

Foundations of Data Science

DS 3001

Data Science Program

Department of Computer Science

Worcester Polytechnic Institute

Instructor: Prof. Kyumin Lee

Upcoming Schedule

- Project Proposal
 - <https://canvas.wpi.edu/courses/18106/assignments/132329>
 - Due date: Today
- HW3 will be out this Friday

Data Science: The Context

Ask question: What data needs to be recorded? or collected?



Real World



Humans behaving
Biology
Finance
Internet
Medicine
Sociology
Olympics



Raw Data is
Collected / Recorded

email
logs
medical records
surveys
blood drawn
(microarray)
olympic records
NYT web pages



Data is
Processed

pipelines
web scraping
cleaning
munging
joining
wrangling



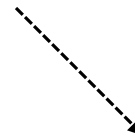
Data Set

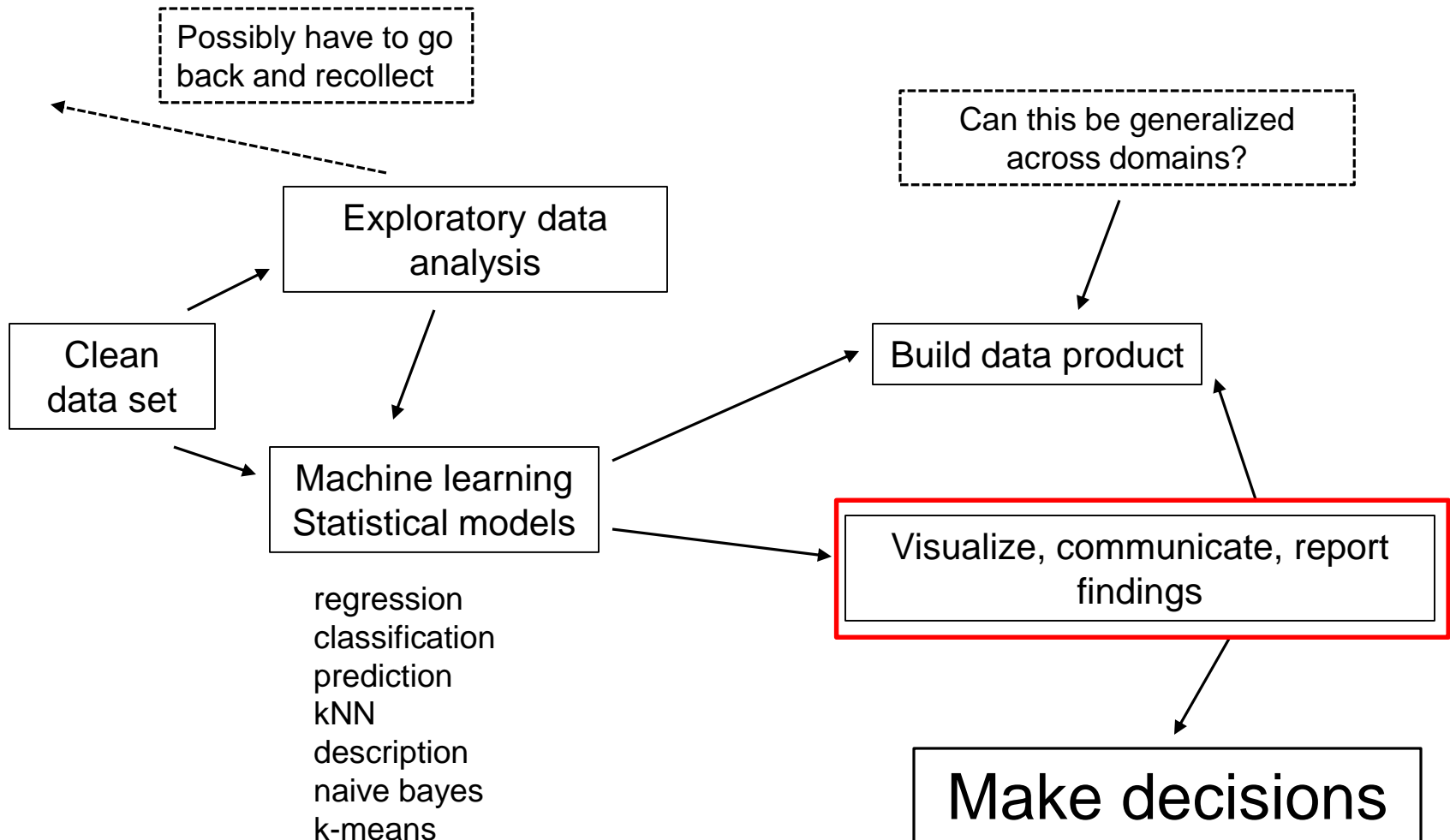
“clean” table

Why? What research question
am I going to answer?



What do I want it to look like?





Data Visualization

The Value of Visualization

- **Record** information
 - Blueprints, photographs, seismographs, ...
- **Analyze** data to support reasoning
 - Develop and assess hypotheses
 - Discover errors in data
 - Expand memory
 - Find patterns
- **Communicate** information to others
 - Share and persuade
 - Collaborate and revise

Tufte: Principles of Graphical Excellence

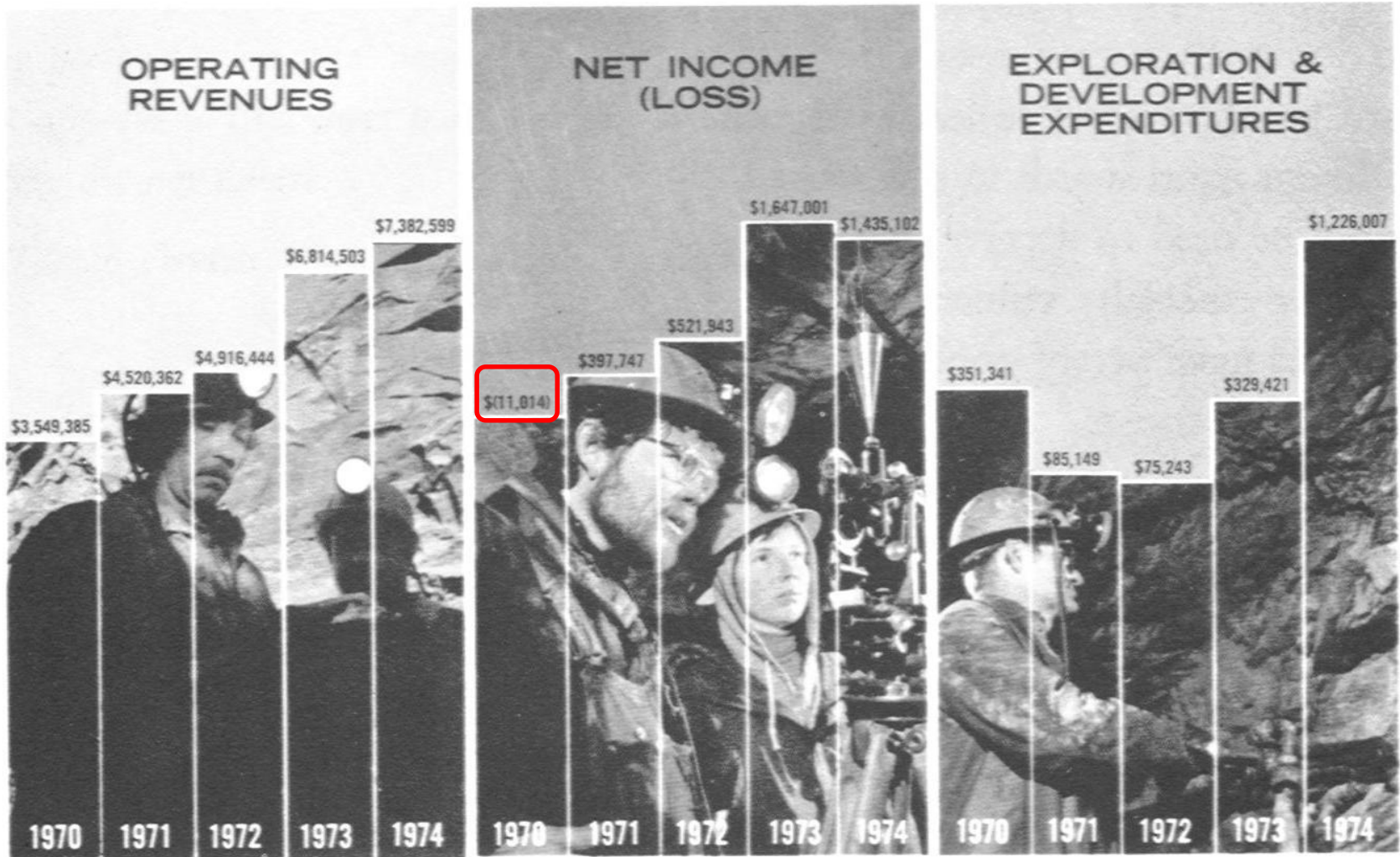
- Graphical excellence is the well-designed presentation of interesting data – a matter of *substance*, *statistics*, and *design*
- Graphical excellence consists of complex ideas communicated with *clarity*, *precision*, and *efficiency*

Tufte: Graphical Integrity

- “not lying with statistics”
- tell the truth about data

Uh oh ...

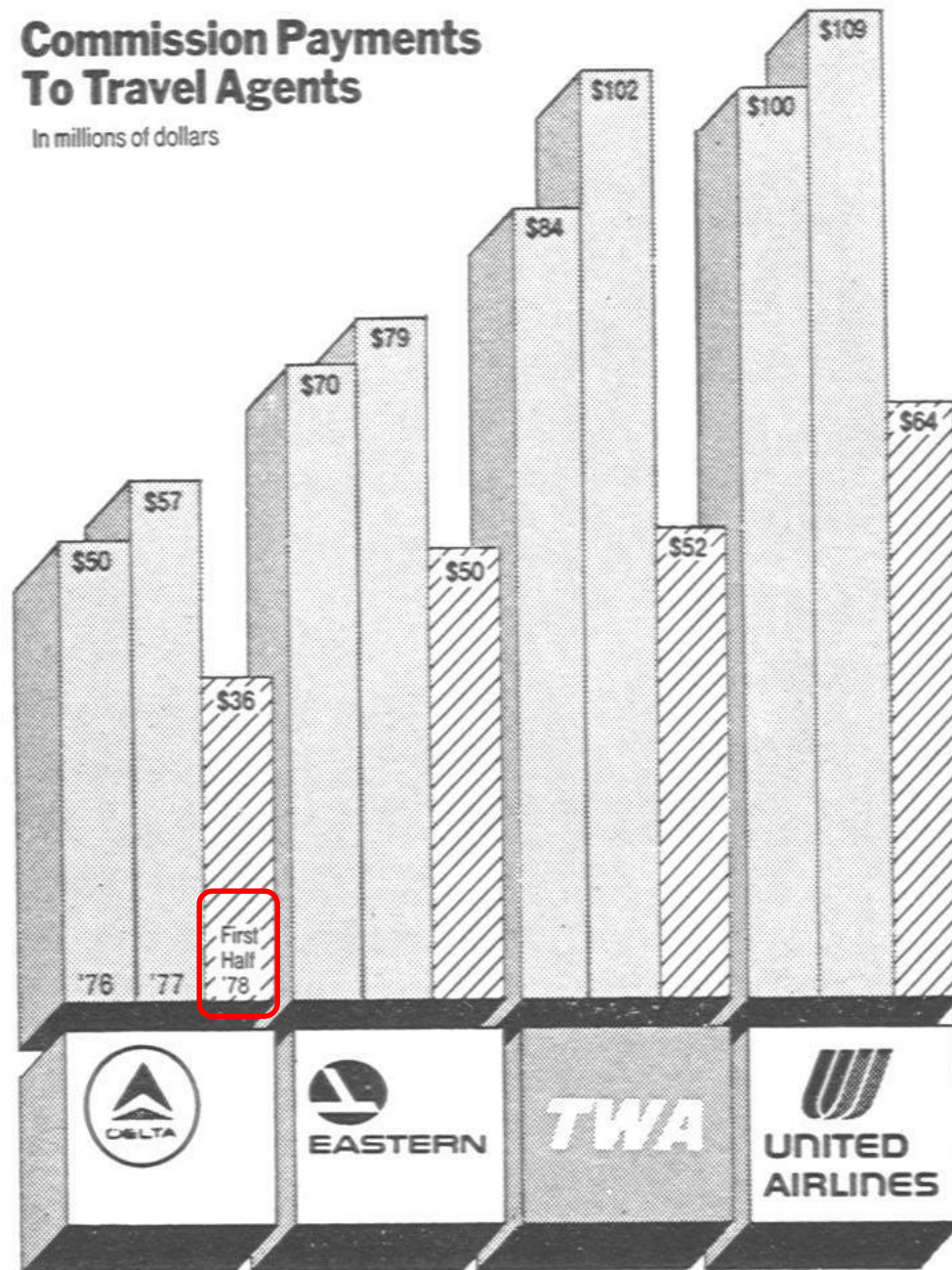
Examples of Infographics lacking integrity



Day Mines, Inc., 1974 Annual Report

Commission Payments To Travel Agents

In millions of dollars



New York Times, 8/8/78

Comparative Annual Cost per Capita for care of Insane in
Pittsburgh City Homes and Pennsylvania State Hospitals.

\$147



South Mountain

\$172



Pittsburgh

\$198



Harrisburg

\$213



Norristown

\$214



Warren

Lie Factor

- Given perceptual difficulties – strive for uniformity (predictability) in graphics (p56)
 - ‘the representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.’
 - ‘Clear, detailed and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.’

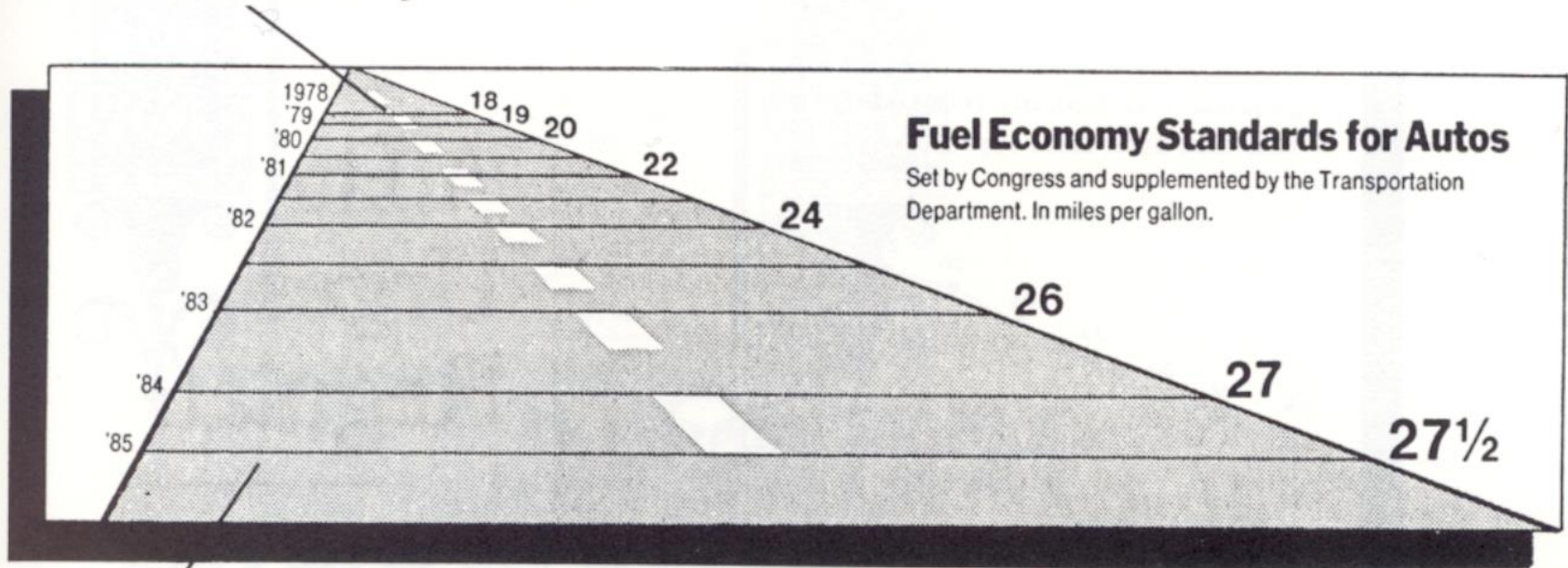
$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

- Lie factor of 1 → is desirable
- Lie factor > 1.05 or < 0.95 go beyond plotting errors

Extreme example

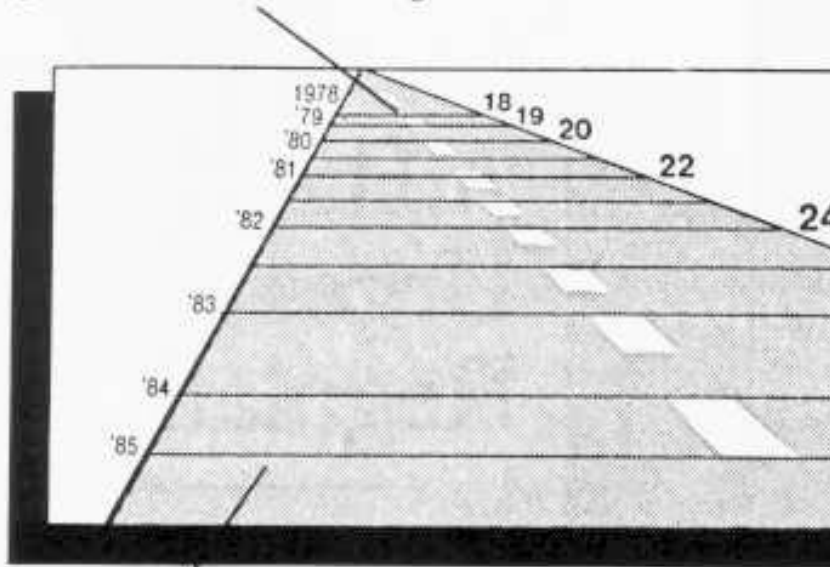
- Fuel economy standards for automobiles
18 miles/gallon in 1978 to 27.5 miles/gallon in 1985
 $(27.5 - 18.0)/(18.0) \times 100 = 53\%$ increase

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

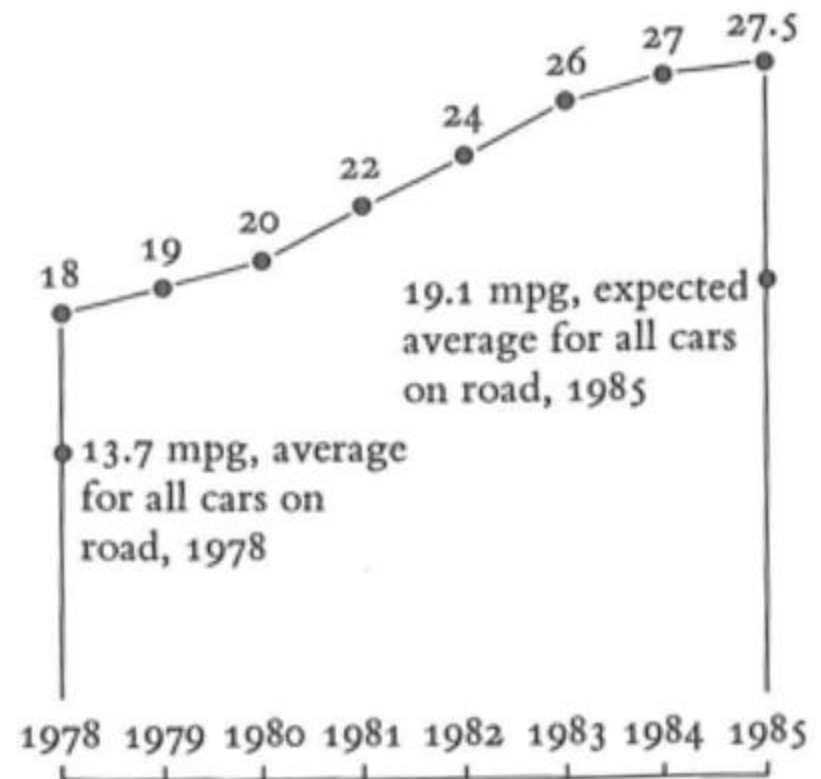


This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Fuel Economy Standards for Autos

