Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 18: Distance models (clustering)

Mar 30, 2016

Clustering

- Clustering (and unsupervised learning more generally) finds *structure* in data, using just X

X = a set of skyscrapers

Flat Clustering

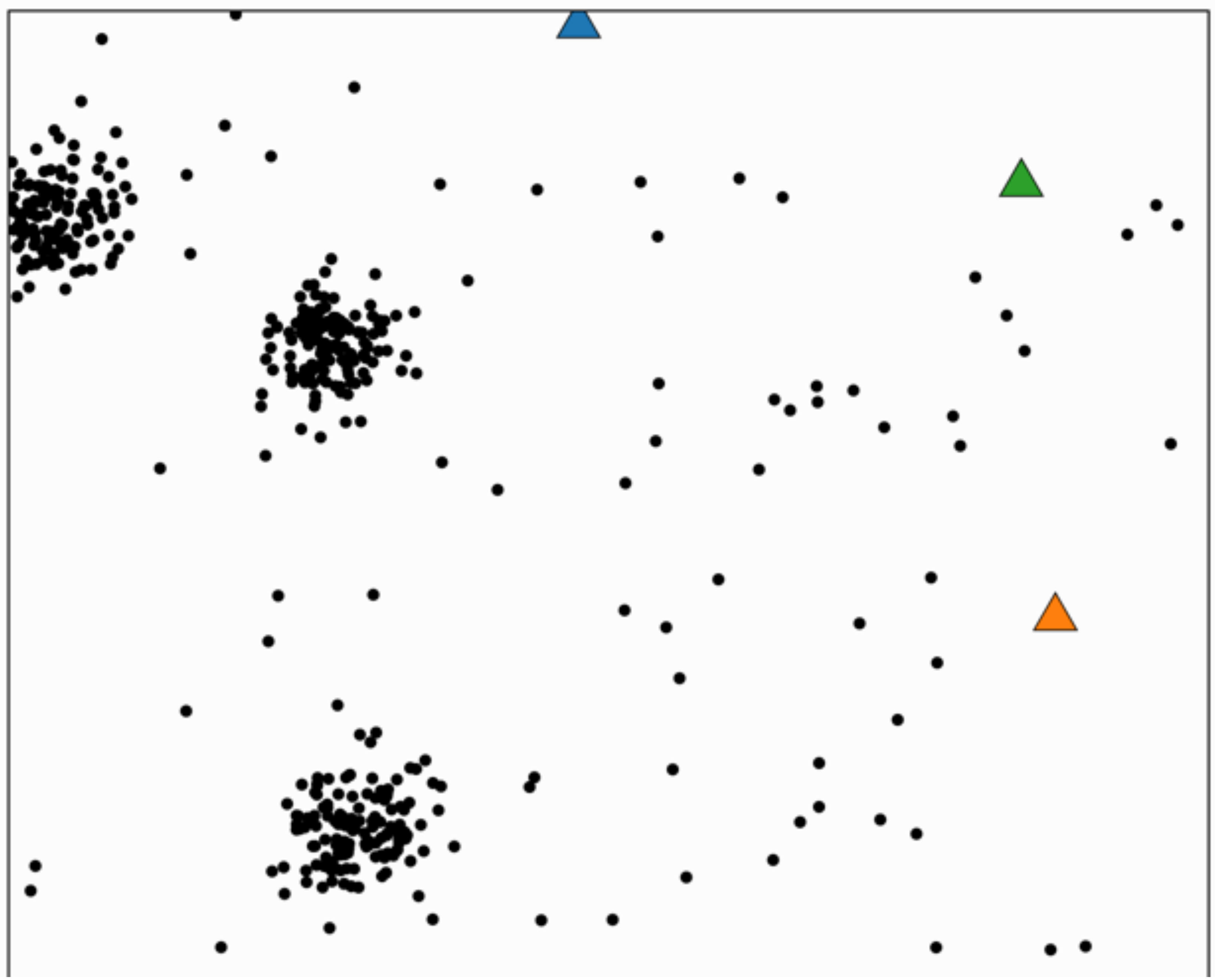
- Partitions the data into a set of *K* clusters

K-means

Algorithm 1 K-means

Data: training data $x \in \mathbb{R}^F$
Given some distance function $d(x, x') \to \mathbb{R}$
Select k initial centers $\{\mu_1, \ldots, \mu_k\}$
while not converged **do**
    for $i = 1$ to N **do**
        Assign x_i to $\arg\min_c d(x_i, \mu_c)$
    end for
    for $i = 1$ to K **do**
        $\mu_i = \frac{1}{D_i} \sum_{j=1}^{D_i} x_i$
    end for
end while

