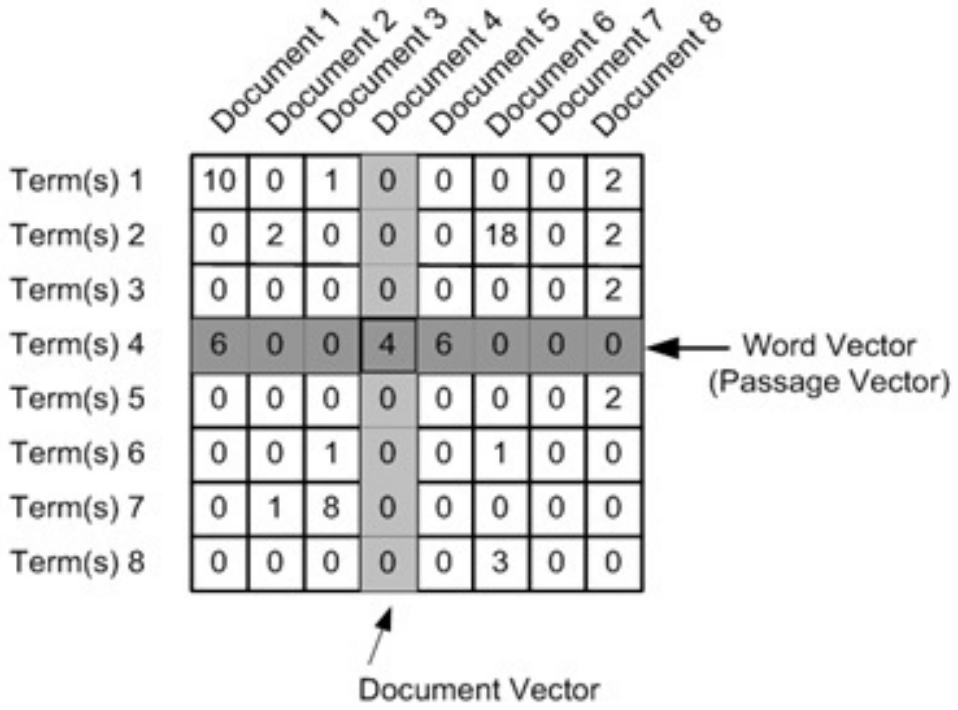
Next up: identifying words that are more unique to documents

- ~~Clustering context~~
- ~~Clustering methods~~
- ~~Text Analysis~~
- ~~Finding patterns in Text~~
- ~~Cleaning text~~
- Clustering Documents
- Homework

Term Frequency-Inverse Document Frequency

TF-IDF makes rare words more prominent and lowers dominance of common words.

document-term matrix



Term Frequency-Inverse Document Frequency

Demo: `tfidf.ipynb`

- ~~Clustering context~~
- ~~Clustering methods~~
- ~~Text Analysis~~
- ~~Finding patterns in Text~~
- ~~Cleaning text~~
- ~~Clustering Documents~~
- Homework

Initial state for text analysis is raw text

- Emails
- Documents (PDF, TXT, DOCX)
- Spreadsheets
- Webpages

https://en.wikipedia.org/wiki/List_of_text_corpora

Remove stop words from txt and docx files

I provide a .zip containing .txt and .docx files

For each file, remove punctuation and stop words

Produce a single .dat file containing the name of the file in quotes, a colon, then a list of words separated by commas. The list of words per file should be unique. Do not include URLs or phone numbers.

```
"File 1.txt" : word1, word2, word3, word7
"name of file.docx" : word8, word2, word1, word10
"another file.doc" : word1, word12, word6
```

Reading assignment

Read Sections 1,2,3,6 of
http://opim.wharton.upenn.edu/~sok/idtresources/python/regex.pdf

Summarize what you learned in a half page (more than 100 words, less than 270).

Do not include your name.

Submit your text to me via Blackboard.

Submit the content as text.

Do not submit a docx or PDF or an image.

References

- See chapter 4 in https://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf page 127
- https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
- https://mubaris.com/2017/10/01/kmeans-clustering-in-python/
- http://brandonrose.org/clustering
- https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html

- LDA using GenSim: https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24