Set by Congress and supplemented by the Transportation

REQUIRED FUEL ECONOMY STANDARDS: NEW CARS BUILT FROM 1978 TO 1985



- Graphic increase

$$(5.3 - 0.6) / (0.6) \times 100 = 783\%$$
- Lie Factor = $783 / 53 = 14.8$
- Additional confounding factors
 Usually the future is in front of us
 Dates remain same size and fuel fac

Visual Area and Numerical Measure

Use of area to portray 1D data can be confusing

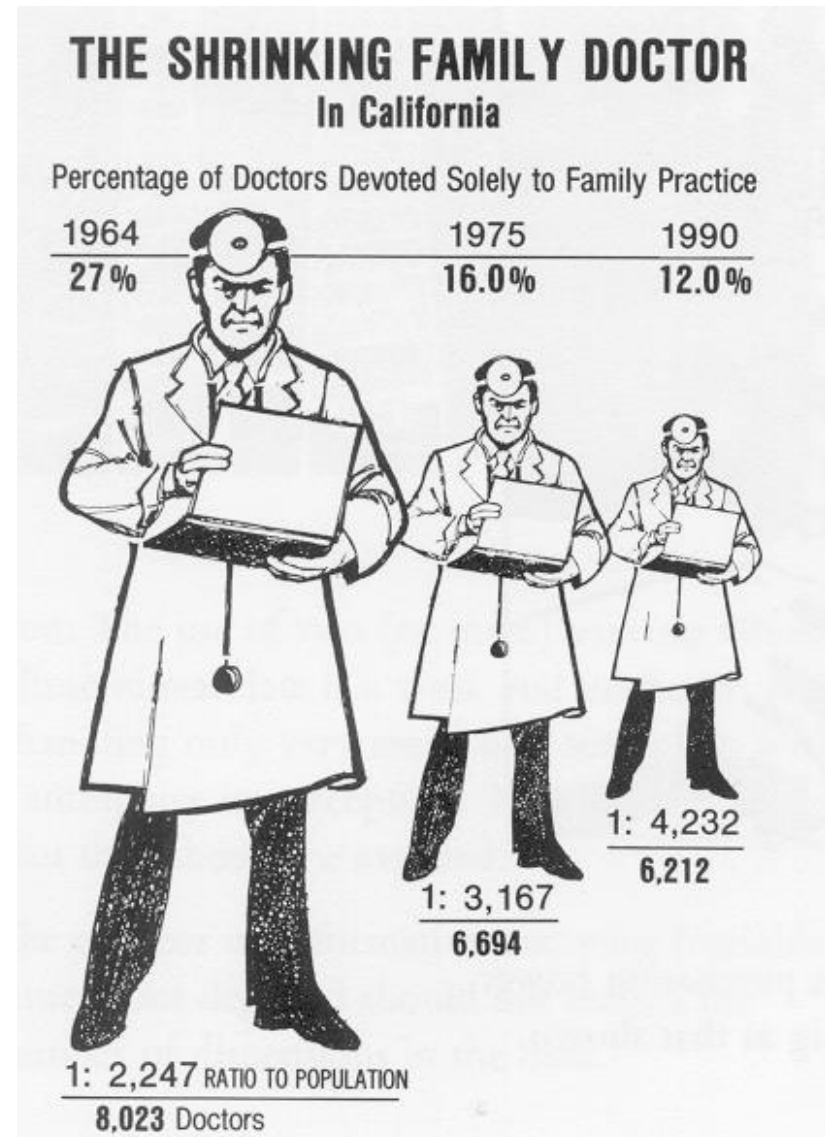
-Area has 2 dimensions

The 'incredible' shrinking family doctor

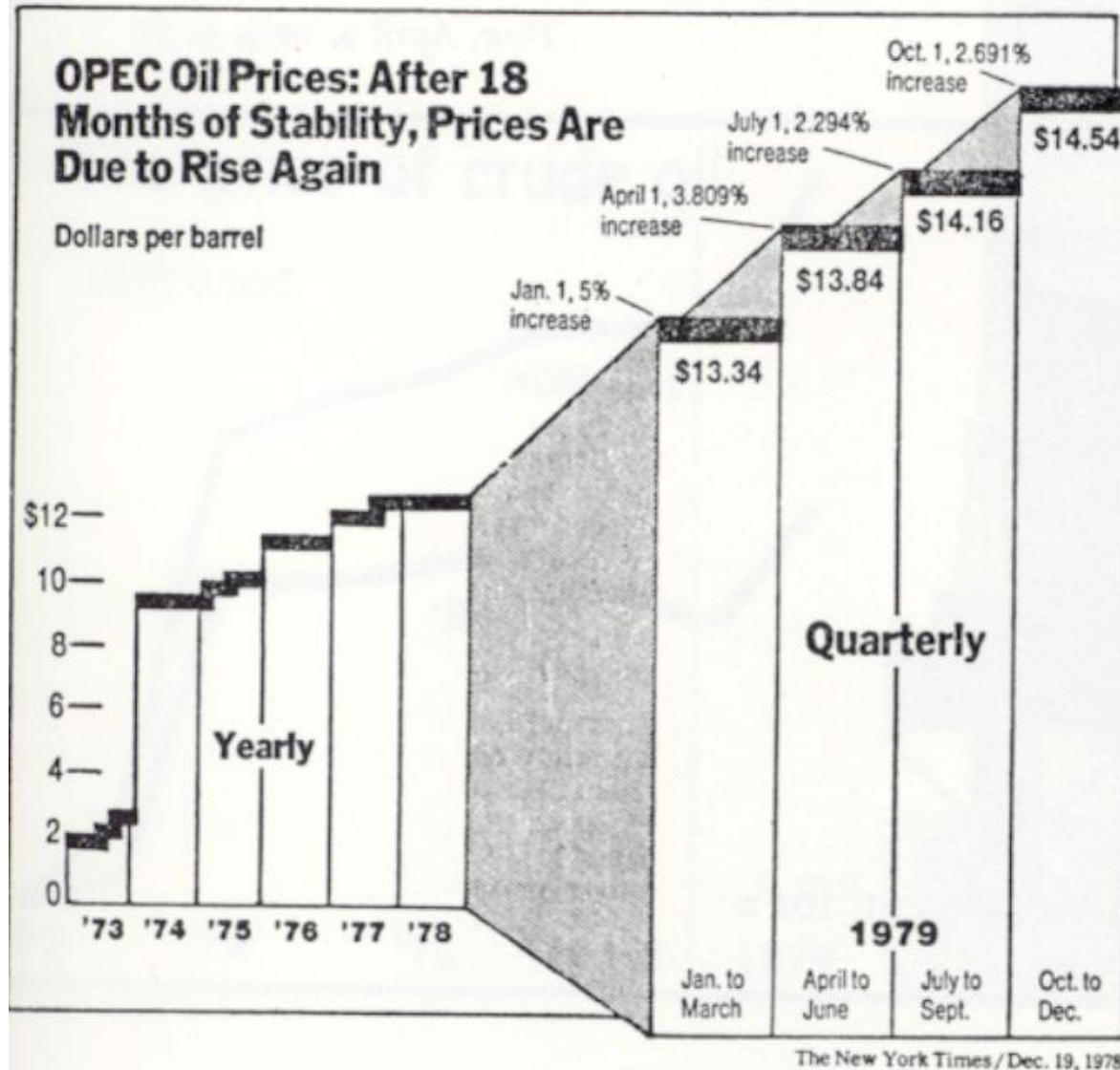
Lie factor of 2.8

Plus incorrect horizontal spacing

Los Angeles Times, August 5, 1979 p.3, (Tufte, 1983, p69)



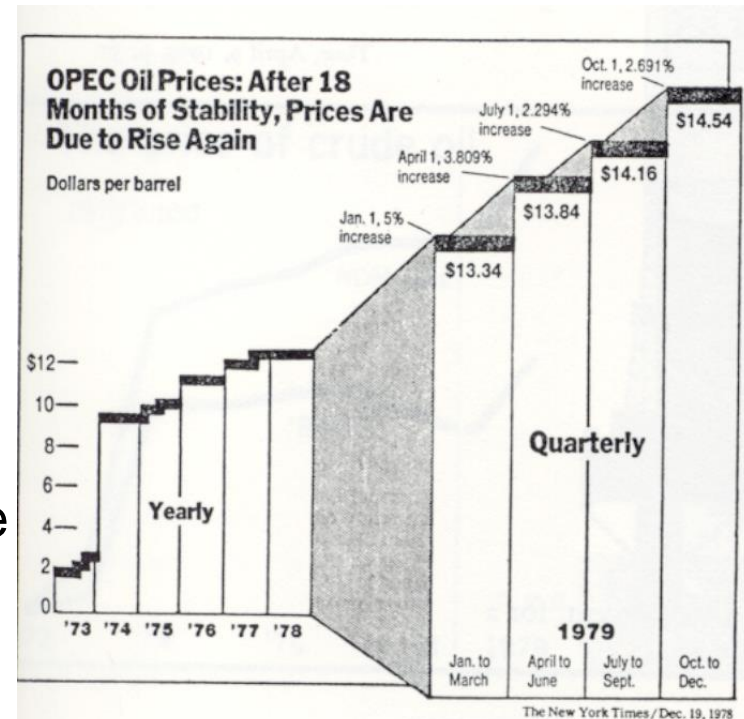
Design Variation vs Data Variation



New York Times, Dec. 19, 1978, p.D-7 (Tufte, 1983, p61)

Design Variation vs Data Variation

- 5 different vertical scales show price
 - During this time one vertical inch equals
 - 1973 -1978 \$8.00
 - Jan. – Mar. 1979 \$4.73
 - Apr. – June 1979 \$4.37
 - Jul. – Sept. 1979 \$4.16
 - Oct. – Dec. 1979 \$3.92
- 2 different horizontal scales show passage time
 - During this time one horizontal inch equals
 - 1973-1978 3.8 years
 - 1979 0.57 years
- With both scales shifting the distortion is multiplicative



National Science Foundation, Science Indicators, 1974
(Washington D.C., 1976), p.15, (Tufte, 1983, p60)

Show data variation, not design variation!

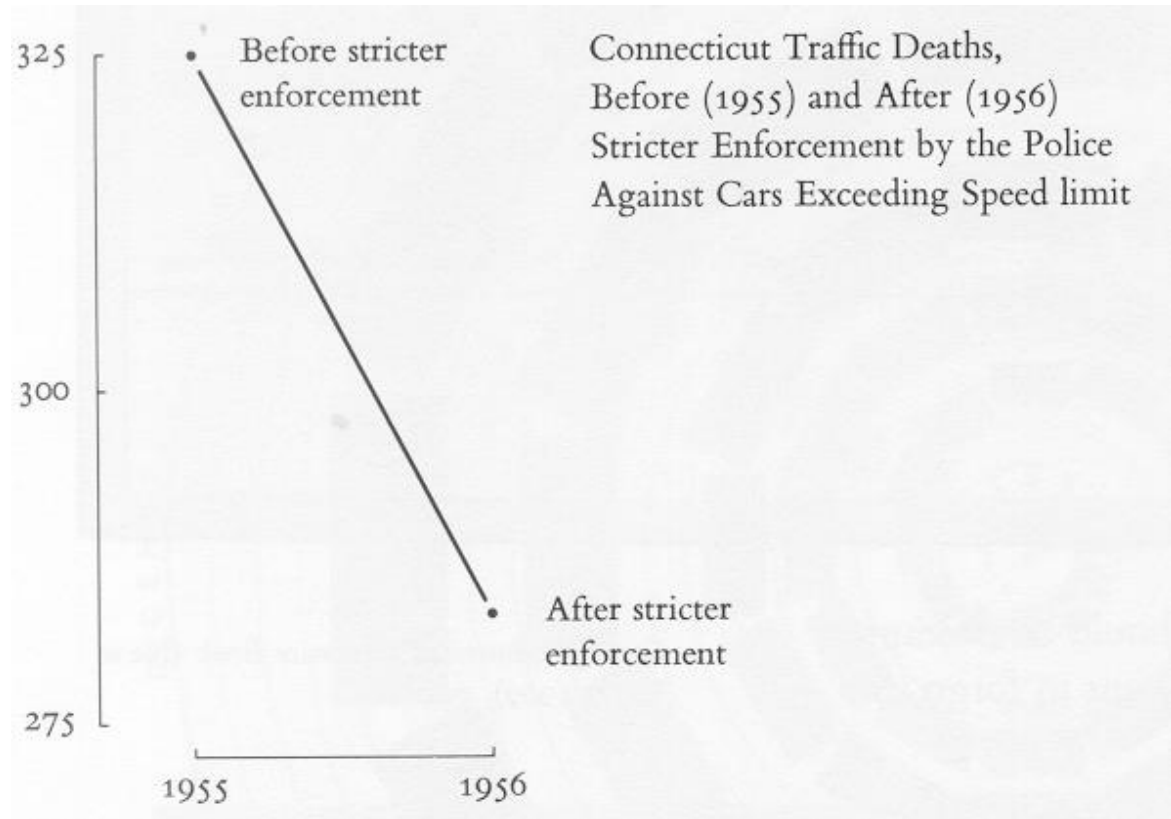
Context is Essential

Graphics must not quote data out of context

Data sparse graphics
should provoke suspicion

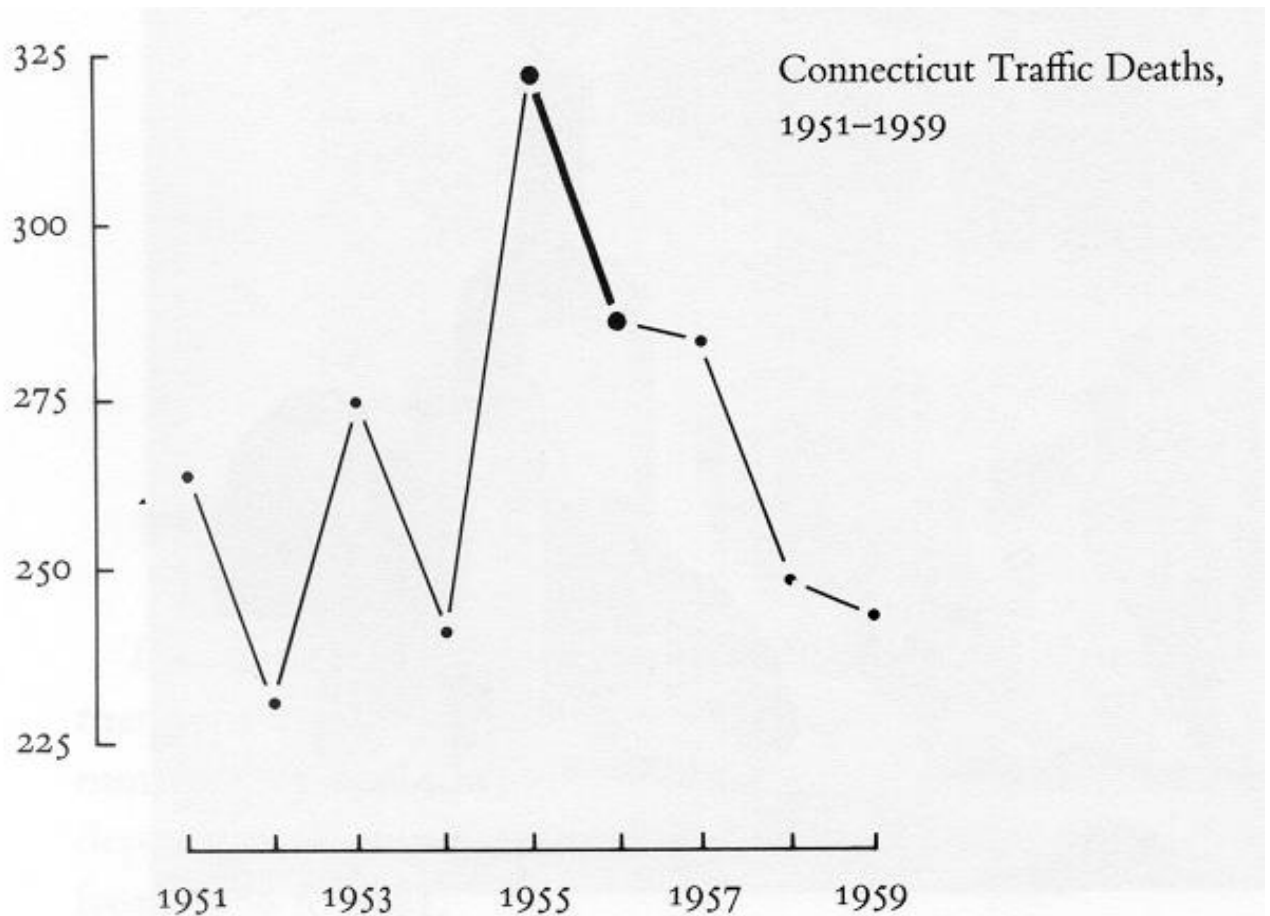
Graphics often lie by
omission

Nearly all important
questions are left
unanswered by this graph



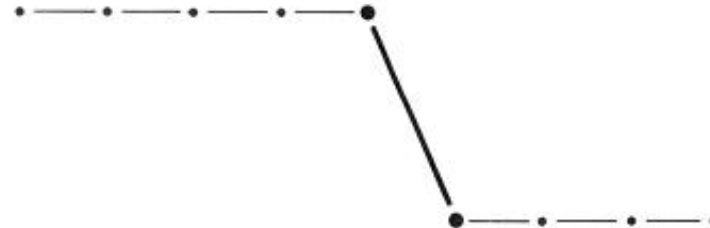
Context is Essential

A few more data points tell a more complete story



Context is Essential

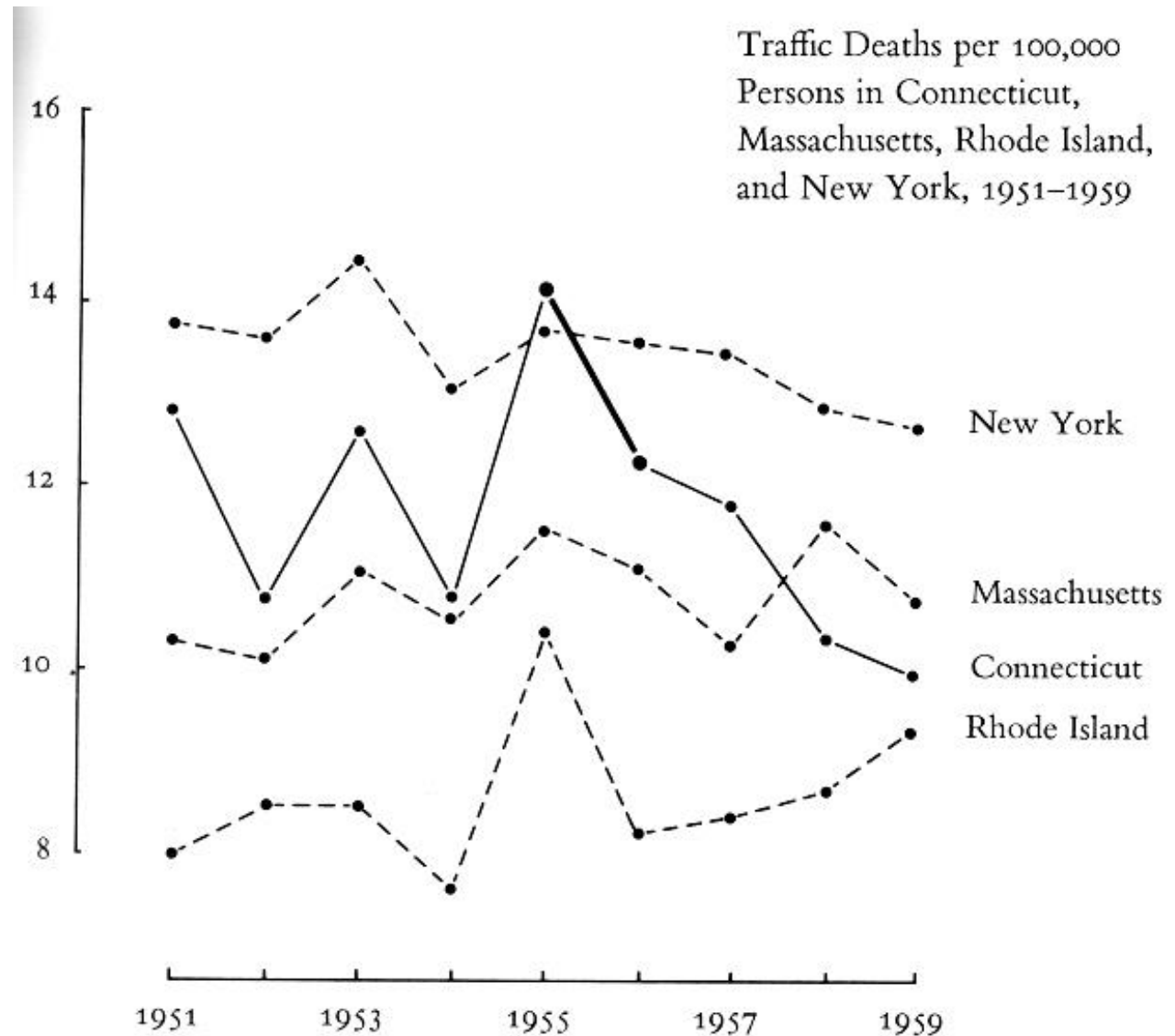
Different data points would tell
a different stories



Context is Essential

Comparisons with adjacent states give more context

Donald T. Campbell and H. Laurence Ross, "The Connecticut Crackdown on Speeding: Time Series Data in Quasi-Experimental Analysis," in Edward R. Tufte, ed., *The Quantitative Analysis of Social Problems* (Reading, Mass., 1970), 110–125.