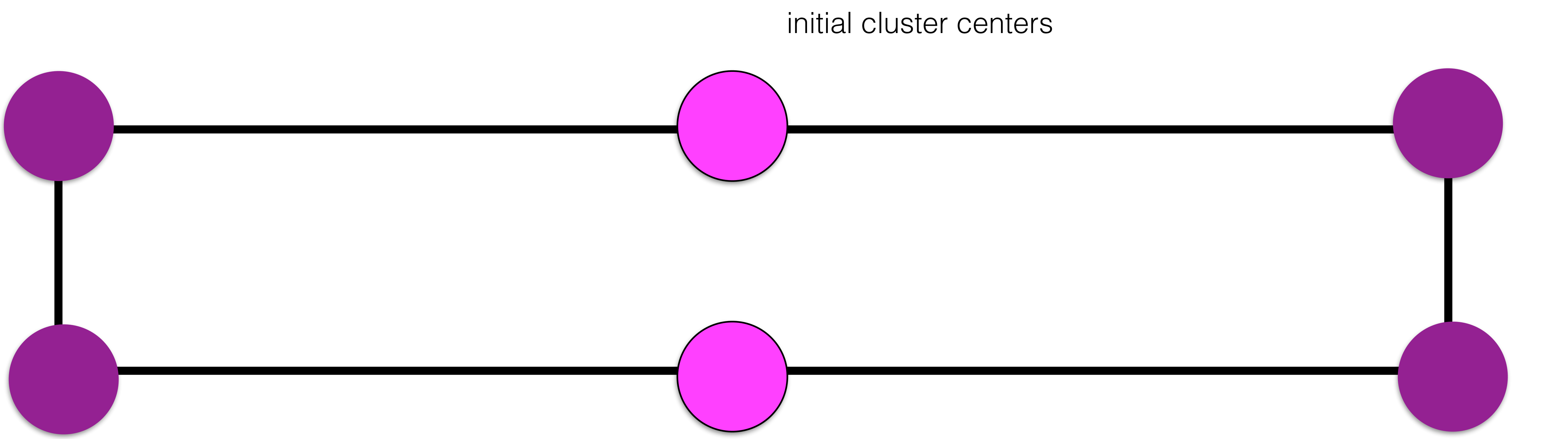
Visualizing K-Means Clustering



http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

Problems

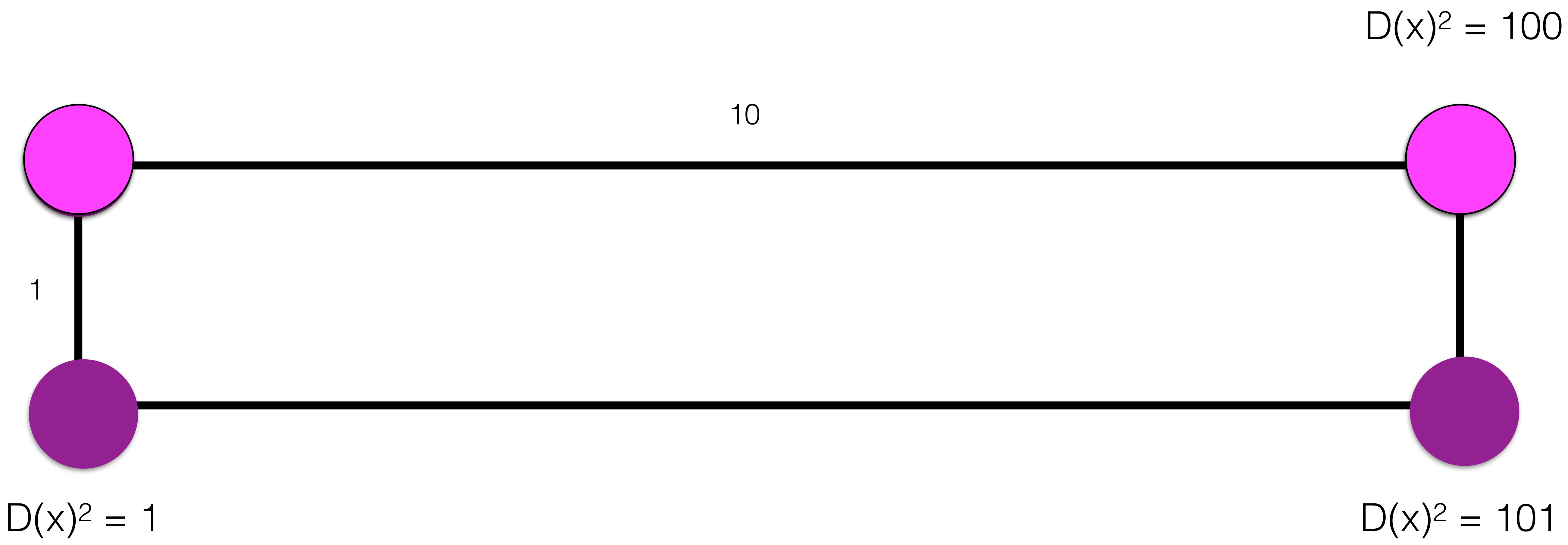
K-means



K-means++

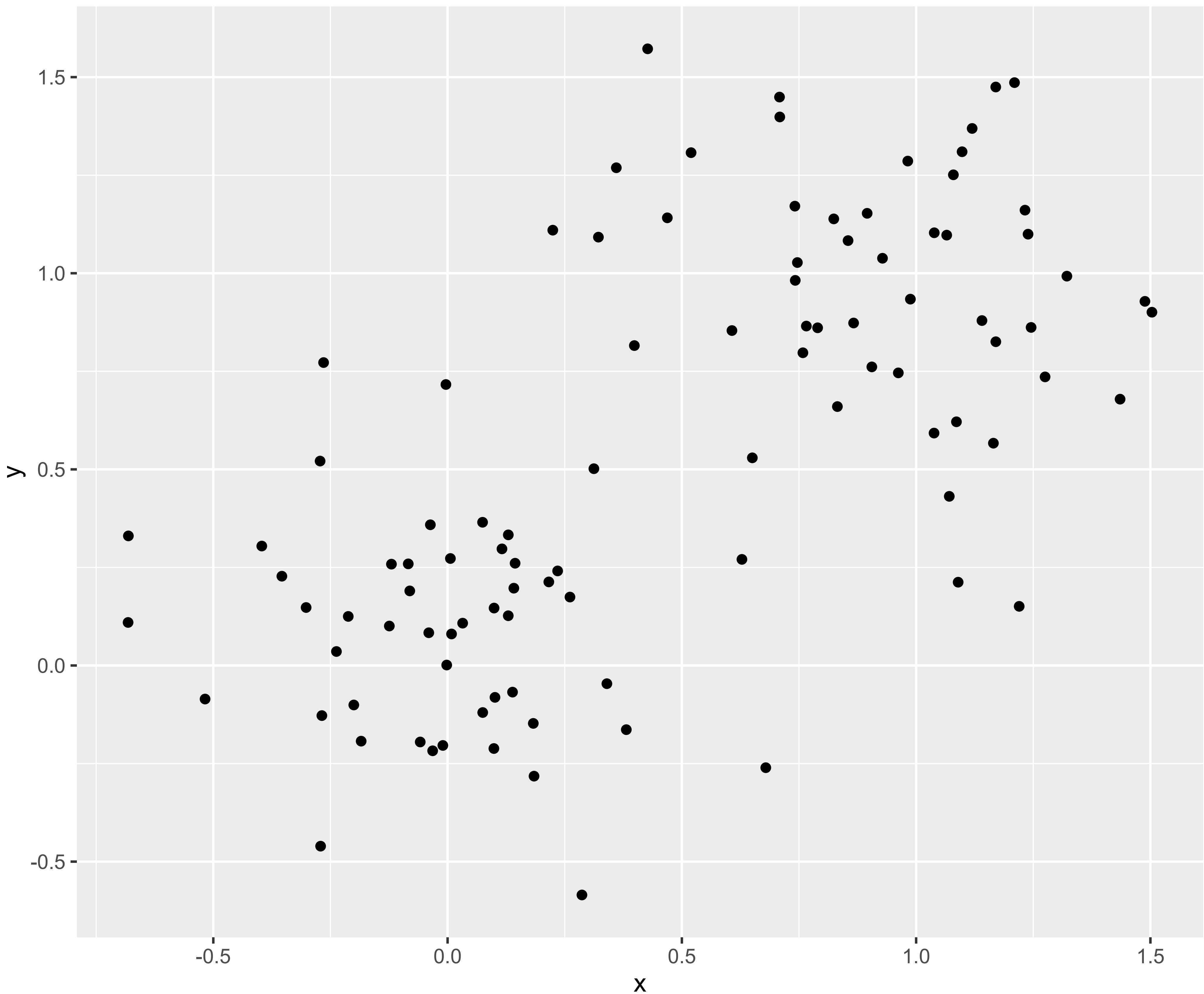
- Improved initialization method for K-means:
  - Choose data point at random as first center
  - For all other data points x, calculate the distance D(x) between x and the nearest cluster center
  - Choose new data point x as next center, with probability proportional to $D(x)^2$
  - Repeat until *K* centers are selected

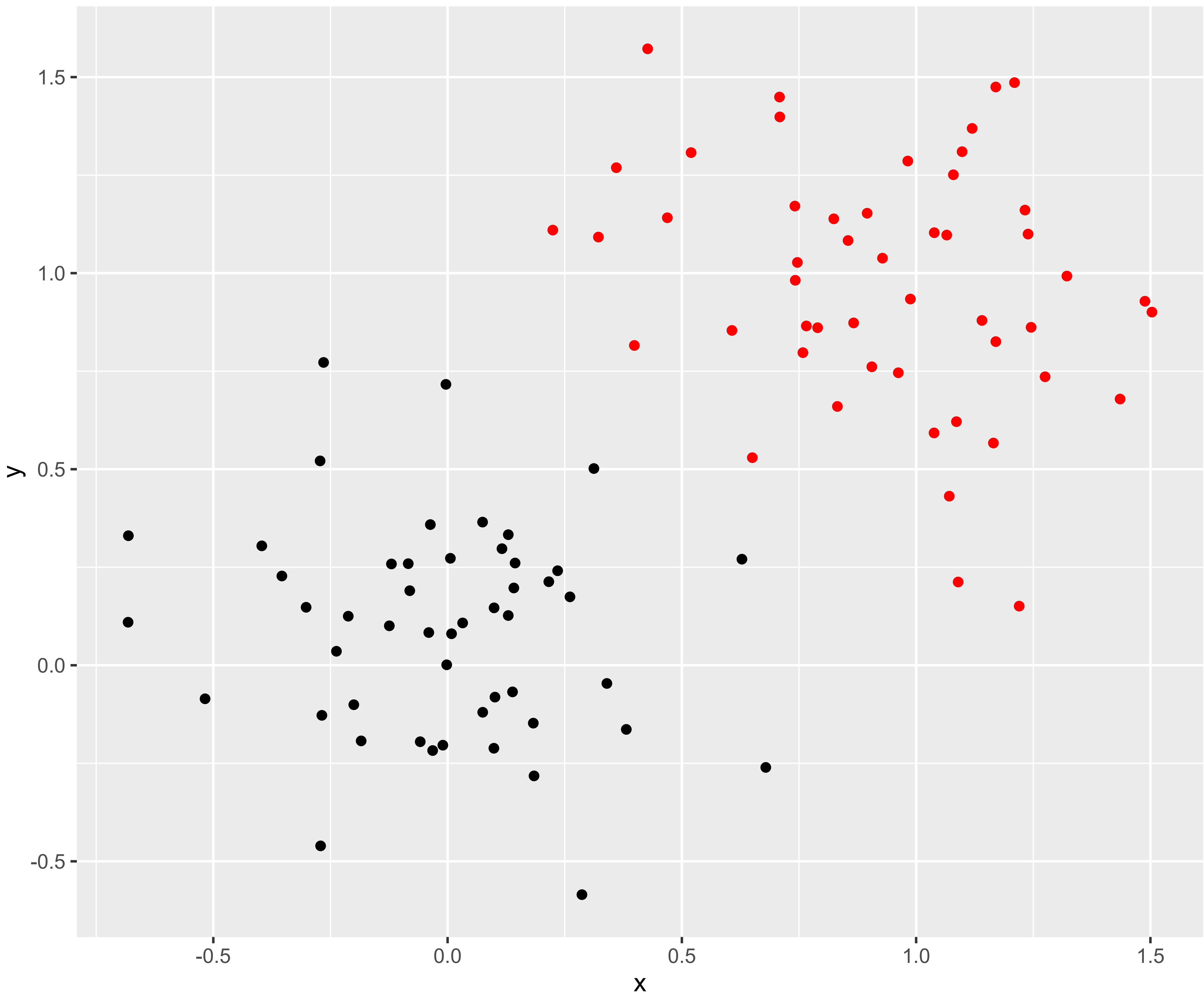
K-means++



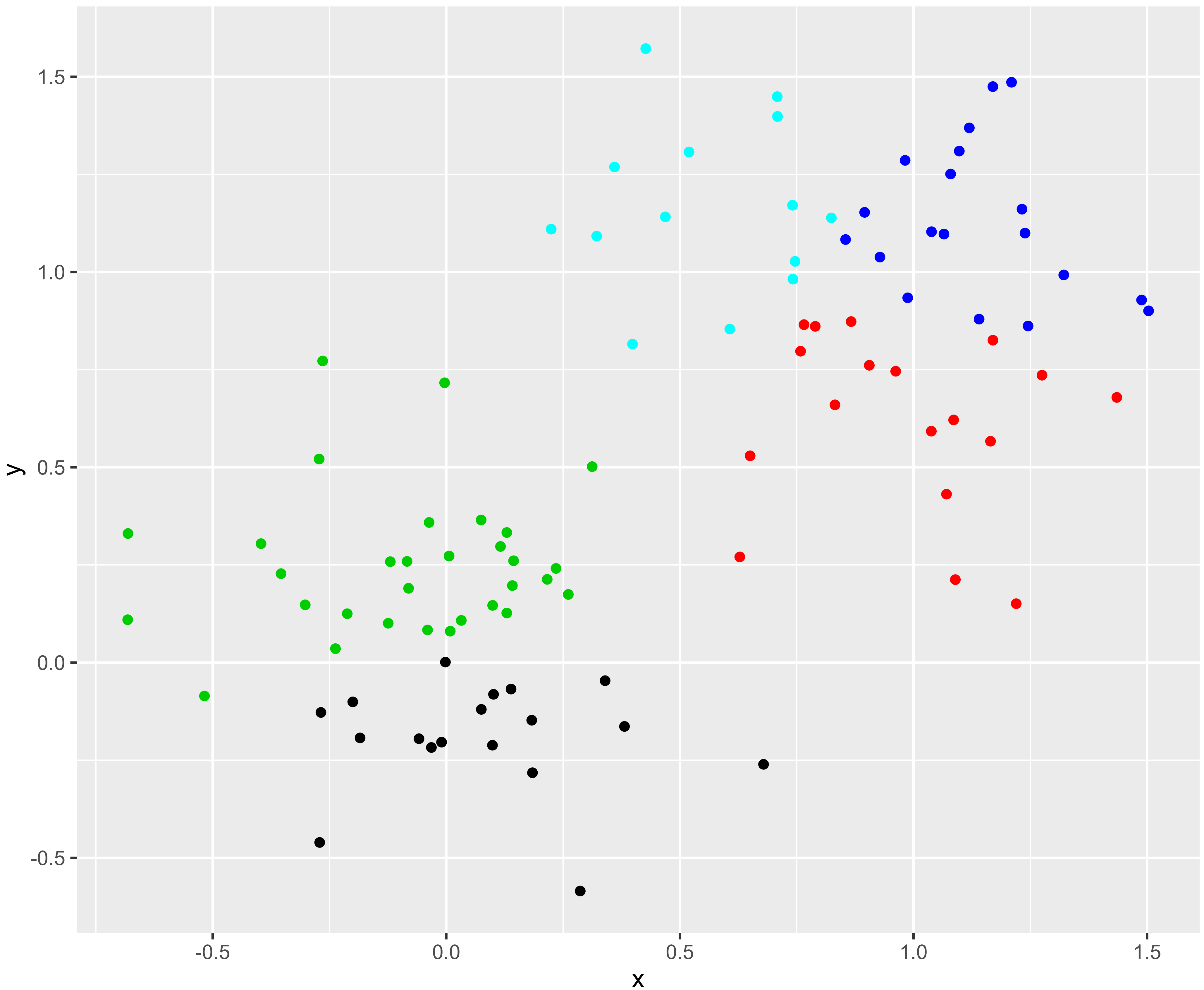
Choosing K

- how do we choose K?


1.5
1.0
0.5
0.0
-0.5
y
-0.5
0.0
0.5
1.0
1.5
x



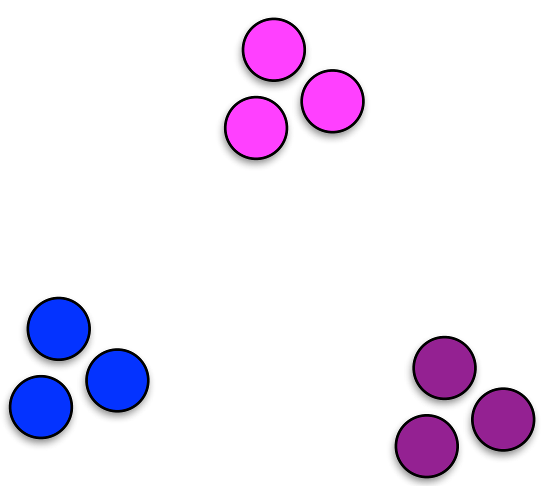
1.5
1.0
0.5
0.0
-0.5
y
-0.5
0.0
0.5
1.0
1.5
x



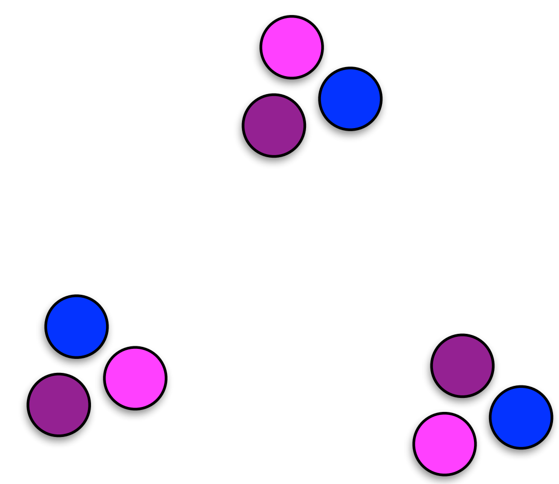
1.5
1.0
0.5
0.0
-0.5
y
-0.5
0.0
0.5
1.0
1.5
x

The “elbow”

Core idea: clusters should minimize the within-cluster variance



good



bad

The “elbow”

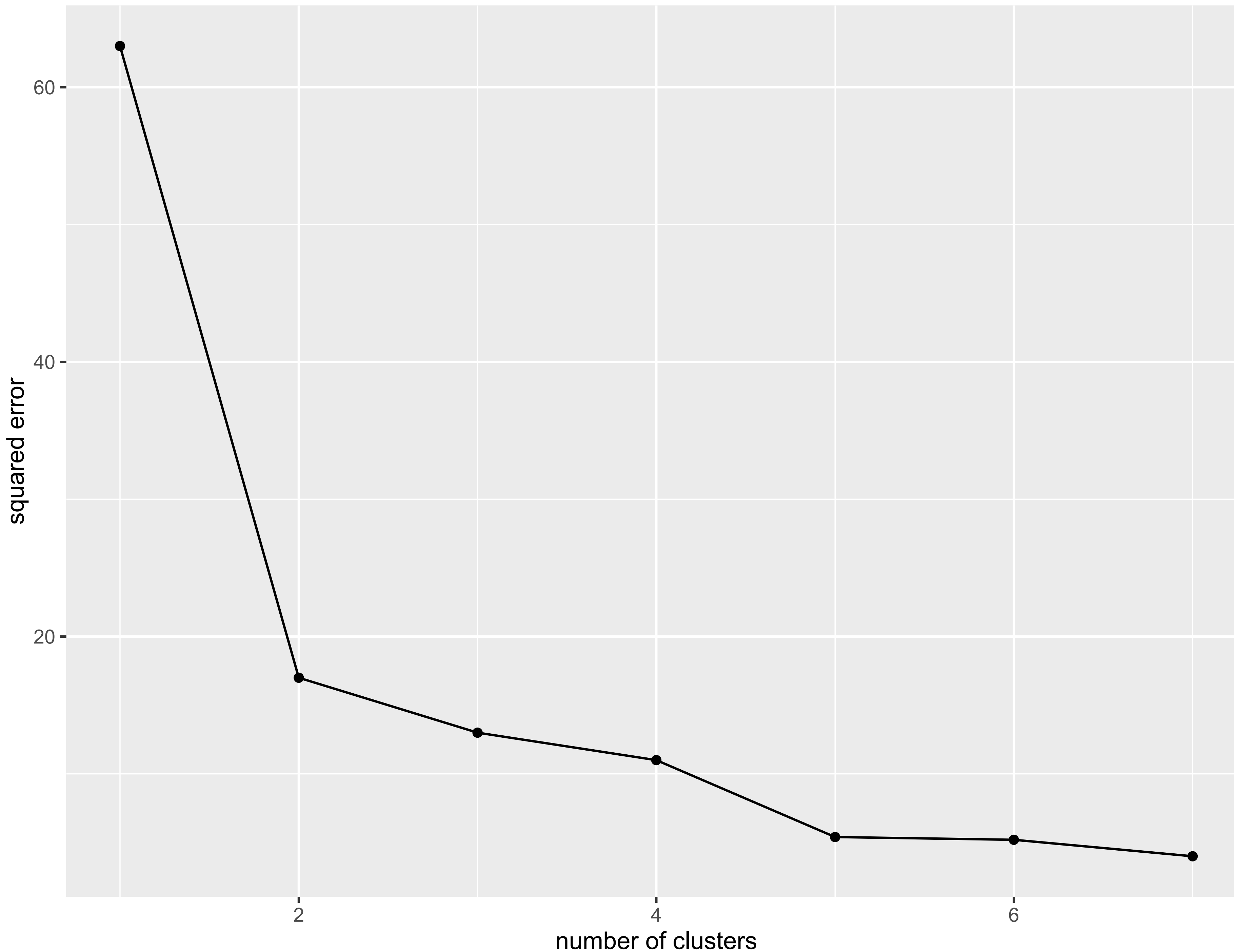
Core idea: clusters should minimize the within-cluster variance