Tufte: Principles of Graphical Excellence

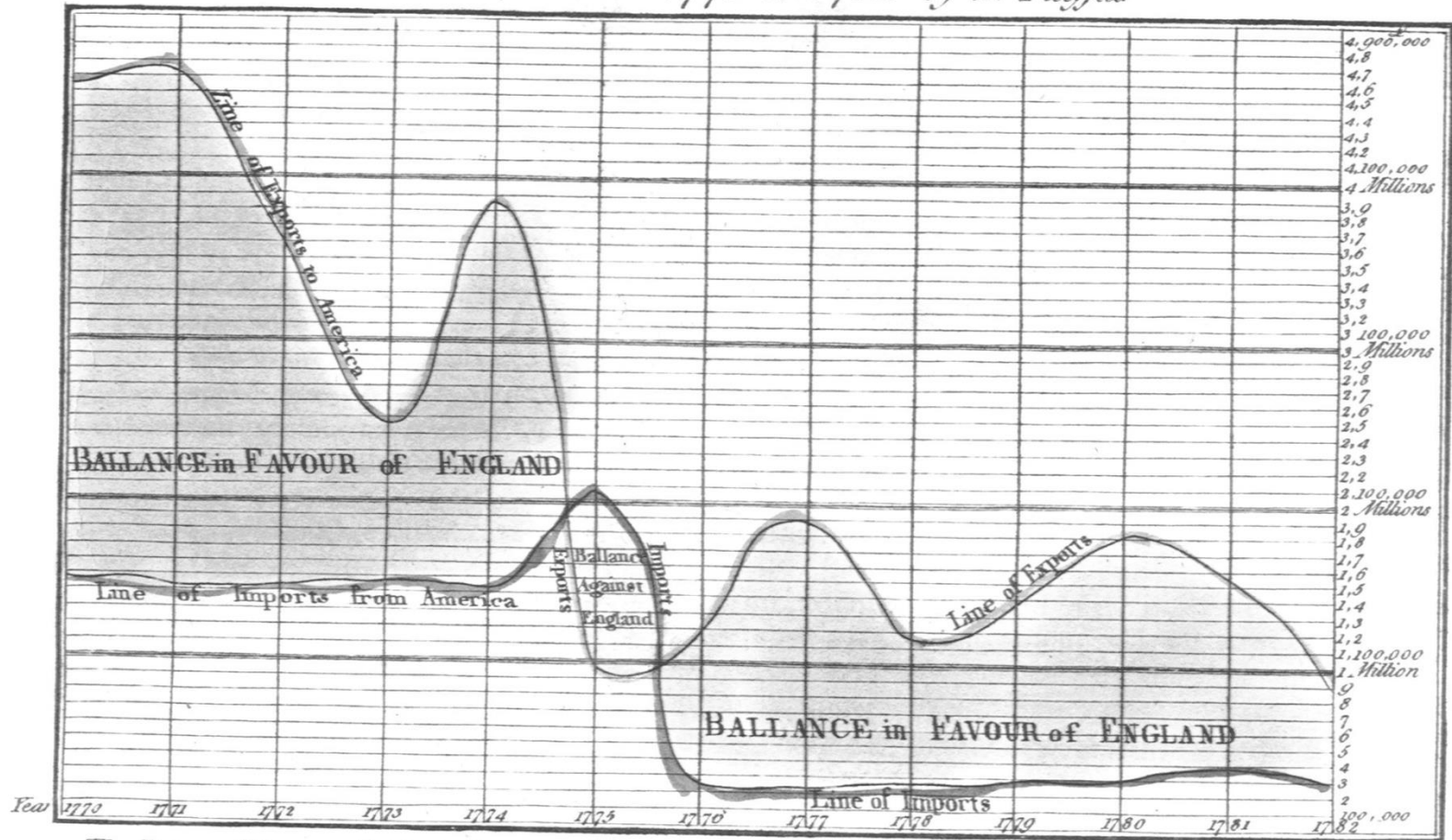
- Graphical excellence is the well-designed presentation of interesting data – a matter of *substance*, *statistics*, and *design*
- Graphical excellence consists of complex ideas communicated with *clarity*, *precision*, and *efficiency*

Tufte's principles for better viz?

- Above all else, show the data
- Maximize the data-ink ratio
 - Erase non-data-ink
 - Erase redundant data-ink
- Revise and edit

Above all else, show the data

*CHART of IMPORTS and EXPORTS of ENGLAND to and from all NORTH AMERICA
From the Year 1770 to 1782 by W. Playfair*



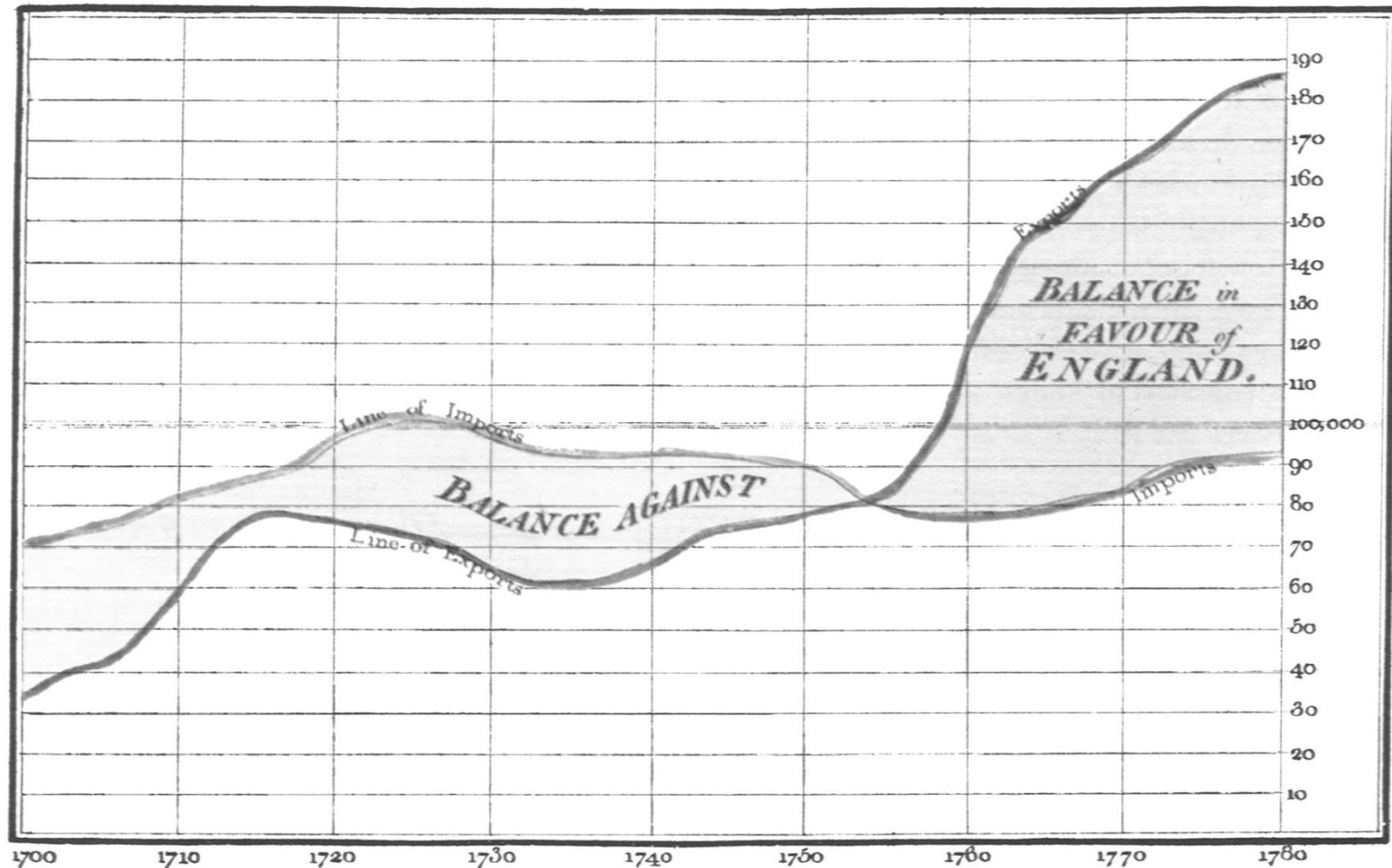
The Bottom Line is divided into Years the right-hand Line into HUNDRED THOUSAND POUNDS

J. Ansell Sculp.

Published as the Act directs 20th Aug^r 1785.

Above all else, show the data

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



The Bottom line is divided into Years, the Right hand line into £10,000 each.

Published as the Act directs, 16th May 1786. by W^m. Playfair

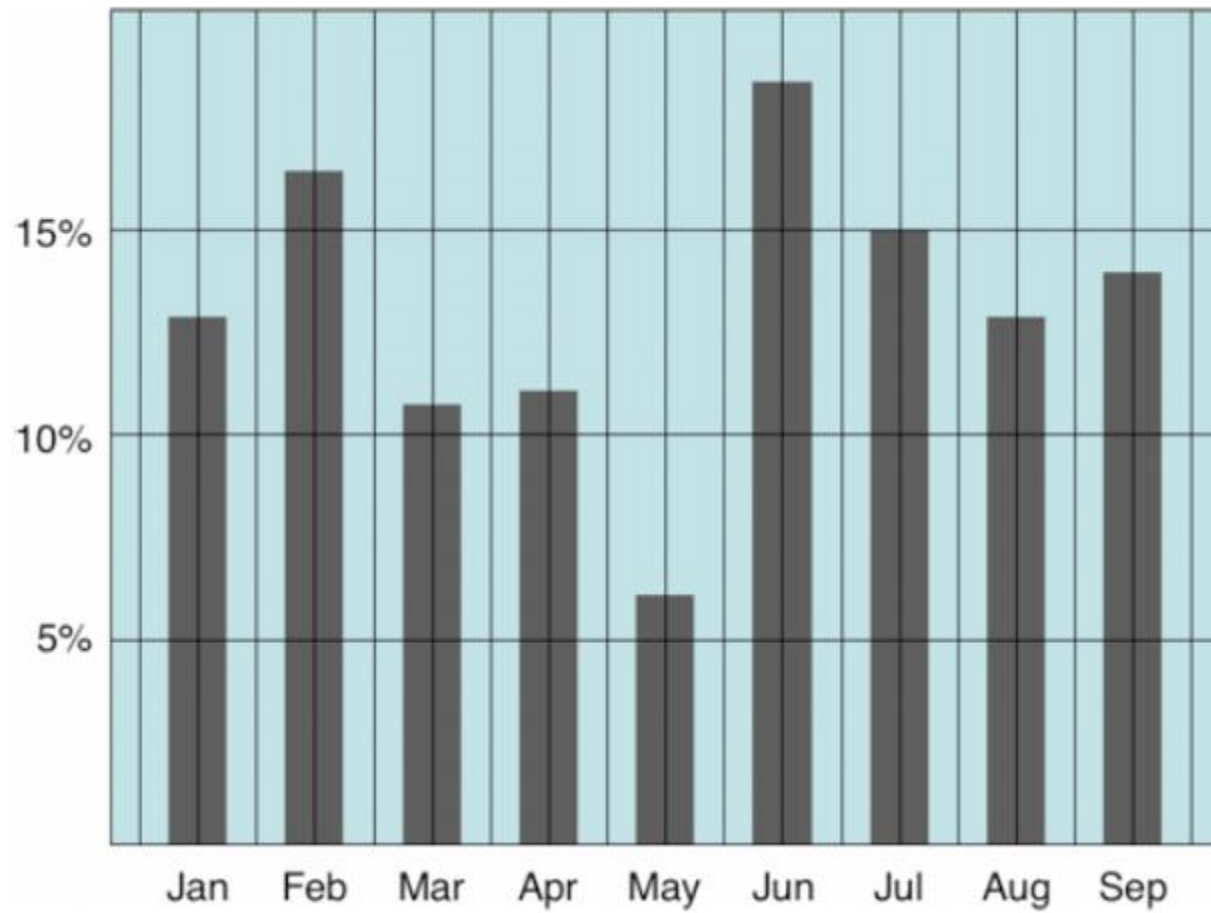
No. 100, Strand, London.

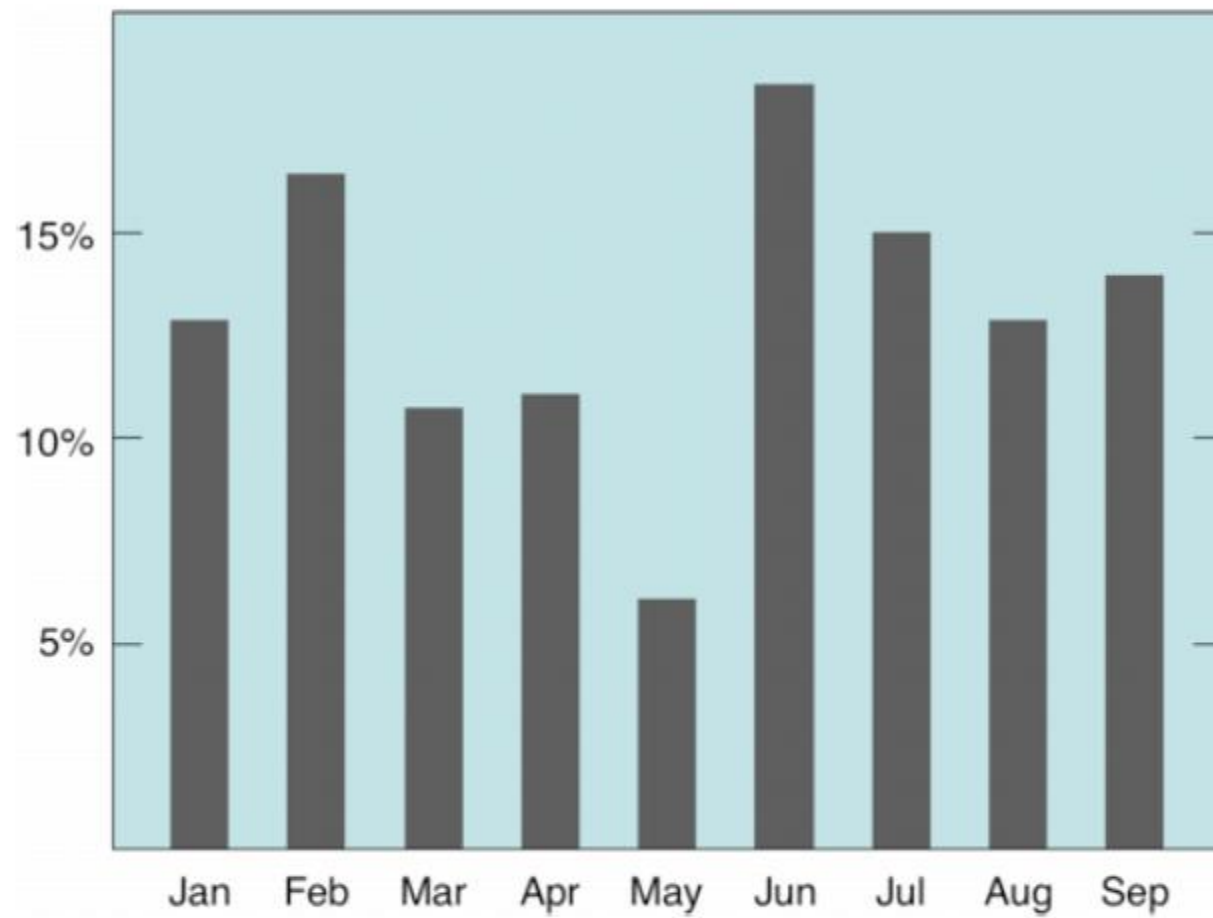
Maximize the data-ink ratio

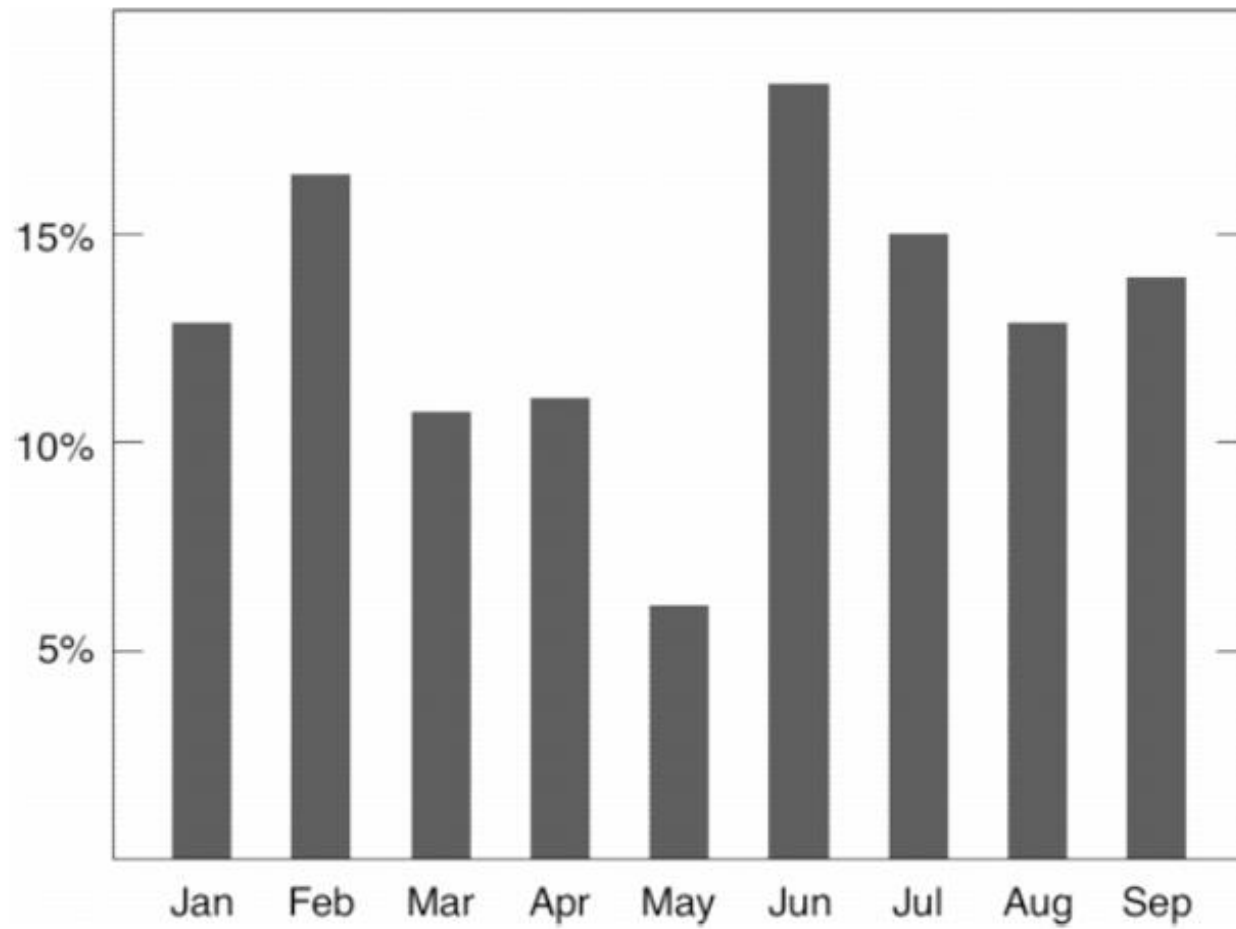
$$\begin{aligned}\text{Data-ink ratio} &= \frac{\text{data-ink}}{\text{Total ink used to print graphic}} \\ &= \text{Proportion of a graphic's ink} \\ &\quad \text{devoted to the non-redundant} \\ &\quad \text{display of data-information.} \\ &= 1.0 - \text{proportion of graphic} \\ &\quad \text{that can be erased without} \\ &\quad \text{the loss of information}\end{aligned}$$