within-cluster
sum of squares

$$\sum_{i=1}^{F}(x_i - \mu_i)^2$$

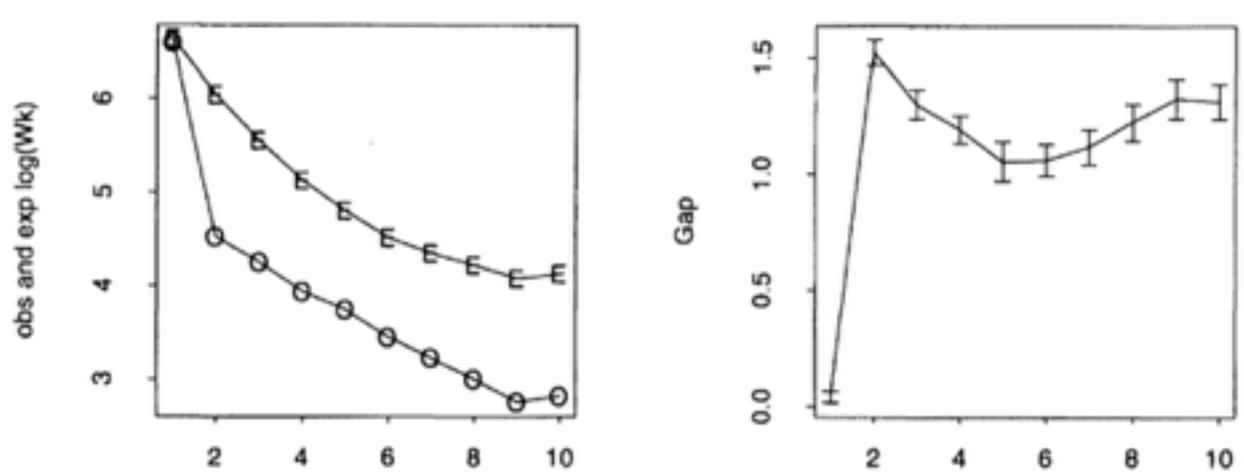
for each cluster

The “elbow”



Gap statistic

- How much variance should we expect to see for a given number of clusters?
- Choose number of clusters that maximizes the “gap” between the observed variance and the expected variance for a given K.



Tibshirani et al., “Estimating the number of clusters in a data set via the gap statistic”
http://web.stanford.edu/~hastie/Papers/gap.pdf

Kernelized K-means

Algorithm 1 K-means

Data: training data $x \in \mathbb{R}^F$
Given some distance function $d(x, x') \to \mathbb{R}$
Select k initial centers $\{\mu_1, \ldots, \mu_k\}$
while not converged **do**
    for $i = 1$ to N **do**
        Assign x_i to $\arg\min_c d(x_i, \mu_c)$
    end for
    for $i = 1$ to K **do**
        $\mu_i = \frac{1}{D_i} \sum_{j=1}^{D_i} x_i$
    end for
end while

Kernelized K-means

$$|\phi(x_i) - \phi(\mu_c)|^2$$

we can kernlize k-means by replacing the original data point x with Φ(x)

$$\left|\phi(x_i) - \frac{\sum_{j=1}^{D_c} \phi(x_j)}{D_c}\right|^2$$

$$|\phi(x_i) - \phi(\mu_c)|^2 \quad \rightarrow \quad \left|\phi(x_i) - \frac{\sum_{j=1}^{D_c} \phi(x_j)}{D_c}\right|^2$$

$$\phi(x_i)\phi(x_i) - \frac{2\phi(x_i)\sum_{j=1}^{D_c} \phi(x_j)}{D_c} + \frac{\sum_{j=1}^{D_c} \phi(x_j) \sum_{k=1}^{D_c} \phi(x_k)}{D_c^2}$$

$$\phi(x_i)\phi(x_i) - \frac{2\sum_{j=1}^{D_c} \phi(x_i)\phi(x_j)}{D_c} + \frac{\sum_{j=1}^{D_c} \sum_{k=1}^{D_c} \phi(x_j)\phi(x_k)}{D_c^2}$$

$$\kappa(x_i, x_i) - \frac{2\sum_{j=1}^{D_c} \kappa(x_i, x_j)}{D_c} + \frac{\sum_{j=1}^{D_c} \sum_{k=1}^{D_c} \kappa(x_j, x_k)}{D_c}$$

Kernelized K-means

Algorithm 3 Kernelized K-means

Data: training data $x \in \mathbb{R}^F$
Given some kernelized distance function $\kappa(x, x') \to \mathbb{R}$
while not converged **do**
    for $i = 1$ to N **do**
        Assign x_i to:
$\arg\min_c \kappa(x_i, x_i) - \frac{2\sum_{j=1}^{D_c} \kappa(x_i, x_j)}{D_c} + \frac{\sum_{j=1}^{D_c} \sum_{k=1}^{D_c} \kappa(x_j, x_k)}{D_c}$
    end for
end while

Hierarchical clustering

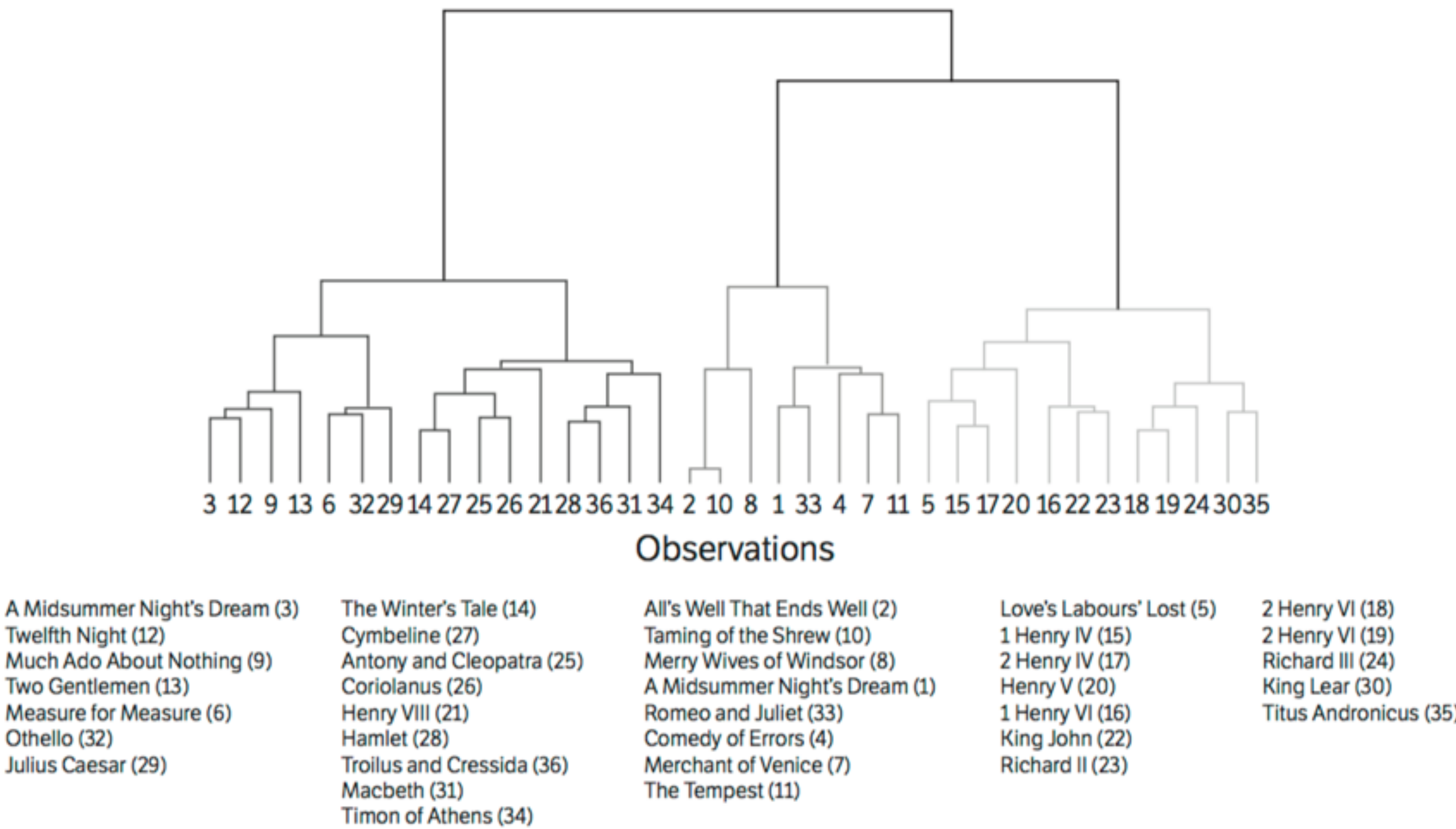
Core idea: build a binary tree of a set of data points by repeatedly merging the two most similar elements