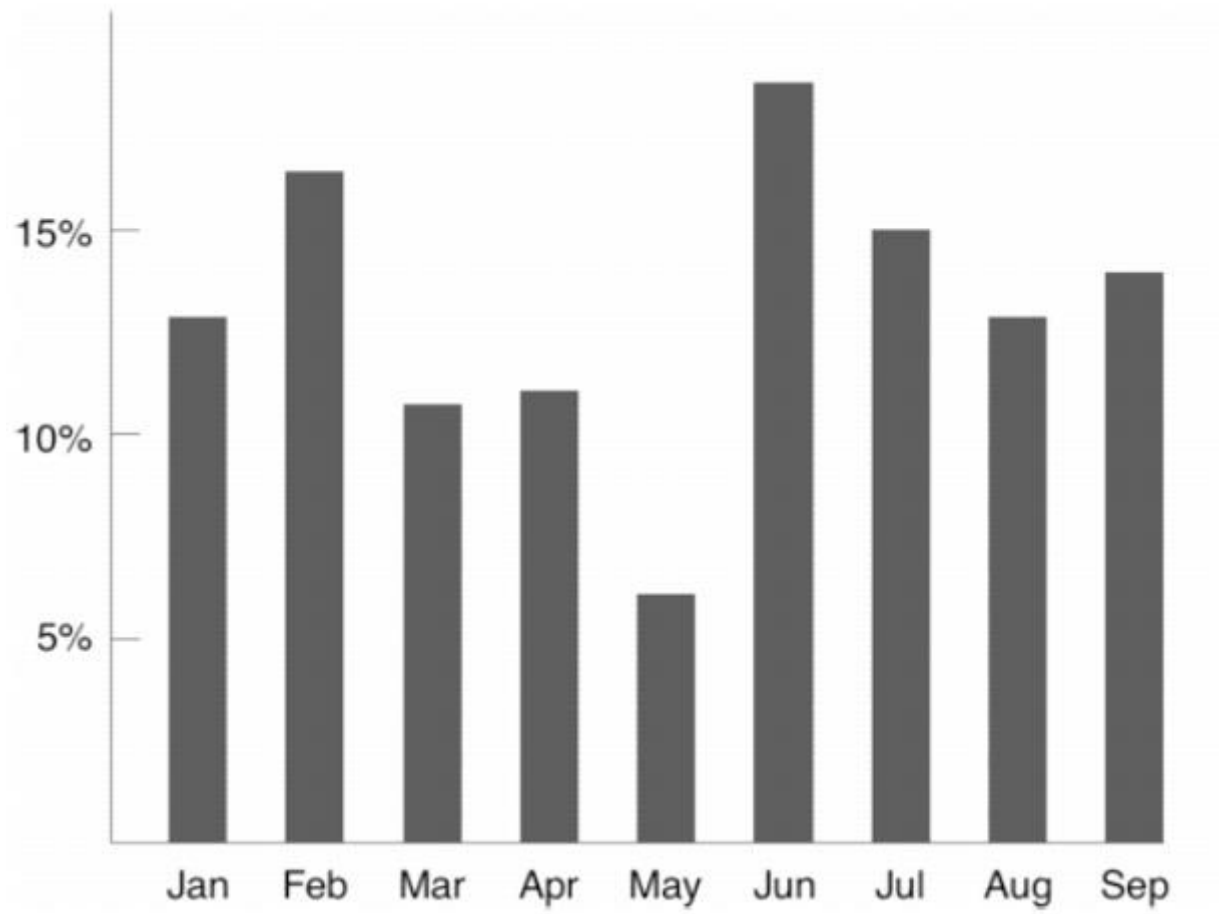
Maximize the data-ink ratio

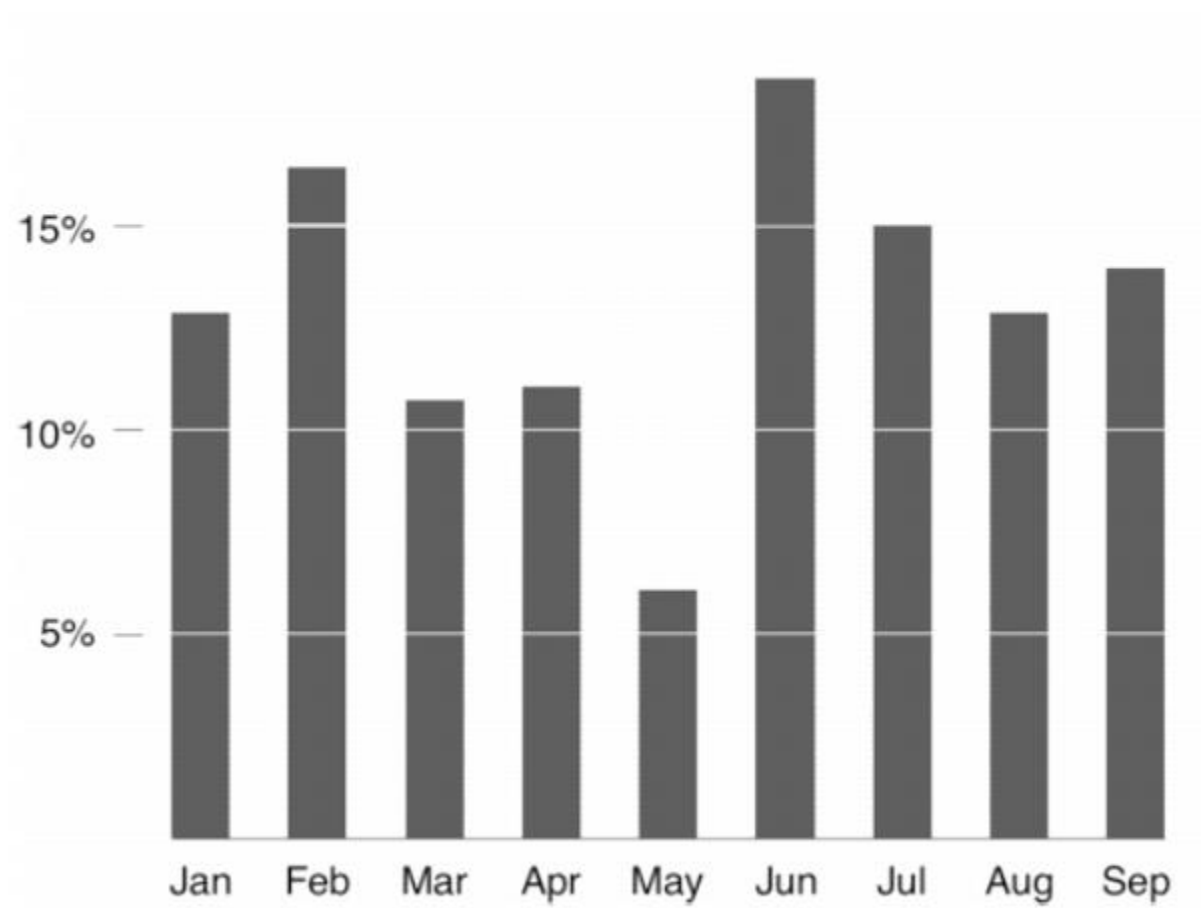
- Within reason
- In essence, you should be able to argue for every pixel
- Starting point:
 - erase non-data ink
 - erase redundant data-ink

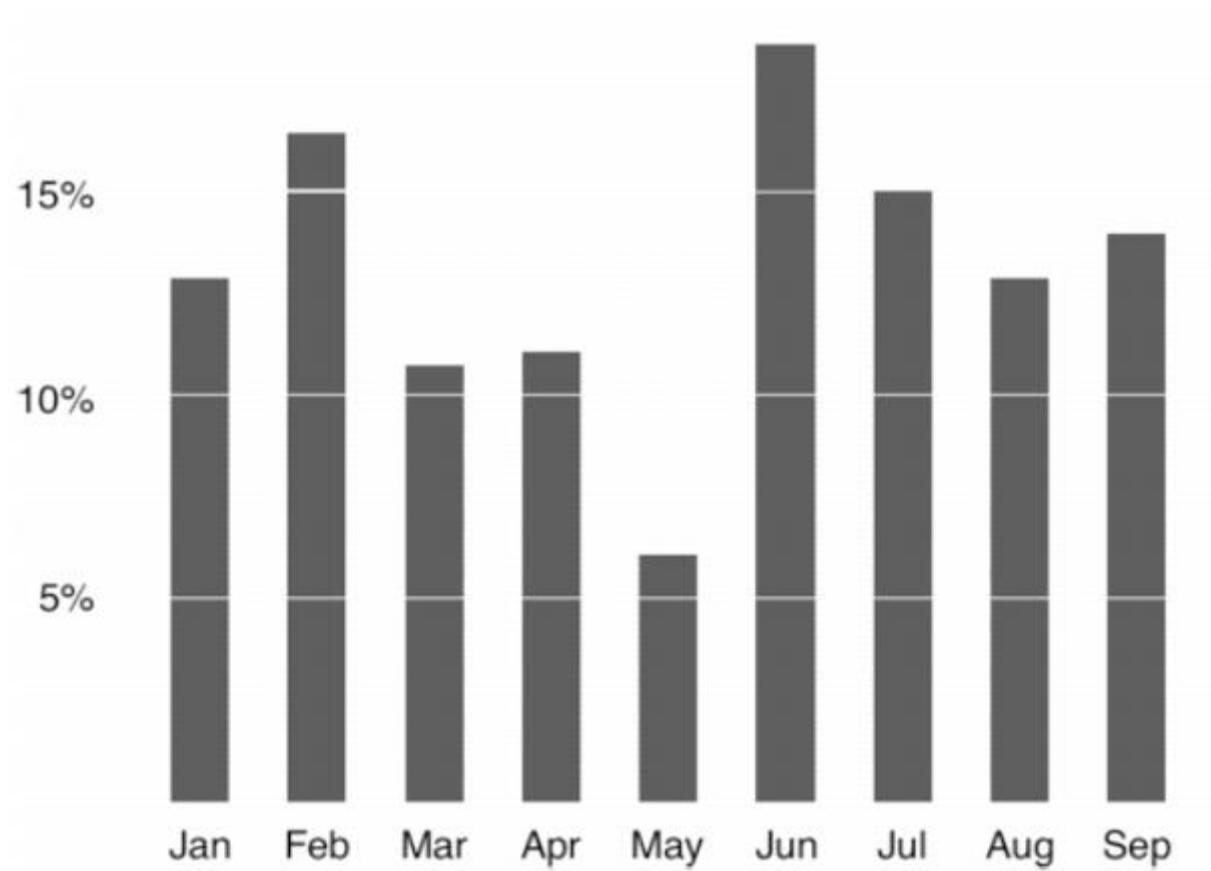












Summary

- Show data variation, not design variation
- Avoid using ink for non-data items
- Avoid redundancy
- Clear and detailed labeling should be used to defeat graphical distortion
- Revise and Edit

Other Examples

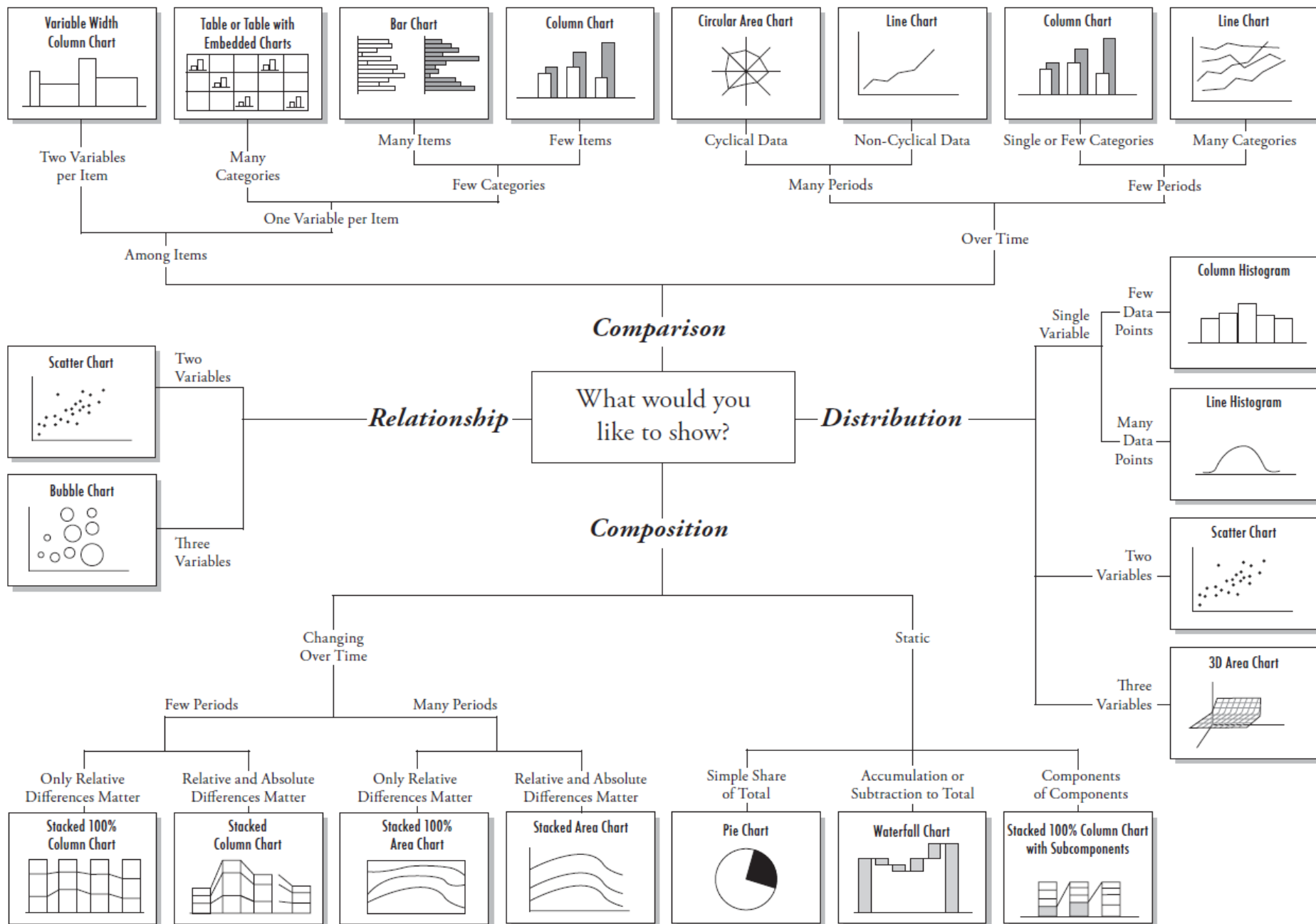
http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html

[http://submarine-cable-map-
2013.telegeography.com/](http://submarine-cable-map-2013.telegeography.com/)

<http://www.npr.org/sections/itsallpolitics/2012/11/01/163632378/a-campaign-map-morphed-by-money>

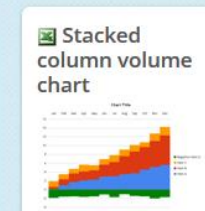
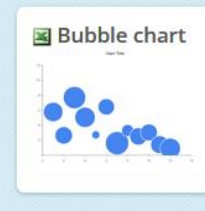
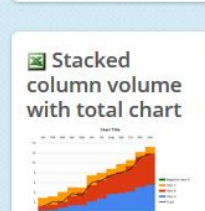
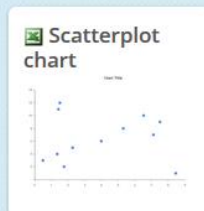
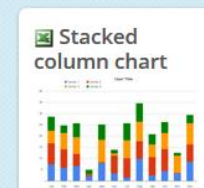
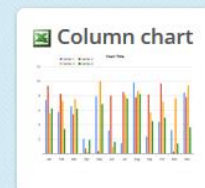
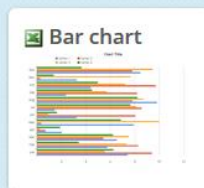
Practical Tips

Chart Suggestions—A Thought-Starter



Welcome to the new Chart Chooser! We've rebuilt our favorite tool for improving Excel and PowerPoint with HTML5 goodness. Use the filters above to find the right chart type for your needs. Then download as Excel or PowerPoint templates and insert your own data. Learn more about Chart Chooser [here](#).