Hierarchical clustering

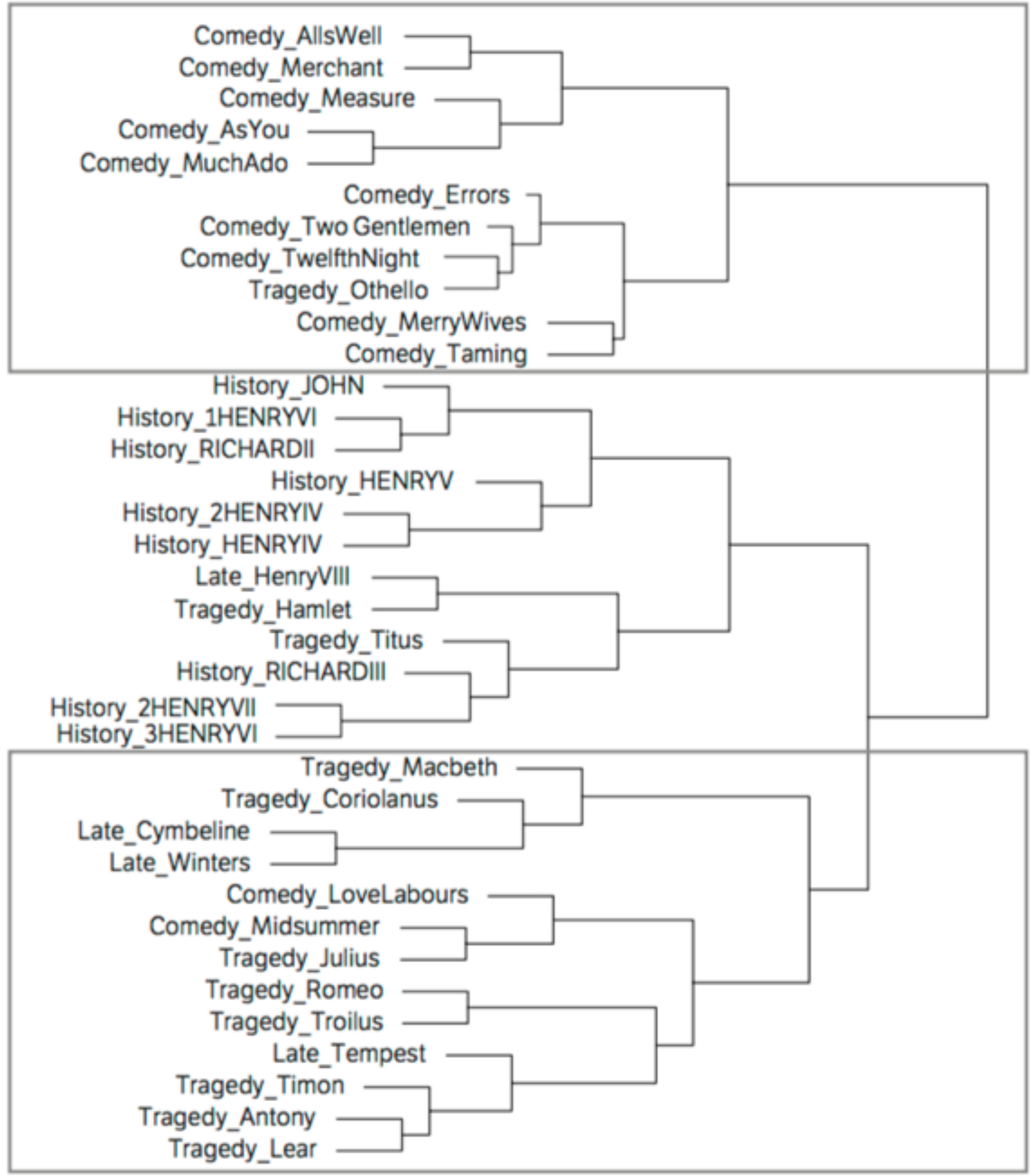
Algorithm 1 Hierarchical agglomerative clustering

Data: N training data points $x \in \mathbb{R}^F$
Let X denote a set of objects x
Given some linkage function $d(X, X') \to \mathbb{R}$
Initialize clusters $\mathcal{C} = \{C_1, \ldots, C_N\}$ to singleton data points
while data points not in one cluster **do**
    Identify X, Y as clusters with smallest linkage function among clusters in $\mathcal{C}$
    Create new cluster $Z = X \cup Y$
    remove X, Y from $\mathcal{C}$
    add Z to $\mathcal{C}$
end while

Hierarchical clustering



Allison et al. 2009



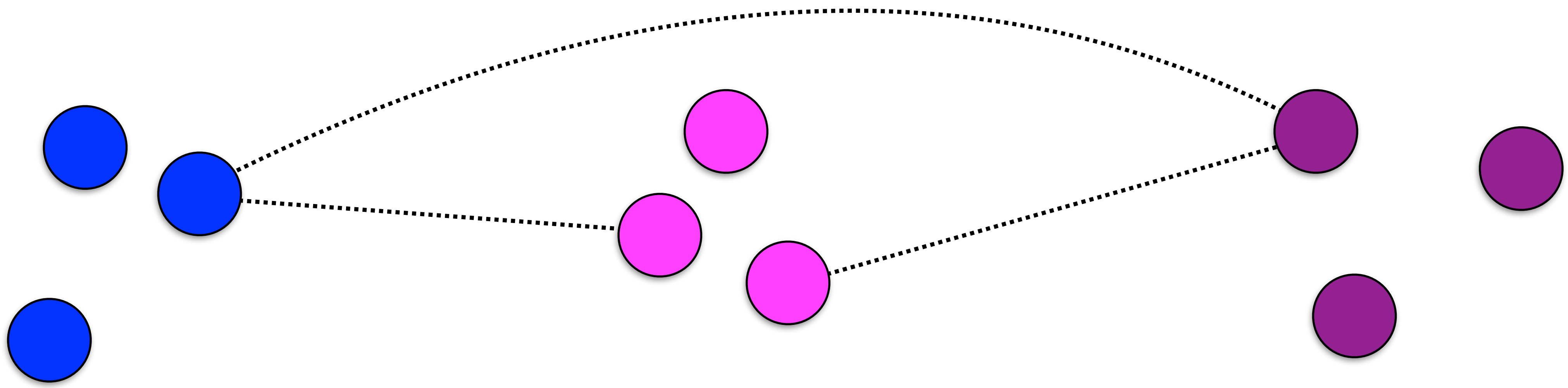
Allison et al. 2009

Hierarchical clustering

We know how to compare data points with distance metrics.