Viewing 17 of 17 **All** Comparison Distribution **Composition** Trend Relationship Table For: Powerpoint **Excel** Share +



Quartiles table

Item	Q1	Q2	Q3	Q4	Q5
Apple	100	150	200	250	300
Banana	120	180	220	280	320
Orange	140	200	240	300	340
Pineapple	160	220	260	320	360
Watermelon	180	240	280	340	380
Grape	200	260	300	360	400
Strawberry	220	280	320	380	420
Peach	240	300	340	400	440
Cherry	260	320	360	420	460
Plum	280	340	380	440	480
Apricot	300	360	400	460	500
Persimmon	320	380	420	480	520
Fig	340	400	440	500	540
Jackfruit	360	420	460	520	560
Papaya	380	440	480	540	580
Mango	400	460	500	560	600
Guava	420	480	520	580	620
Lemon	440	500	540	600	640
Lime	460	520	560	620	660
Pomegranate	480	540	580	640	680
Coconut	500	560	600	660	700
Avocado	520	580	620	680	720
Blackberry	540	600	640	700	740
Raspberry	560	620	660	720	760
Blueberry	580	640	680	740	780
Blackcurrant	600	660	700	760	800
Rosehip	620	680	720	780	820
Gooseberry	640	700	740	800	840
Elderberry	660	720	760	820	860
Cloudberry	680	740	780	840	880
Huckleberry	700	760	800	860	900
Marionberry	720	780	820	880	920
Tart cherry	740	800	840	900	940
Montmorency	760	820	860	920	960
Smoked cherry	780	840	880	940	980
Amarelle	800	860	900	960	1000
White cherry	820	880	920	980	1020
Black cherry	840	900	940	1000	1040
Red cherry	860	920	960	1020	1060
Yellow cherry	880	940	980	1040	1080
Green cherry	900	960	1000	1060	1100
Pink cherry	920	980	1020	1080	1120
White cherry	940	1000	1040	1100	1140
Black cherry	960	1020	1060	1120	1160
Red cherry	980	1040	1080	1140	1180
Yellow cherry	1000	1060	1100	1160	1200
Green cherry	1020	1080	1120	1180	1220
Pink cherry	1040	1100	1140	1200	1240
White cherry	1060	1120	1160	1220	1260
Black cherry	1080	1140	1180	1240	1280
Red cherry	1100	1160	1200	1260	1300
Yellow cherry	1120	1180	1220	1280	1320
Green cherry	1140	1200	1240	1300	1340
Pink cherry	1160	1220	1260	1320	1360
White cherry	1180	1240	1280	1340	1380
Black cherry	1200	1260	1300	1360	1400
Red cherry	1220	1280	1320	1380	1420
Yellow cherry	1240	1300	1340	1400	1440
Green cherry	1260	1320	1360	1420	1460
Pink cherry	1280	1340	1380	1440	1480
White cherry	1300	1360	1400	1460	1500
Black cherry	1320	1380	1420	1480	1520
Red cherry	1340	1400	1440	1500	1540
Yellow cherry	1360	1420	1460	1520	1560
Green cherry	1380	1440	1480	1540	1580
Pink cherry	1400	1460	1500	1560	1600
White cherry	1420	1480	1520	1580	1620
Black cherry	1440	1500	1540	1600	1640
Red cherry	1460	1520	1560	1620	1660
Yellow cherry	1480	1540	1580	1640	1680
Green cherry	1500	1560	1600	1660	1700
Pink cherry	1520	1580	1620	1680	1720
White cherry	1540	1600	1640	1700	1740
Black cherry	1560	1620	1660	1720	1760
Red cherry	1580	1640	1680	1740	1780
Yellow cherry	1600	1660	1700	1760	1800
Green cherry	1620	1680	1720	1780	1820
Pink cherry	1640	1700	1740	1800	1840
White cherry	1660	1720	1760	1820	1860
Black cherry	1680	1740	1780	1840	1880
Red cherry	1700	1760	1800	1860	1900
Yellow cherry	1720	1780	1820	1880	1920
Green cherry	1740	1800	1840	1900	1940
Pink cherry	1760	1820	1860	1920	1960
White cherry	1780	1840	1880	1940	1980
Black cherry	1800	1860	1900	1960	2000
Red cherry	1820	1880	1920	1980	2020
Yellow cherry	1840	1900	1940	2000	2040
Green cherry	1860	1920	1960	2020	2060
Pink cherry	1880	1940	1980	2040	2080
White cherry	1900	1960	2000	2060	2100
Black cherry	1920	1980	2020	2080	2120
Red cherry	1940	2000	2040	2100	2140
Yellow cherry	1960	2020	2060	2120	2160
Green cherry	1980	2040	2080	2140	2180
Pink cherry	2000	2060	2100	2160	2200
White cherry	2020	2080	2120	2180	2220
Black cherry	2040	2100	2140	2200	2240
Red cherry	2060	2120	2160	2220	2260
Yellow cherry	2080	2140	2180	2240	2280
Green cherry	2100	2160	2200	2260	2300
Pink cherry	2120	2180	2220	2280	2320
White cherry	2140	2200	2240	2300	2340
Black cherry	2160	2220	2260	2320	2360
Red cherry	2180	2240	2280	2340	2380
Yellow cherry	2200	2260	2300	2360	2400
Green cherry	2220	2280	2320	2380	2420
Pink cherry	2240	2300	2340	2400	2440
White cherry	2260	2320	2360	2420	2460
Black cherry	2280	2340	2380	2440	2480
Red cherry	2300	2360	2400	2460	2500
Yellow cherry	2320	2380	2420	2480	2520
Green cherry	2340	2400	2440	2500	2540
Pink cherry	2360	2420	2460	2520	2560
White cherry	2380	2440	2480	2540	2580
Black cherry	2400	2460	2500	2560	2600
Red cherry	2420	2480	2520	2580	2620
Yellow cherry	2440	2500	2540	2600	2640
Green cherry	2460	2520	2560	2620	2660
Pink cherry	2480	2540	2580	2640	2680
White cherry	2500	2560	2600	2660	2700
Black cherry	2520	2580	2620	2680	2720
Red cherry	2540	2600	2640	2700	2740
Yellow cherry	2560	2620	2660	2720	2760
Green cherry	2580	2640	2680	2740	2780
Pink cherry	2600	2660	2700	2760	2800
White cherry	2620	2680	2720	2780	2820
Black cherry	2640	2700	2740	2800	2840
Red cherry	2660	2720	2760	2820	2860
Yellow cherry	2680	2740	2780	2840	2880
Green cherry	2700	2760	2800	2860	2900
Pink cherry	2720	2780	2820	2880	2920
White cherry	2740	2800	2840	2900	2940
Black cherry	2760	2820	2860	2920	2960
Red cherry	2780	2840	2880	2940	2980
Yellow cherry	2800	2860	2900	2960	3000
Green cherry	2820	2880	2920	2980	3020
Pink cherry	2840	2900	2940	3000	3040
White cherry	2860	2920	2960	3020	3060
Black cherry	2880	2940	2980	3040	3080
Red cherry	2900	2960	3000	3060	3100
Yellow cherry	2920	2980	3020	3080	3120
Green cherry	2940	3000	3040	3100	3140
Pink cherry	2960	3020	3060	3120	3160
White cherry	2980	3040	3080	3140	3180
Black cherry	3000	3060	3100	3160	3200
Red cherry	3020	3080	3120	3180	3220
Yellow cherry	3040	3100	3140	3200	3240
Green cherry	3060	3120	3160	3220	3260
Pink cherry	3080	3140	3180	3240	3280
White cherry	3100	3160	3200	3260	3300
Black cherry	3120	3180	3220	3280	3320
Red cherry	3140	3200	3240	3300	3340
Yellow cherry	3160	3220	3260	3320	3360
Green cherry	3180	3240	3280	3340	3380
Pink cherry	3200	3260	3300	3360	3400
White cherry	3220	3280	3320	3380	3420
Black cherry	3240	3300	3340	3400	3440
Red cherry	3260	3320	3360	3420	3460
Yellow cherry	3280	3340	3380	3440	3480
Green cherry	3300	3360	3400	3460	3500
Pink cherry	3320	3380	3420	3480	3520
White cherry	3340	3400	3440	3500	3540
Black cherry	3360	3420	3460	3520	3560
Red cherry	3380	3440	3480	3540	3580
Yellow cherry	3400	3460	3500	3560	3600
Green cherry	3420	3480	3520	3580	3620
Pink cherry	3440	3500	3540	3600	3640
White cherry	3460	3520	3560	3620	3660
Black cherry	3480	3540	3580	3640	3680
Red cherry	3500	3560	3600	3660	3700
Yellow cherry	3520	3580	3620	3680	3720
Green cherry	3540	3600	3640	3700	3740
Pink cherry	3560	3620	3660	3720	3760
White cherry	3580	3640	3680	3740	3780
Black cherry	3600	3660	3700	3760	3800
Red cherry	3620	3680	3720	3780	3820
Yellow cherry	3640	3700	3740	3800	3840
Green cherry	3660	3720	3760	3820	3860
Pink cherry	3680	3740	3780	3840	3880
White cherry	3700	3760	3800	3860	3900
Black cherry	3720	3780	3820	3880	3920
Red cherry	3740	3800	3840	3900	3940
Yellow cherry	3760	3820	3860	3920	3960
Green cherry	3780	3840	3880	3940	3980
Pink cherry	3800	3860	3900	3960	4000
White cherry	3820	3880	3920	3980	4020
Black cherry	3840	3900	3940	4000	4040
Red cherry	3860	3920	3960	4020	4060
Yellow cherry	3880	3940	3980	4040	4080
Green cherry	3900	3960	4000	4060	4100
Pink cherry	3920	3980	4020	4080	4120
White cherry	3940	4000	4040	4100	4140
Black cherry	3960	4020	4060	4120	4160
Red cherry	3980	4040	4080	4140	4180</

Tools

- Tableau
 - <https://www.tableau.com/solutions/gallery>
- Highcharts
 - <https://www.highcharts.com/demo>
- Matplotlib
 - A library for plotting data in Python
 - <https://matplotlib.org/examples/index.html>

Data Science: The Context

Ask question: What data needs to be recorded? or collected?



Real World



Humans behaving
Biology
Finance
Internet
Medicine
Sociology
Olympics



Raw Data is
Collected / Recorded

email
logs
medical records
surveys
blood drawn
(microarray)
olympic records
NYT web pages



Data is
Processed

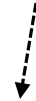
pipelines
web scraping
cleaning
munging
joining
wrangling



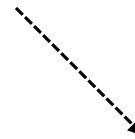
Data Set

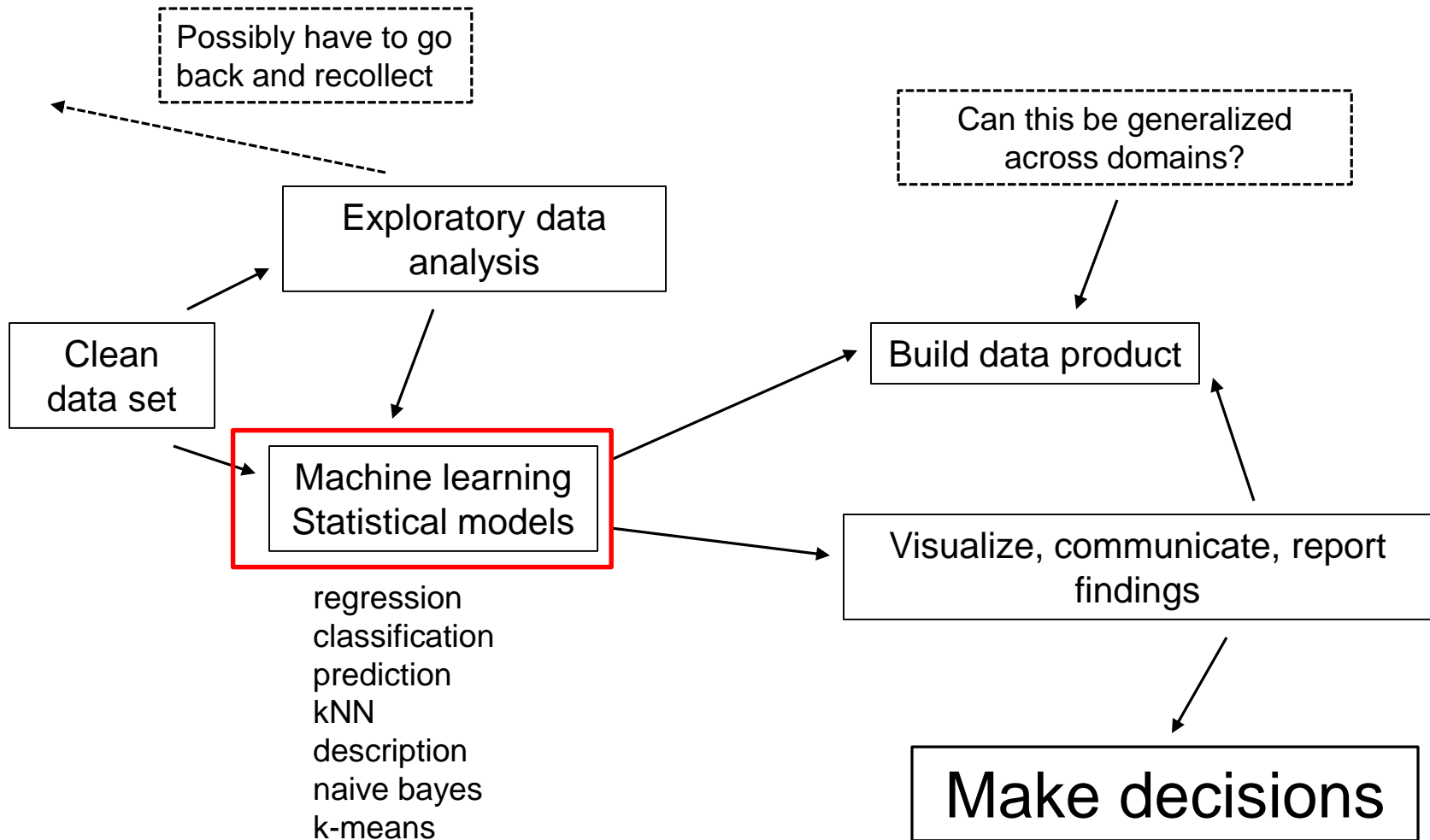
“clean” table

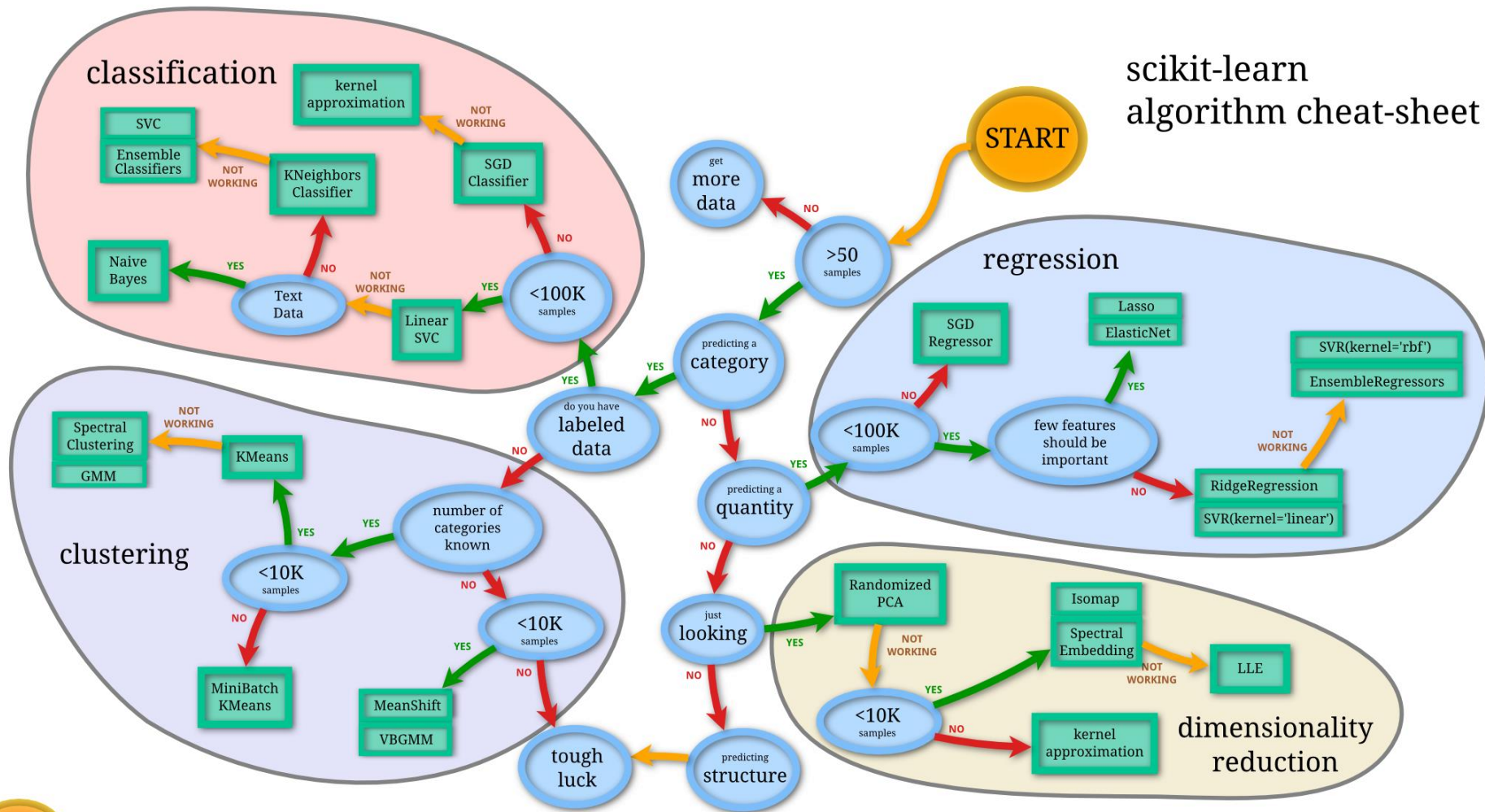
Why? What research question
am I going to answer?



What do I want it to look like?





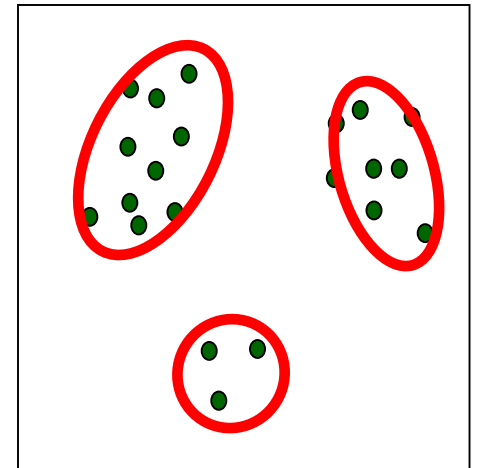
scikit-learn
algorithm cheat-sheet

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Clustering

Unsupervised Learning

- Supervised learning
 - Predict target value (“y”) given features (“x”)
 - E.g., classification and regression
- Unsupervised learning
 - Understand patterns of data (just “x”)
 - Useful for many reasons
 - Data mining (“explain”)
 - Representation (feature generation or selection)
 - E.g., **Clustering**



High Dimensional Data

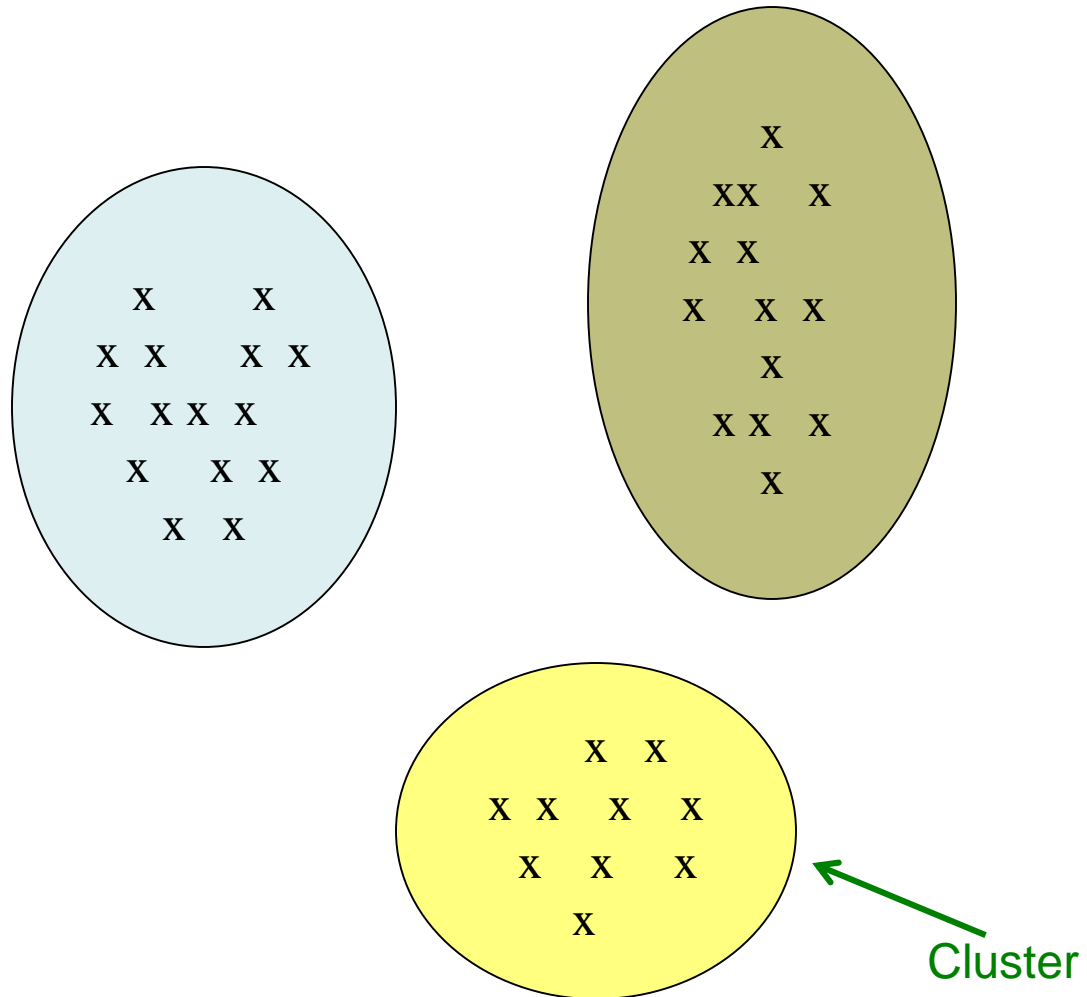
- Given a cloud of data points, we want to understand their structure



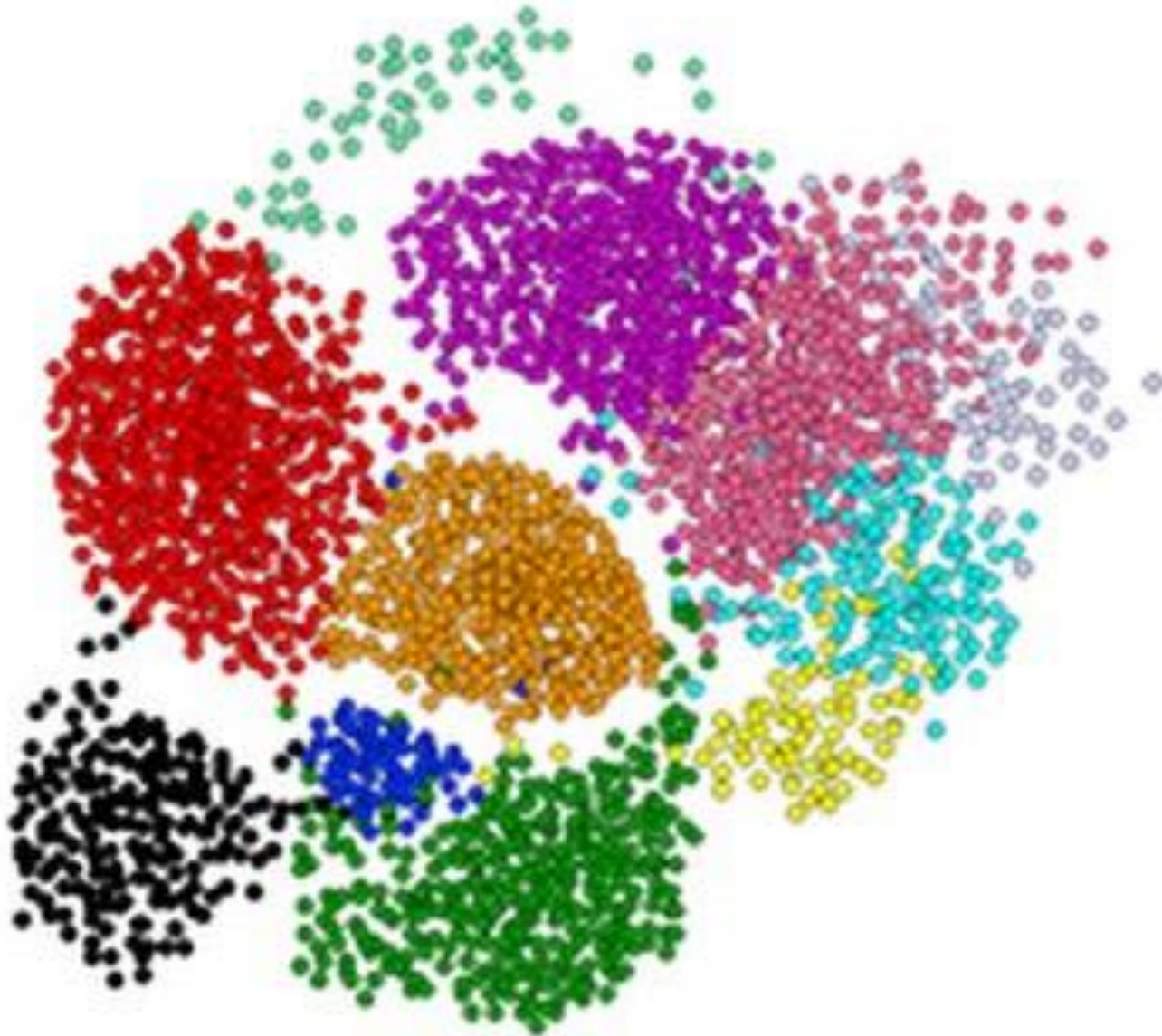
Clustering

- Given a **set of points**, with a notion of **distance** between points, **group the points** into some number of *clusters*, so that
 - Members of a cluster are close/similar to each other
 - Members of different clusters are dissimilar
- Usually:
 - Points are in a high-dimensional space
 - Similarity is defined using a distance measure
 - Euclidean, Cosine, Jaccard, edit distance, ...

Example Clusters



Clustering is Hard!



Why is it hard?

- Clustering in two dimensions looks easy
- Clustering small amounts of data looks easy
- And in most cases, looks are **not** deceiving
- Many applications involve not 2, but 10 or 10,000 dimensions
- **High-dimensional spaces look different:**
Almost all pairs of points are at about the same distance

- Typical applications
 - As a **stand-alone tool** to get insight into data distribution
 - As a **preprocessing step** for other algorithms

Clustering: Application Examples

- **Biology**: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- **Information retrieval**: document clustering
- **Land use**: Identification of areas of similar land use in an earth observation database
- **Marketing**: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **City-planning**: Identifying groups of houses according to their house type, value, and geographical location
- **Earth-quake studies**: Observed earth quake epicenters should be clustered along continent faults
- **Climate**: understanding earth climate, find patterns of atmospheric and ocean
- **Economic Science**: market research

Example: Clustering Songs

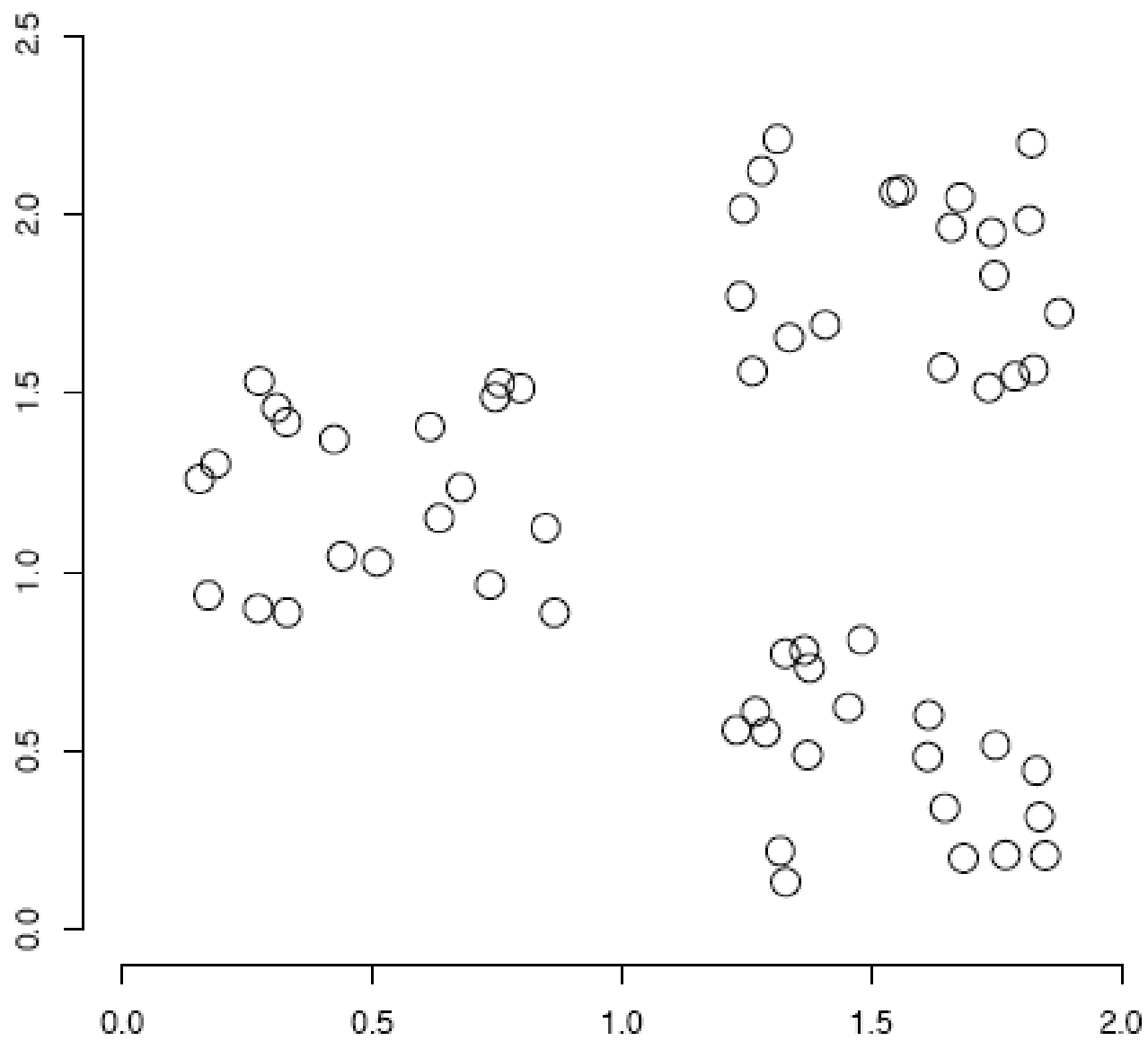
- Intuitively: Music divides into categories, and customers prefer a few categories
 - But what are categories really?
- Represent a song by a set of customers who downloaded it
- Similar songs have similar sets of downloaders, and vice-versa
- **Goal: Find clusters of similar songs**

Challenge

- To cluster songs:
 - How do we define the problem?
 - How do we tackle it?
- Hint: Represent a song by a set of customers who downloaded it

- k-means
- k-medoids
- Naive Bayes
- EM clustering (probabilistic)

K-means



K-means (in one slide!)

Input is **k** (the number of clusters), **data points** in Euclidean space

0. Initialize clusters by picking one point per cluster

Loop:

1. Place each point in the cluster whose current centroid is nearest
2. Find the new centroid for each cluster

K-means

- Objective/partitioning criterion: minimize the average squared difference from the centroid
- Assumes documents are real-valued vectors
- Clusters based on *centroids* (aka the *center of gravity* or mean) of points in a cluster, ω :

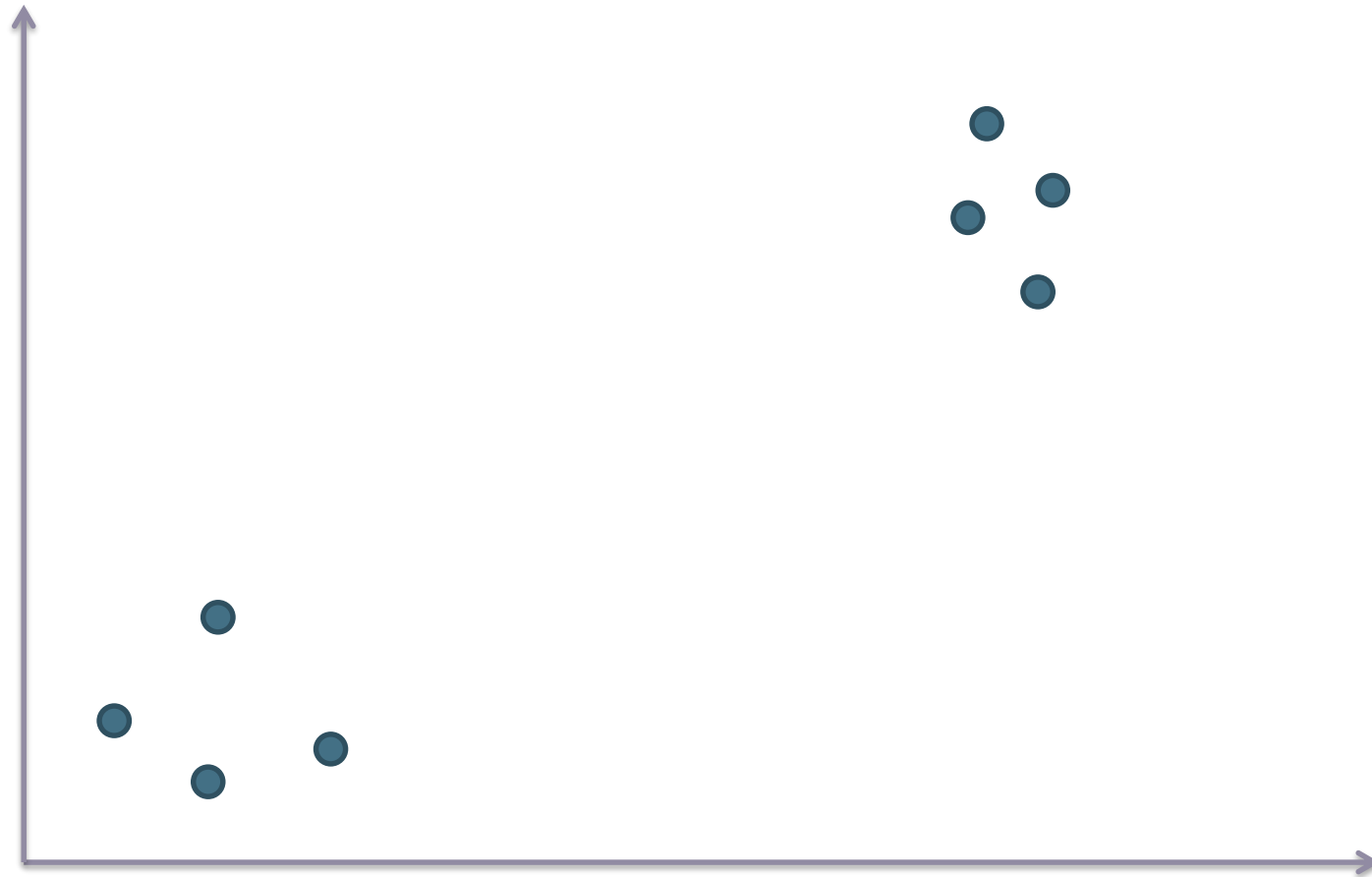
$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x}$$

- We try to find the minimum average squared difference by iterating two steps:
 - **reassignment**: assign each vector to its closest centroid
 - **recomputation**: recompute each centroid as the average of the vectors that were assigned to it in reassignment

K -MEANS($\{\vec{x}_1, \dots, \vec{x}_N\}, K$)

```
1   $(\vec{s}_1, \vec{s}_2, \dots, \vec{s}_K) \leftarrow \text{SELECTRANDOMSEEDS}(\{\vec{x}_1, \dots, \vec{x}_N\}, K)$ 
2  for  $k \leftarrow 1$  to  $K$ 
3  do  $\vec{\mu}_k \leftarrow \vec{s}_k$ 
4  while stopping criterion has not been met
5  do for  $k \leftarrow 1$  to  $K$ 
6      do  $\omega_k \leftarrow \{\}$ 
7      for  $n \leftarrow 1$  to  $N$ 
8      do  $j \leftarrow \arg \min_{j'} |\vec{\mu}_{j'} - \vec{x}_n|$ 
9           $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$  (reassignment of vectors)
10     for  $k \leftarrow 1$  to  $K$ 
11     do  $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$  (recomputation of centroids)
12 return  $\{\vec{\mu}_1, \dots, \vec{\mu}_K\}$ 
```