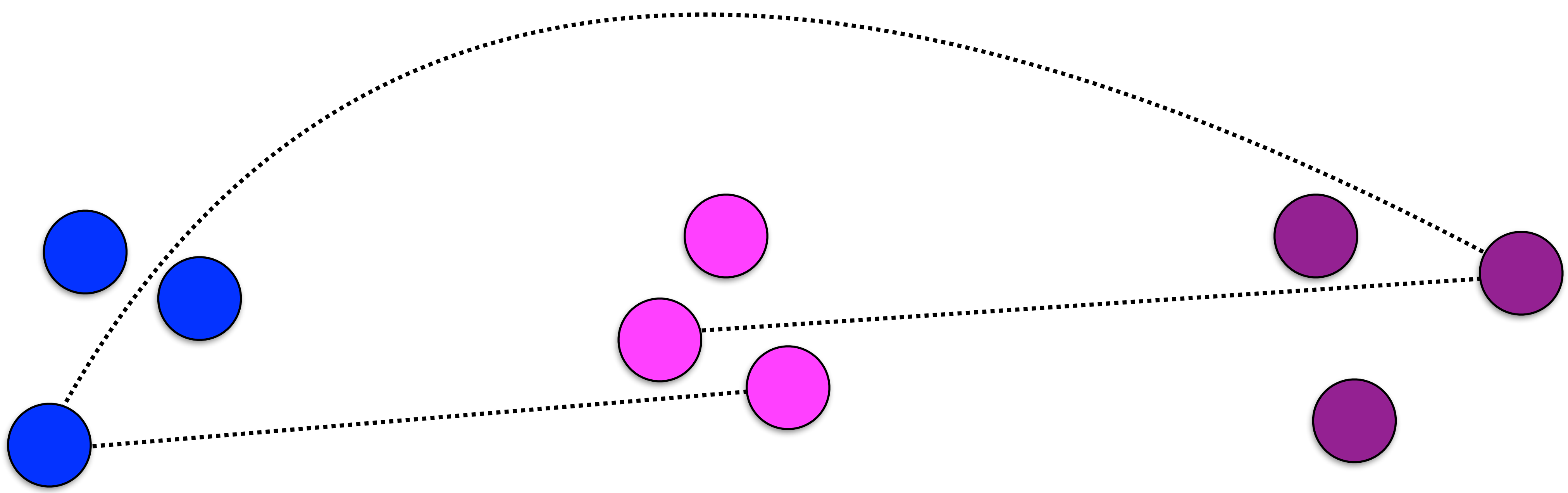
How do we compare sets of data points?

Single linkage



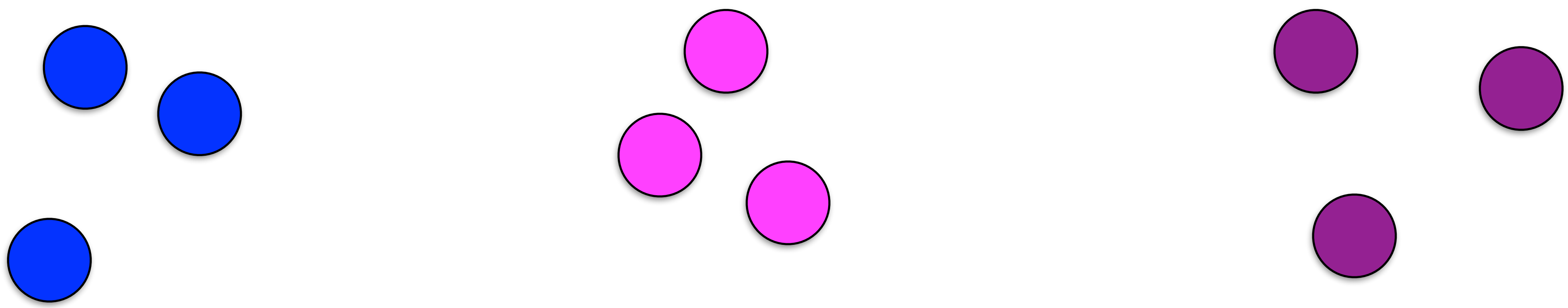
$$\min_{x \in A,\, y \in B} \mathrm{Dis}(x, y)$$

Complete linkage

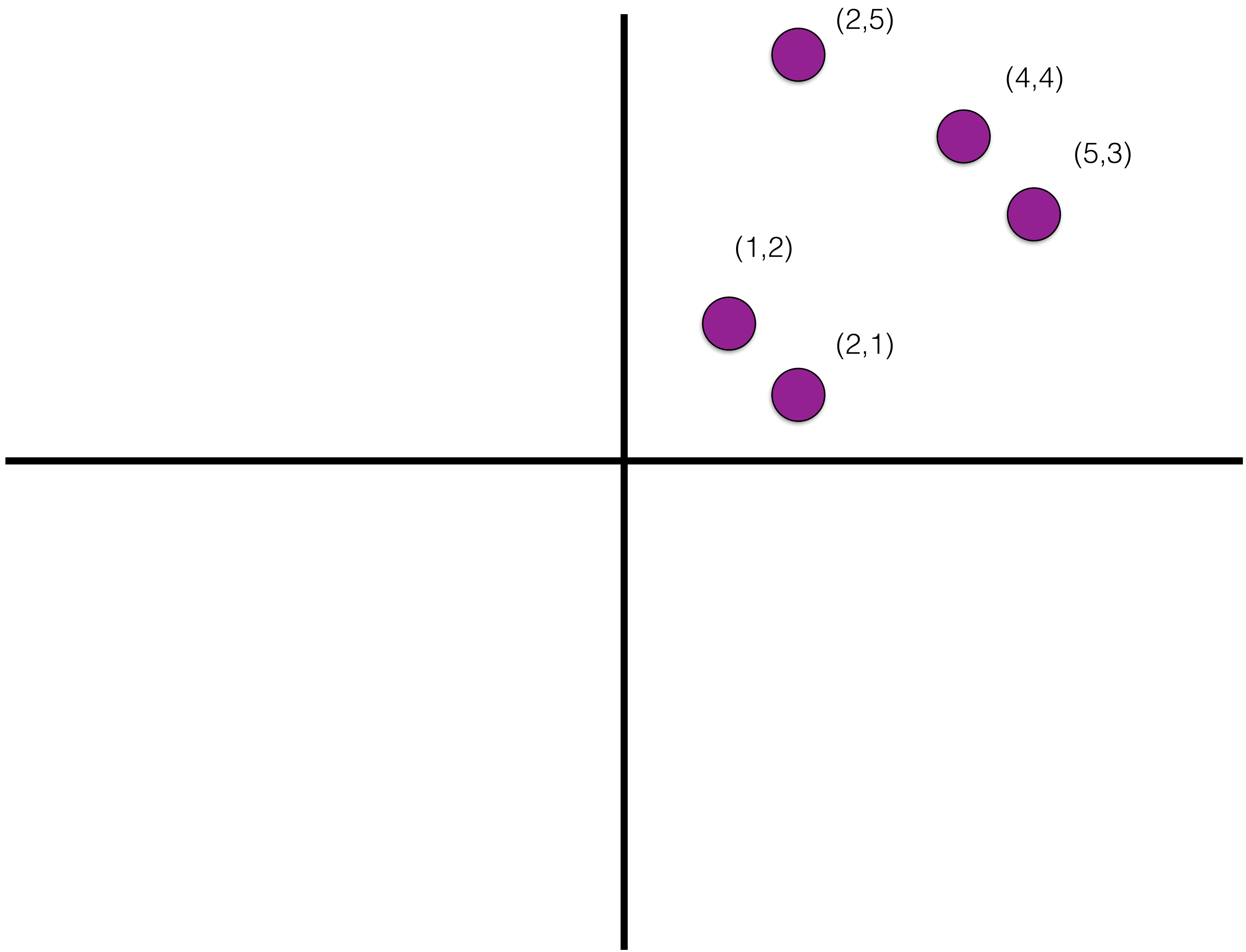


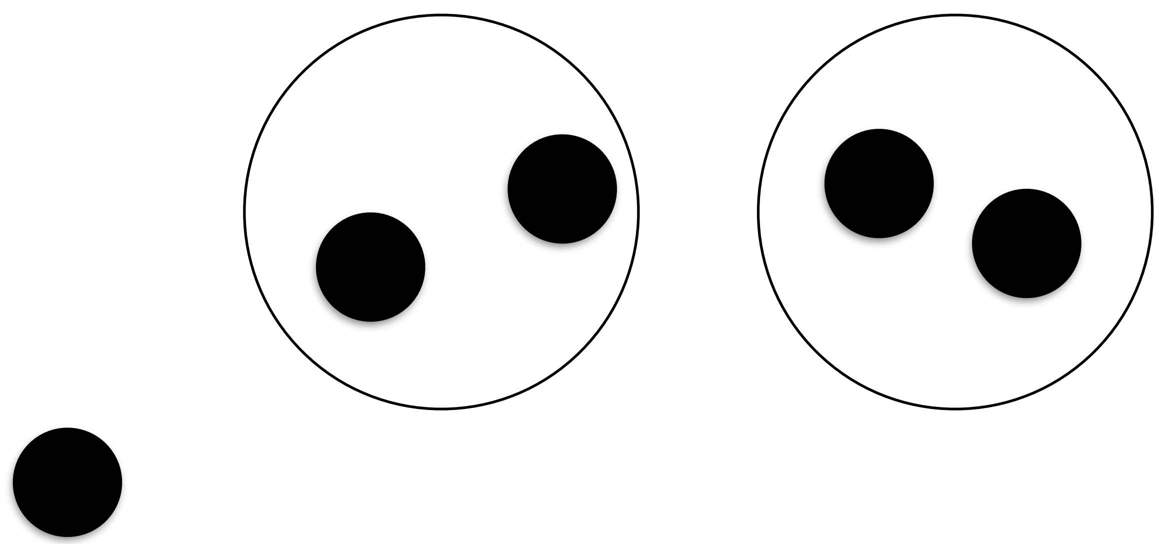
$$\max_{x \in A,\, y \in B} \mathrm{Dis}(x, y)$$