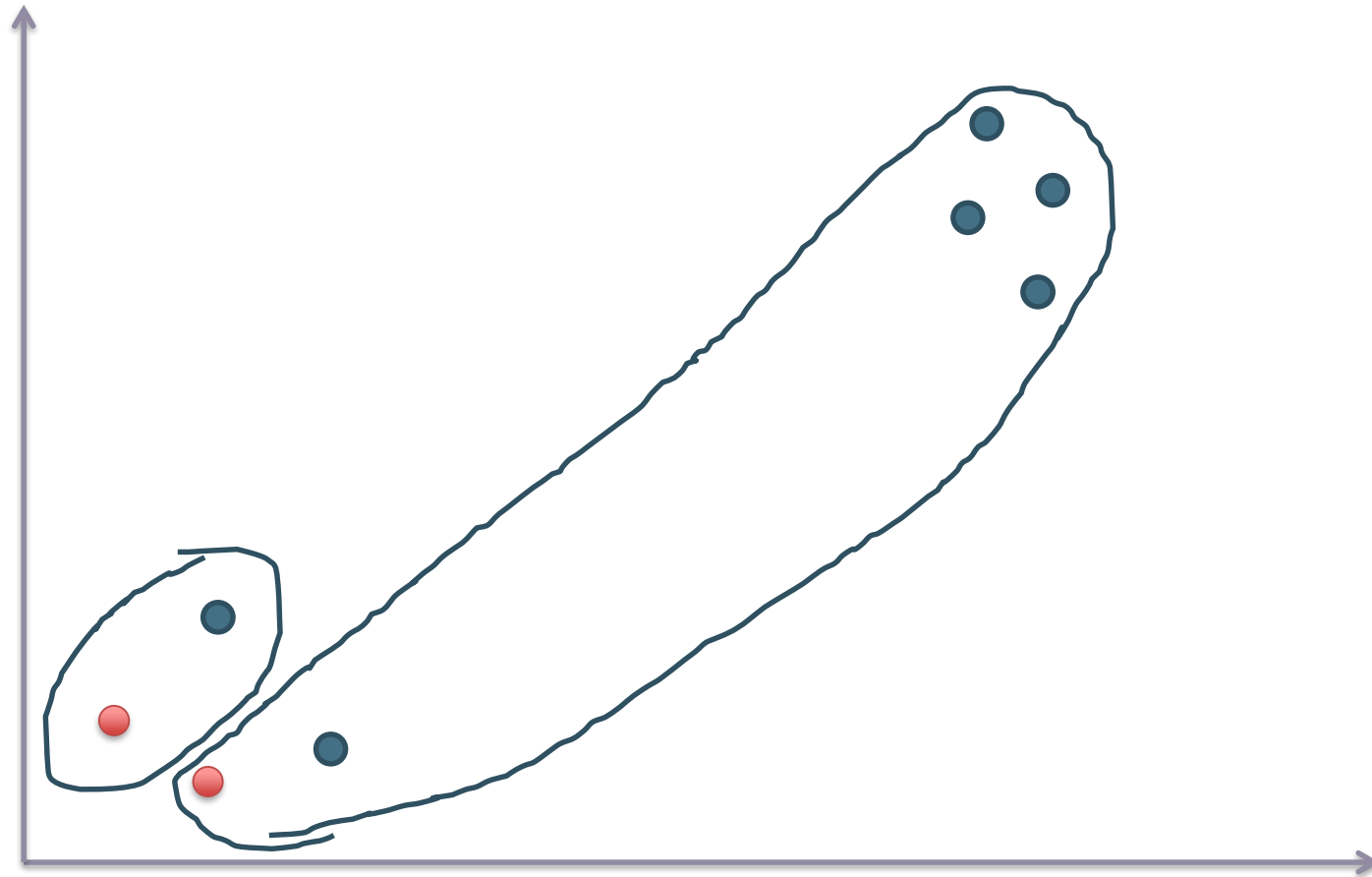
K-Means Clustering Example



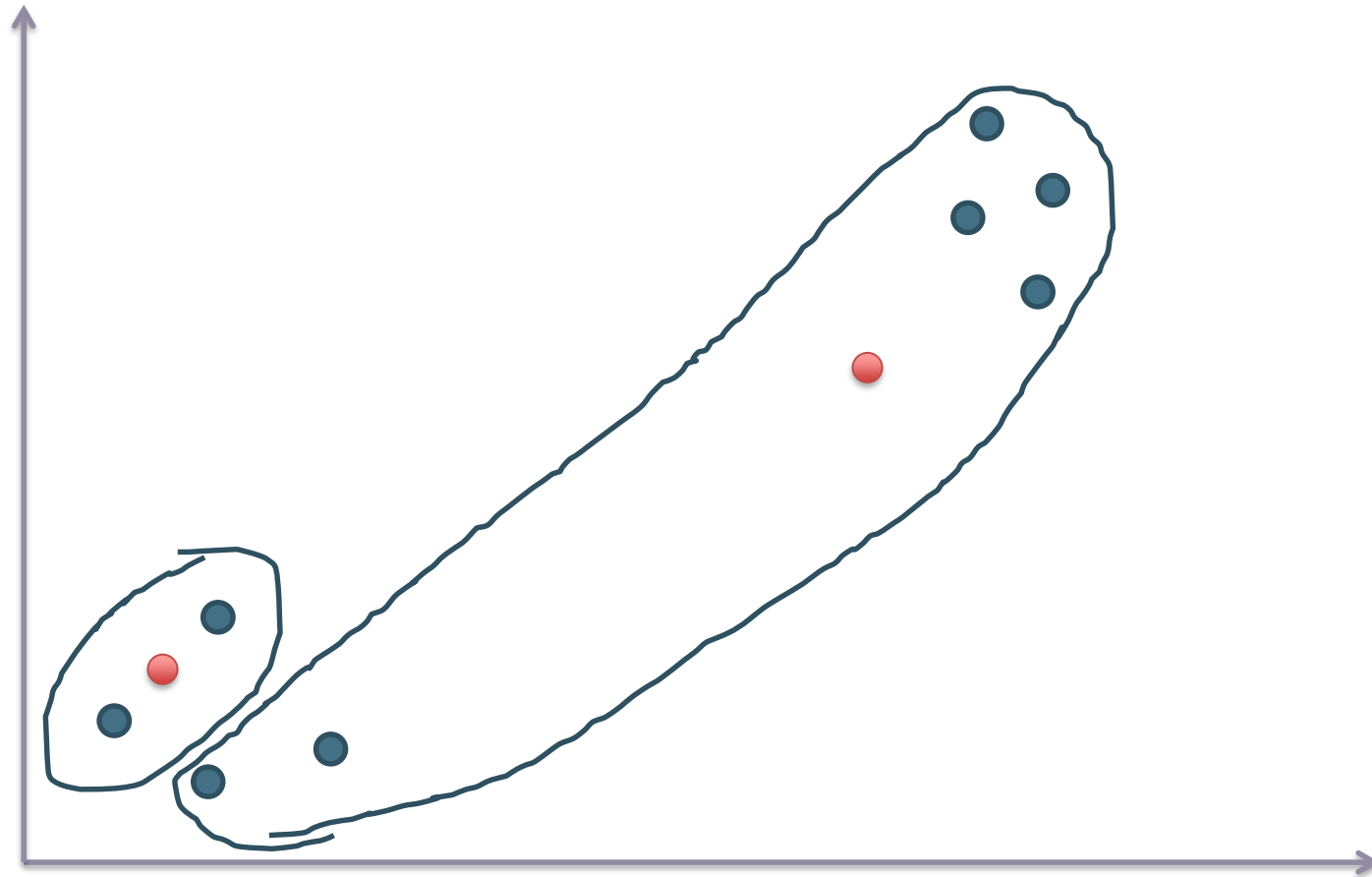
K-Means Clustering Example



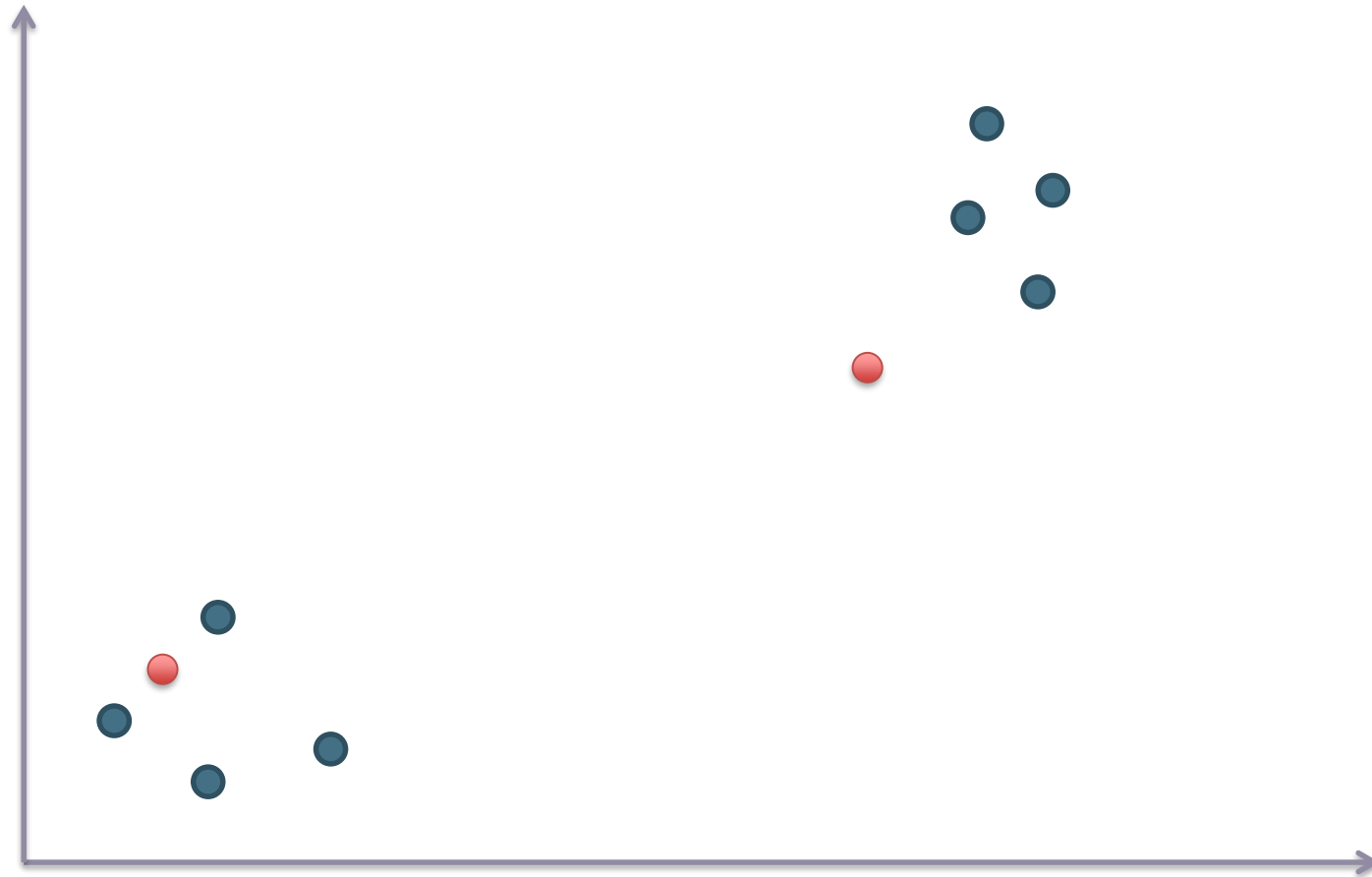
K-Means Clustering Example



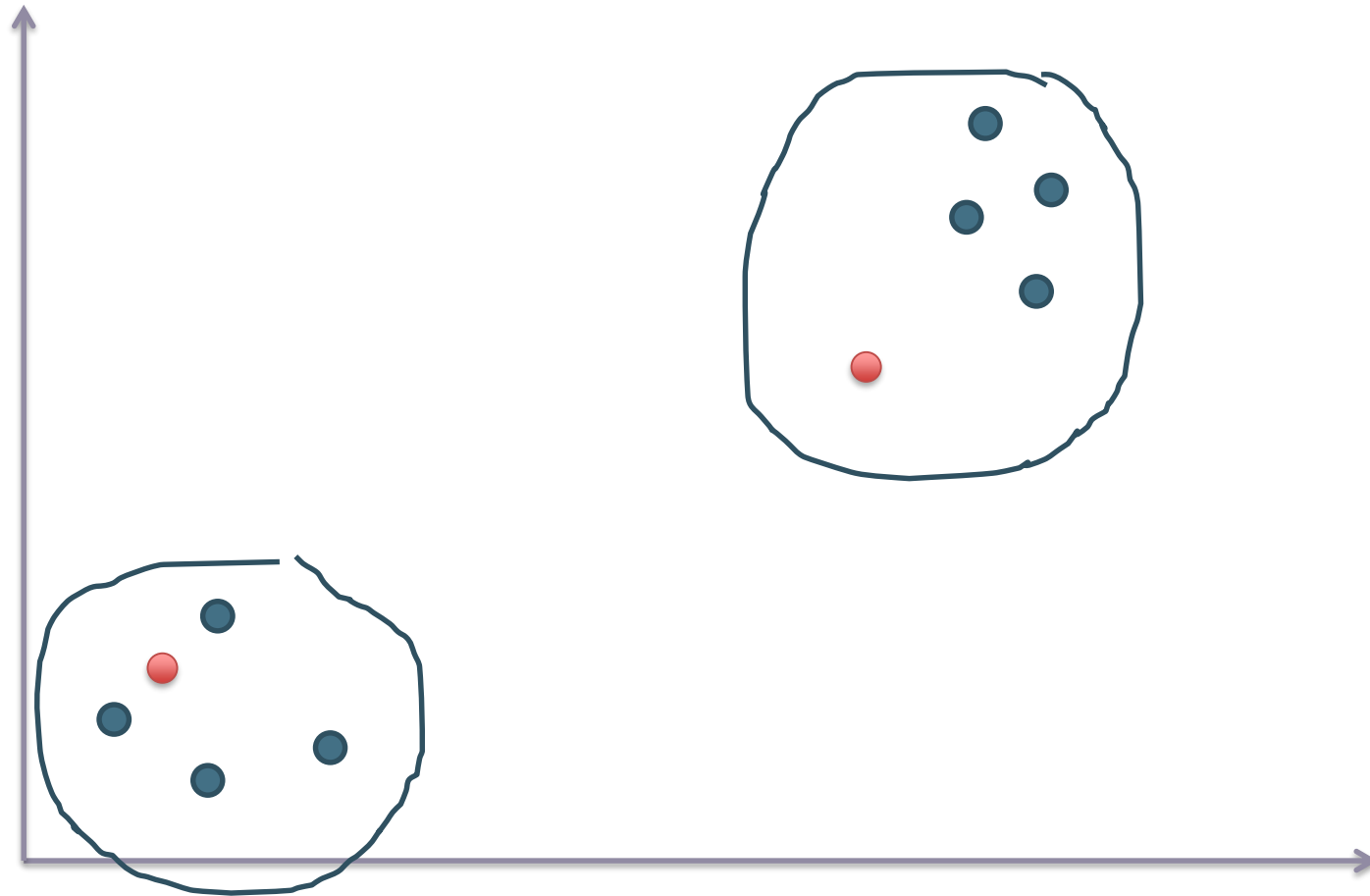
K-Means Clustering Example



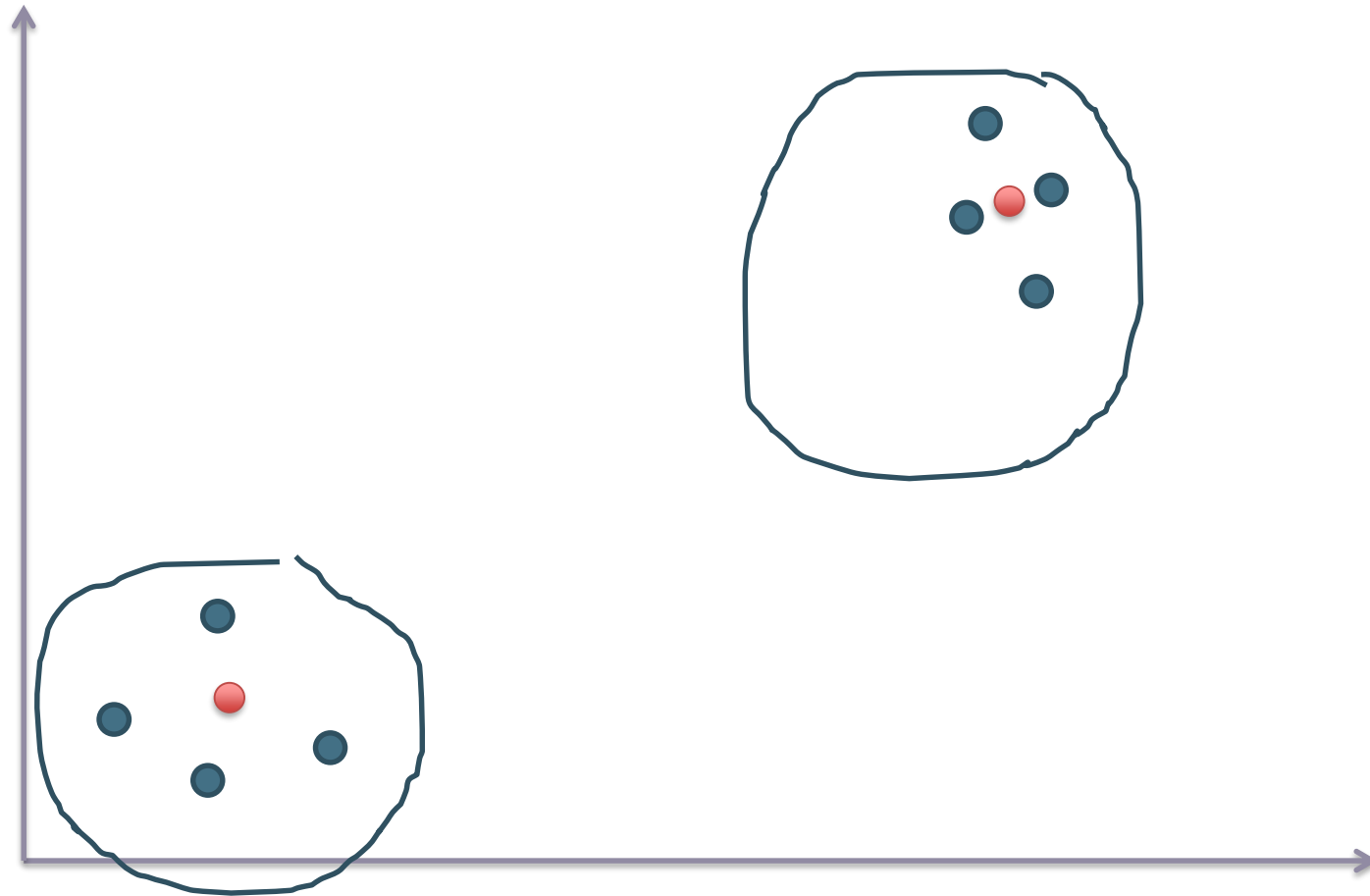
K-Means Clustering Example



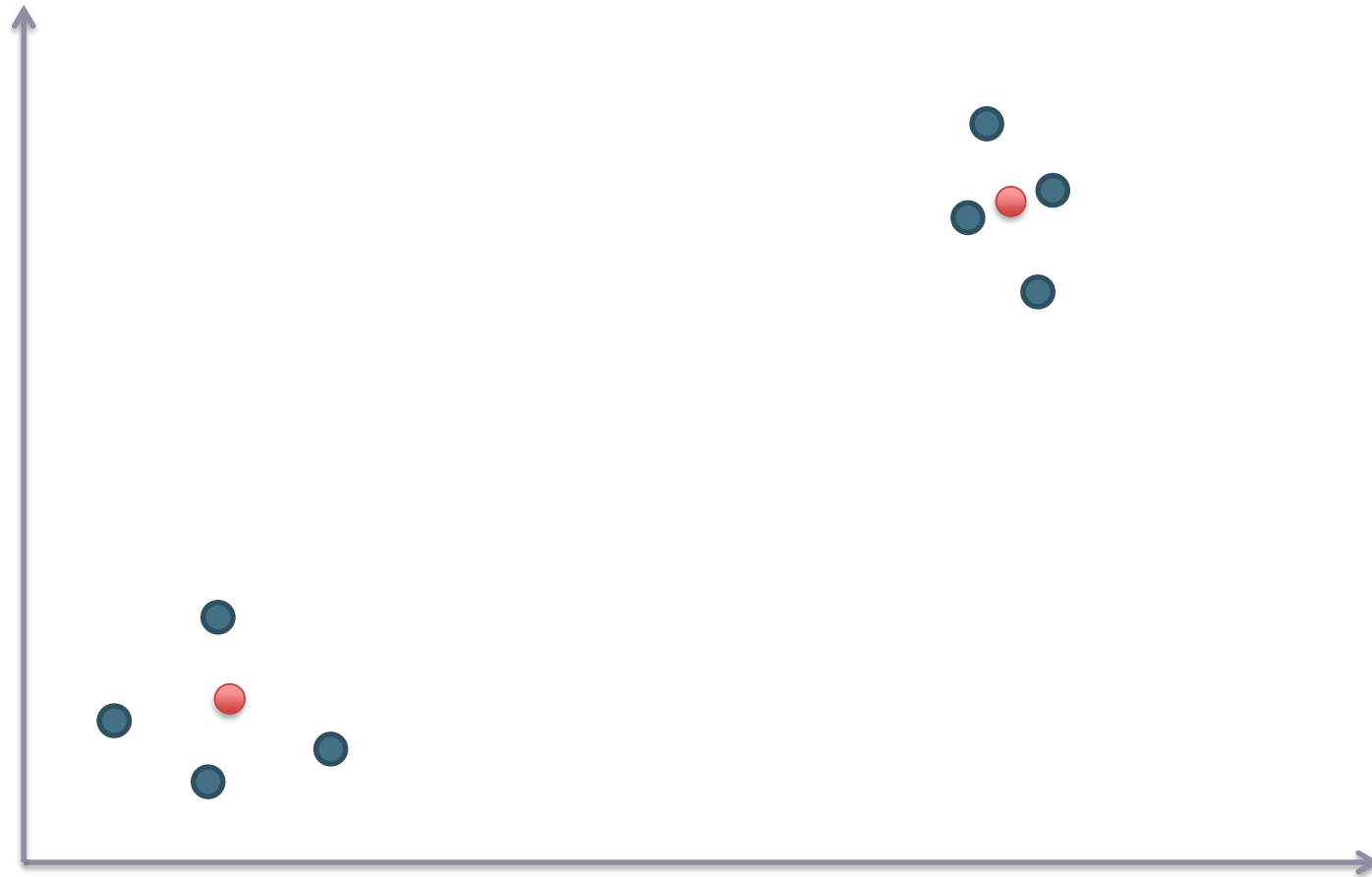
K-Means Clustering Example



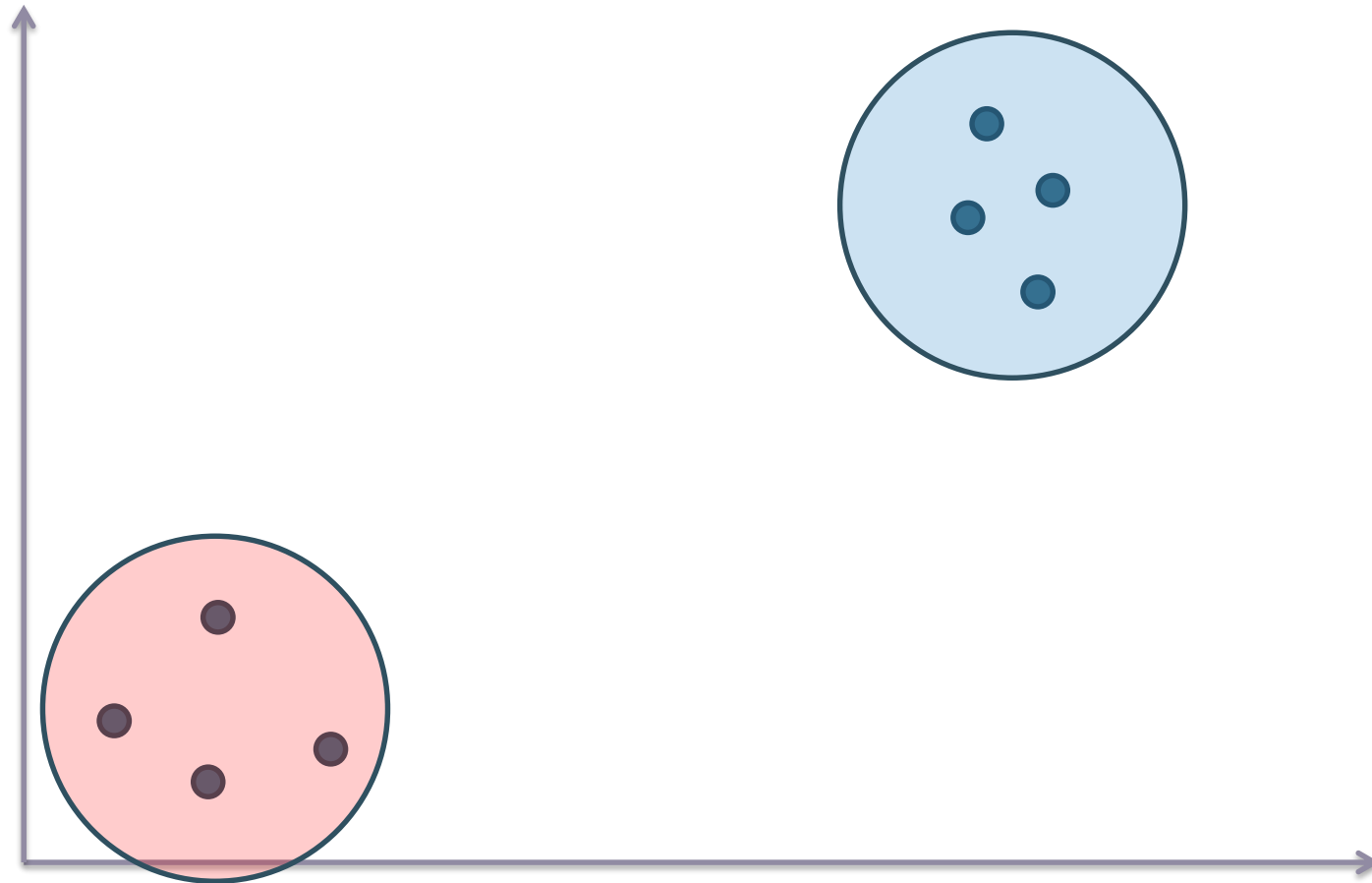
K-Means Clustering Example



K-Means Clustering Example

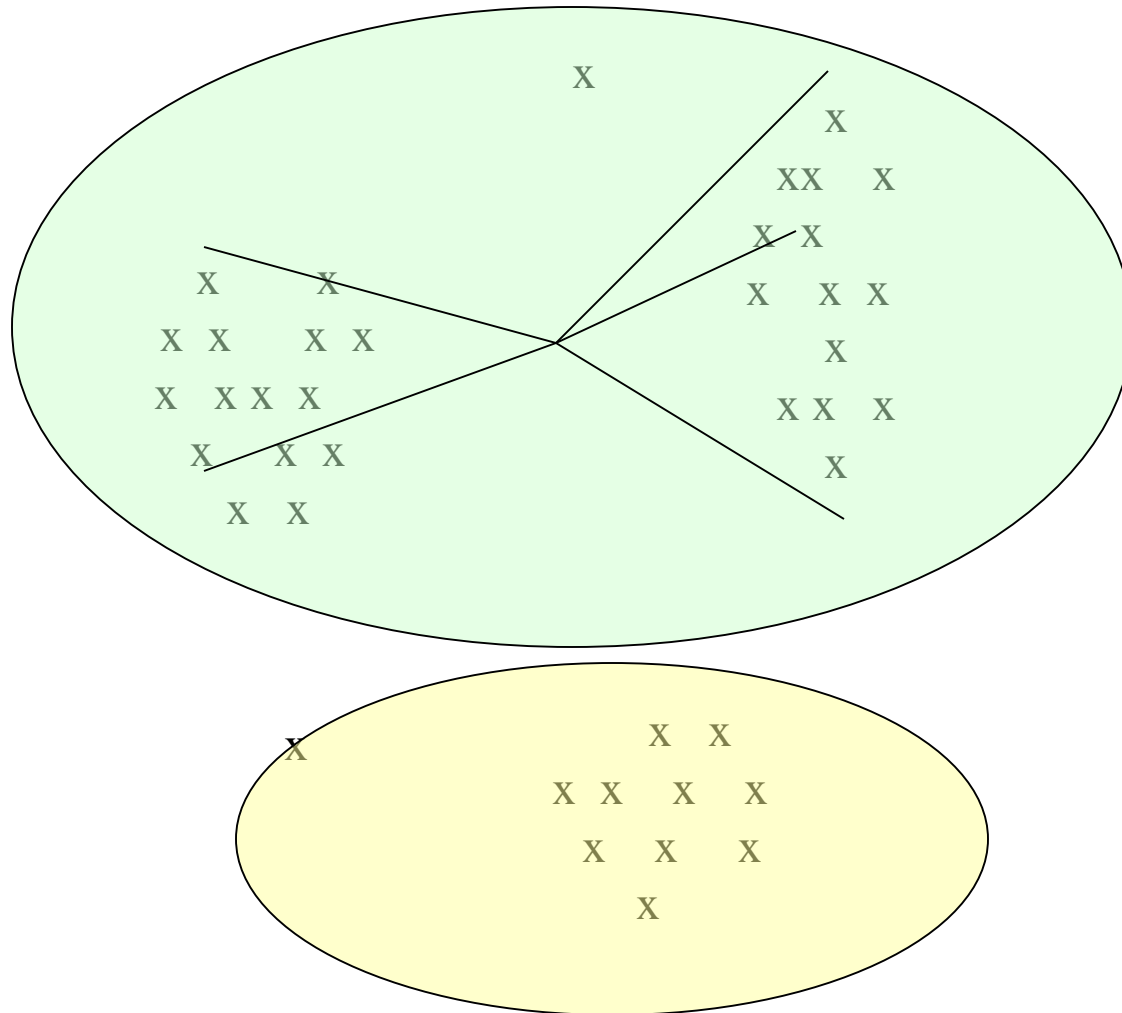


K-Means Clustering Example



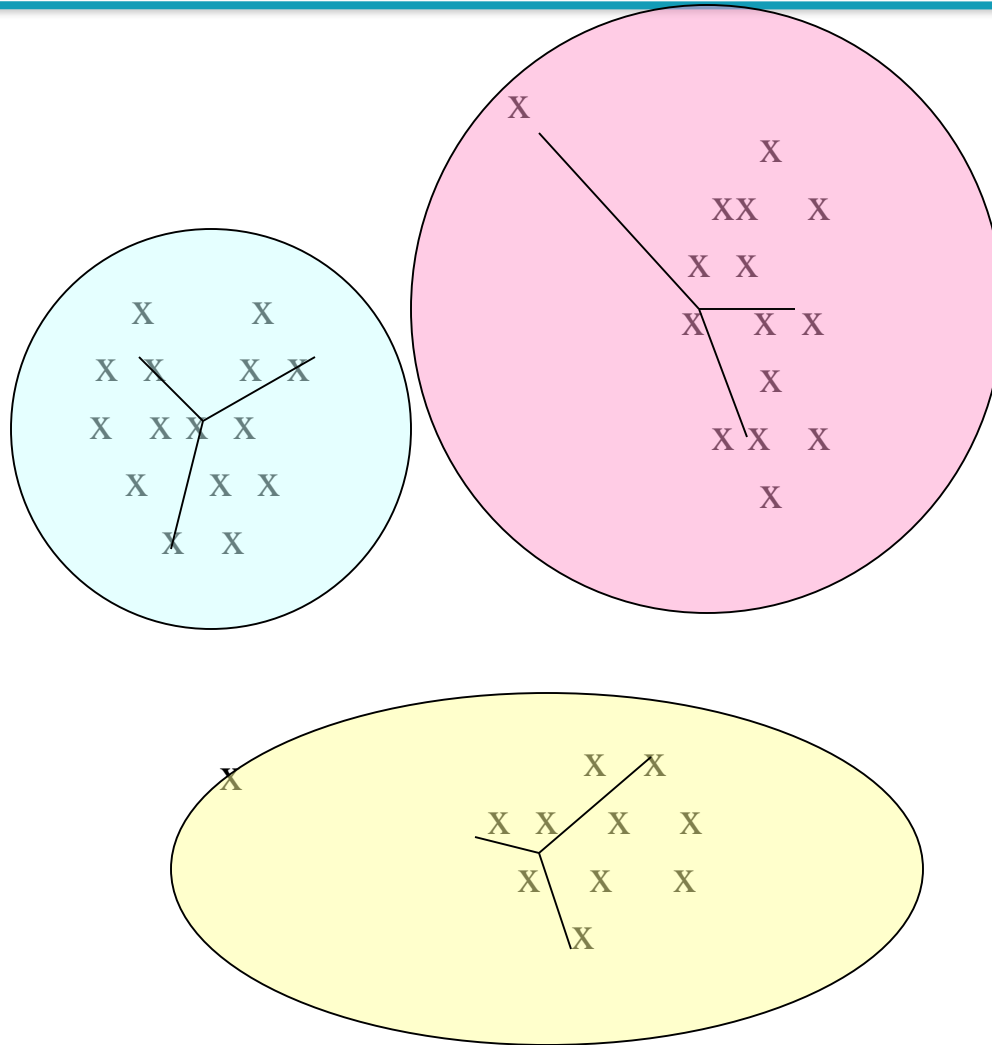
Example: Picking k

Too few;
many long
distances
to centroid.



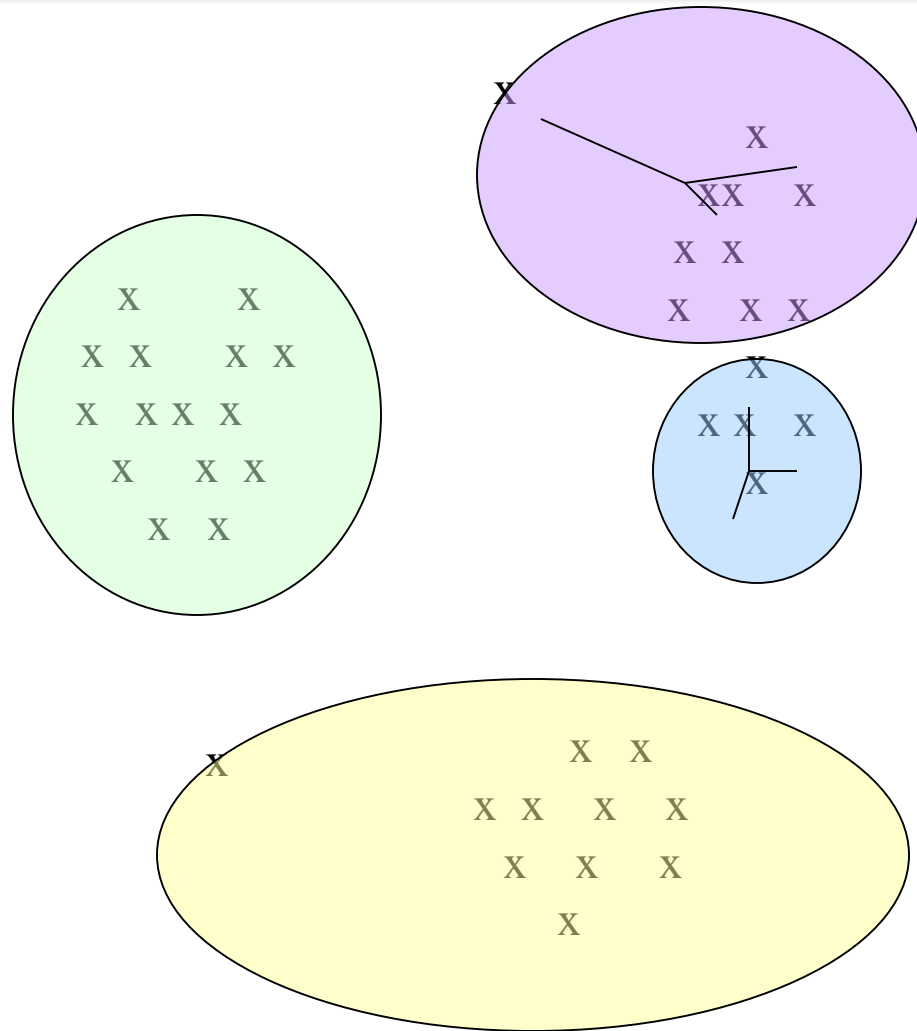
Example: Picking k

Just right;
distances
rather short.



Example: Picking k

Too many;
little improvement
in average
distance.



Convergence of K Means

- K-means converges to a fixed point in a finite number of iterations.
- Proof:
 - The sum of squared distances (RSS) decreases during reassignment.
 - (because each vector is moved to a closer centroid)
 - RSS decreases during recomputation.
 - There is only a finite number of clusterings
 - Thus: We must reach a fixed point.
- But we don't know how long convergence will take!
- If we don't care about a few docs switching back and forth, then convergence is usually fast (< 10-20 iterations).

Recomputation decreases average distance

- RSS = residual sum of squares (the “goodness” measure G)

$$\text{RSS}_k(\vec{v}) = \sum_{\vec{x} \in \omega_k} \|\vec{v} - \vec{x}\|^2 = \sum_{\vec{x} \in \omega_k} \sum_{m=1}^M (v_m - x_m)^2$$

$$\frac{\partial \text{RSS}_k(\vec{v})}{\partial v_m} = \sum_{\vec{x} \in \omega_k} 2(v_m - x_m) = 0$$

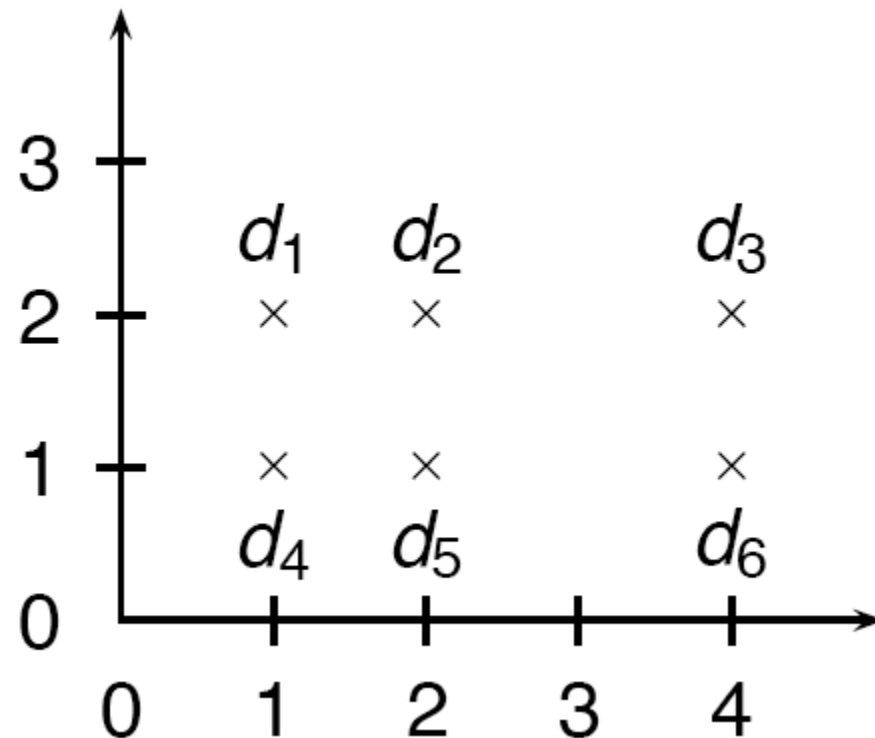
$$v_m = \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} x_m$$

- The last line is the componentwise definition of the centroid! We minimize RSS_k when the old centroid is replaced with the new centroid. RSS, the sum of the RSS_k , must then also decrease during recomputation.

Optimality of K -means

- Convergence does not mean that we converge to the optimal clustering!
- This is the great weakness of K -means.
- If we start with a bad set of seeds, the resulting clustering can be horrible.

Example of suboptimal clustering!!!!



- What is the optimal clustering for $K=2$?
- What happens when our seeds are: d_2, d_5 ?

Initialization of K -means

- Results can vary based on random seed selection.
- Some seeds can result in poor convergence rate, or convergence to sub-optimal clusterings.
 - Select good seeds using a heuristic (e.g., doc least similar to any existing mean)
 - Try out multiple starting points
 - Initialize with the results of another method.

How many clusters?

Hmm...

- **Either: Number of clusters K is given.**
 - Then partition into K clusters
 - K might be given because there is some external constraint. Example: You cannot show more than 10–20 clusters on a screen.
- **Or: Finding the “right” number of clusters is part of the problem.**
 - Given docs, find K for which an optimum is reached.
 - How to define “optimum”?
 - Why can’t we use RSS or average distance from centroid?

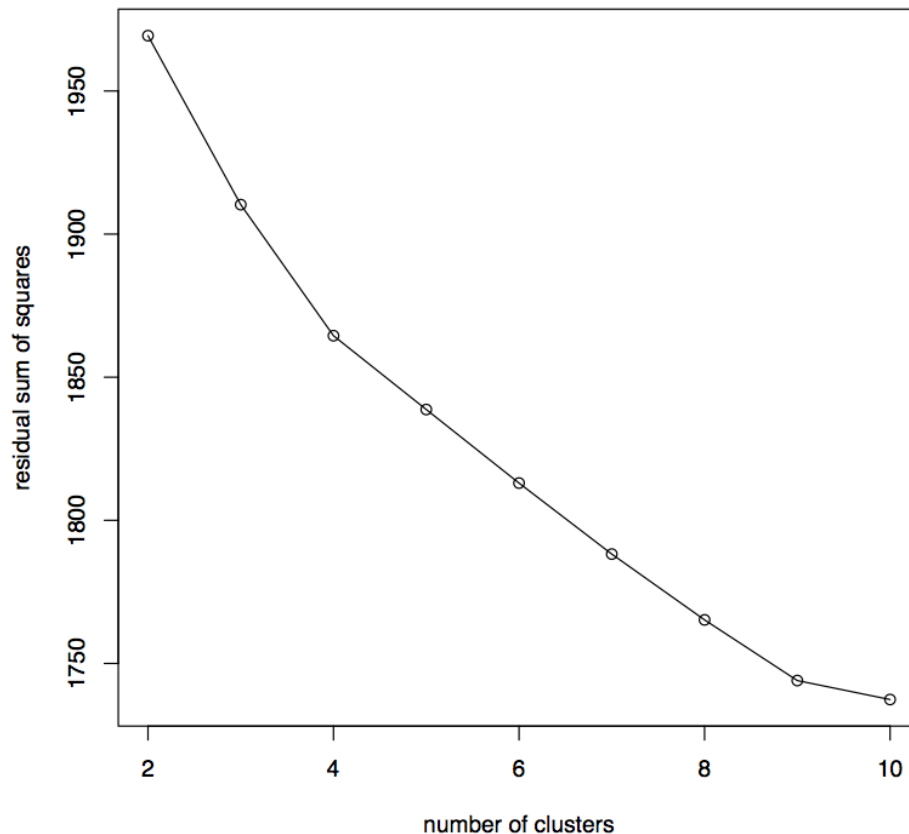
Simple objective function for K

- Basic idea:
 - Start with 1 cluster ($K = 1$)
 - Keep adding clusters (= keep increasing K)
 - Add a penalty for each new cluster
- Trade off cluster penalties against average squared distance from centroid
- Choose the value of K with the best tradeoff

Simple objective function for K

- Given a clustering, define the cost for a document as (squared) distance to centroid
- Define total **distortion** $RSS(K)$ as sum of all individual document costs (corresponds to average distance)
- Then: penalize each cluster with a cost λ
- Thus for a clustering with K clusters, total cluster penalty is $K\lambda$
- Define the total cost of a clustering as distortion plus total cluster penalty: $RSS(K) + K\lambda$
- Select K that minimizes $(RSS(K) + K\lambda)$
- Still need to determine good value for λ . . .

Finding the “knee” in the curve



Pick the number of clusters where curve “flattens”. Here: 4 or 9.

What Is the Problem of the K-Means Method?

- The k-means algorithm is sensitive to outliers !
 - Since an object with an extremely large value may substantially distort the distribution of the data
- K-Medoids: instead of finding a **centroid** for each cluster, use one of the points (the **medoid**) as the cluster “center”