Average linkage



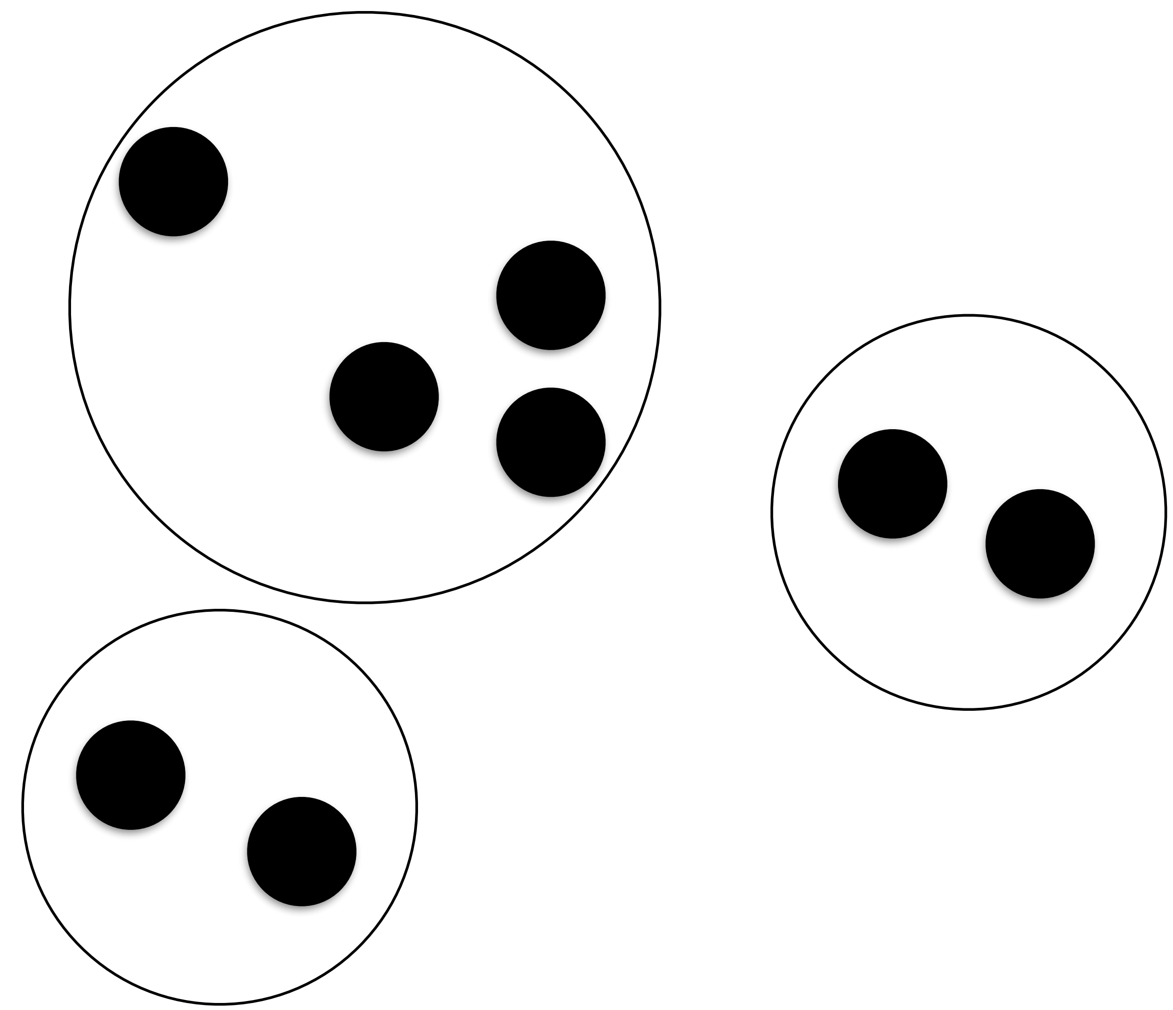
$$\frac{\sum_{x \in A,\, y \in B} \mathrm{Dis}(x, y)}{|A| \times |B|}$$


(2,5)
(4,4)
(5,3)
(1,2)
(2,1)



Single linkage may link bigger clusters together before outliers

Complete linkage



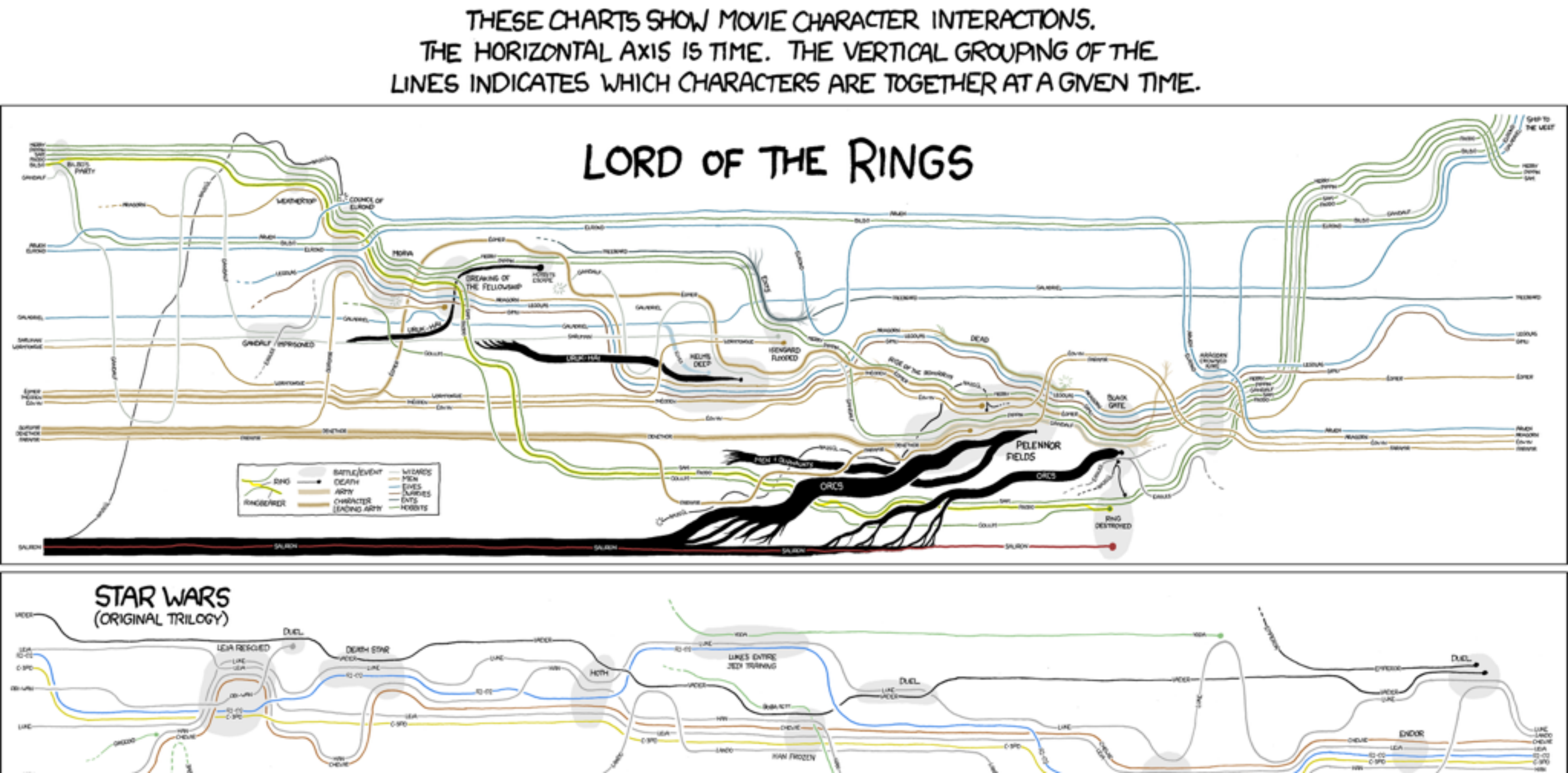
Complete linkage may *not* link close clusters together because of outliers

Digital Humanities

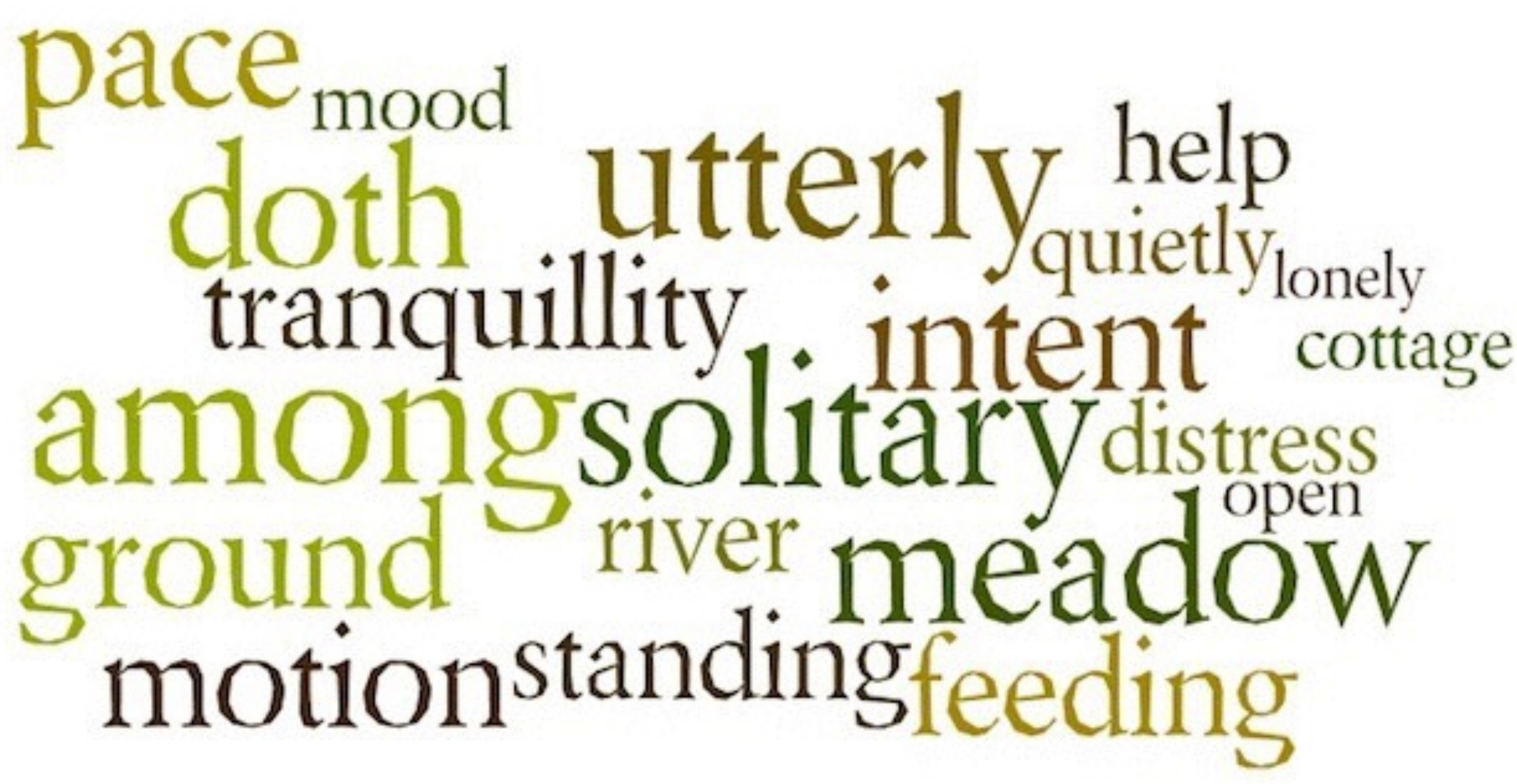
- Marche (2012), Literature Is not Data: Against Digital Humanities

- Underwood (2015), Seven ways humanists are using computers to understand text.

Text visualization



Characteristic vocabulary



Characteristic words by William Wordsworth (in comparison to other contemporary poets) [Underwood 2015]

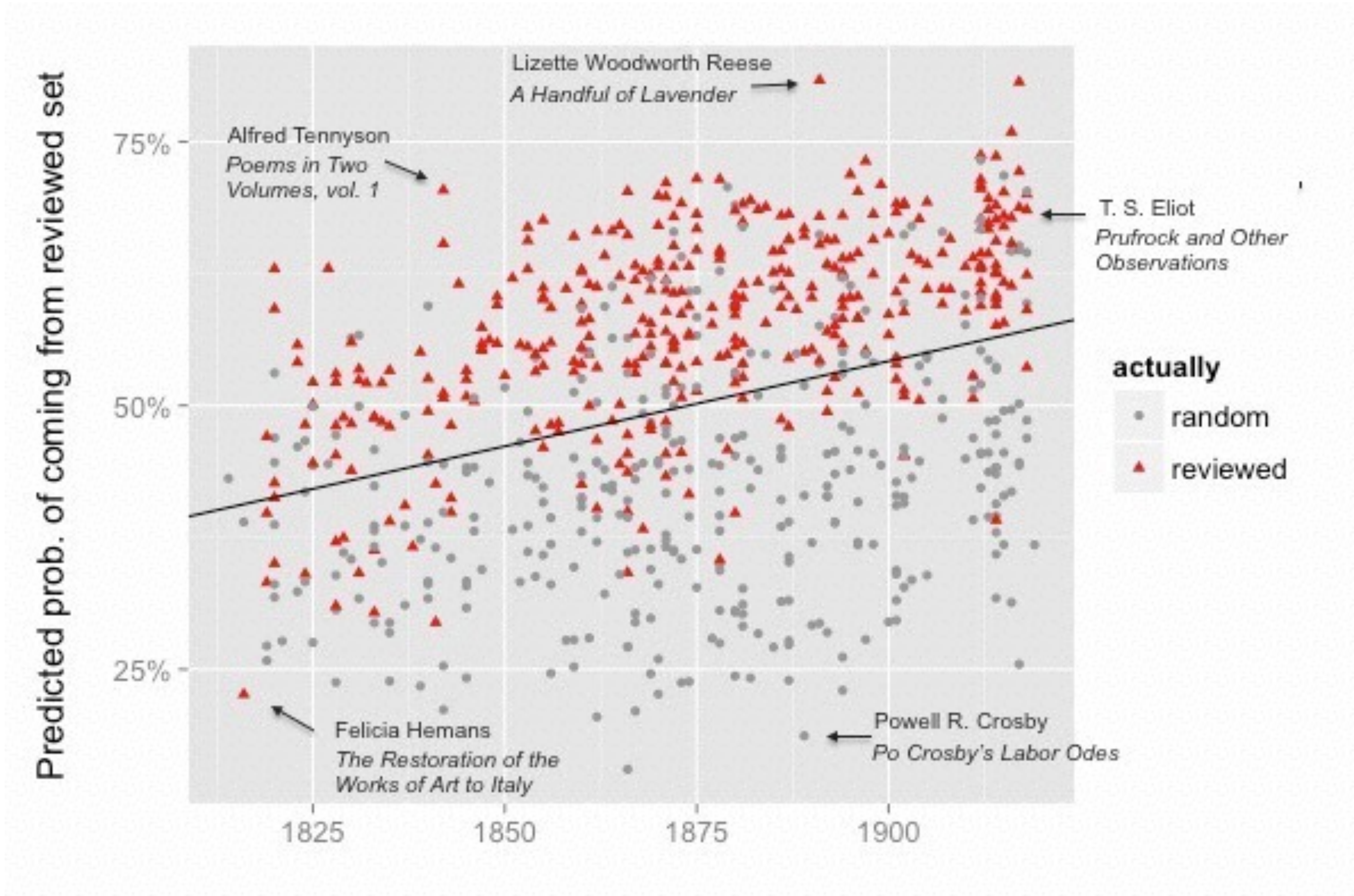
Finding and organizing texts

- e.g., finding all examples of a complex literary form (Haiku).
- Supplement traditional searches: book catalogues, search engines.

Modeling literary forms

- What features of a text are predictive of Haiku?

Modeling social boundaries



Predicting reviewed texts [Underwood and Sellers (2015)]

Unsupervised modeling

A Topic Model of Literary Studies Journals | Overview | Topic | Article | Word | Bibliography | Word index | Settings | About

List | Grid | Years

click a column label to sort; click a row for more about a topic

topic	1889—2013	top words	proportion of corpus
1		see both own view role university further account critical particular	2.5%
2		other both two form same even each part experience process	2.6%
3		old beowulf english ic mid swa pe poet ond grendel	0.3%
4		law legal justice rights laws right state court case common	0.3%
5		voltaire rousseau mme corneille french diderot moliere france lettres paris	0.3%
6		shakespeare play hamlet scene king plays elizabethan lear speech see	0.4%
7		like other voice even speech same words much way well	1.1%
8		other derrida even first like same two text man way	0.9%
9		new public city world urban space everyday american york life	0.4%
10		own power text form subject order discourse becomes authority figure	2.3%

- Allison et al., “Quantitative Formalism: an Experiment”

DocuScope

Dictionary mapping ngrams to classes

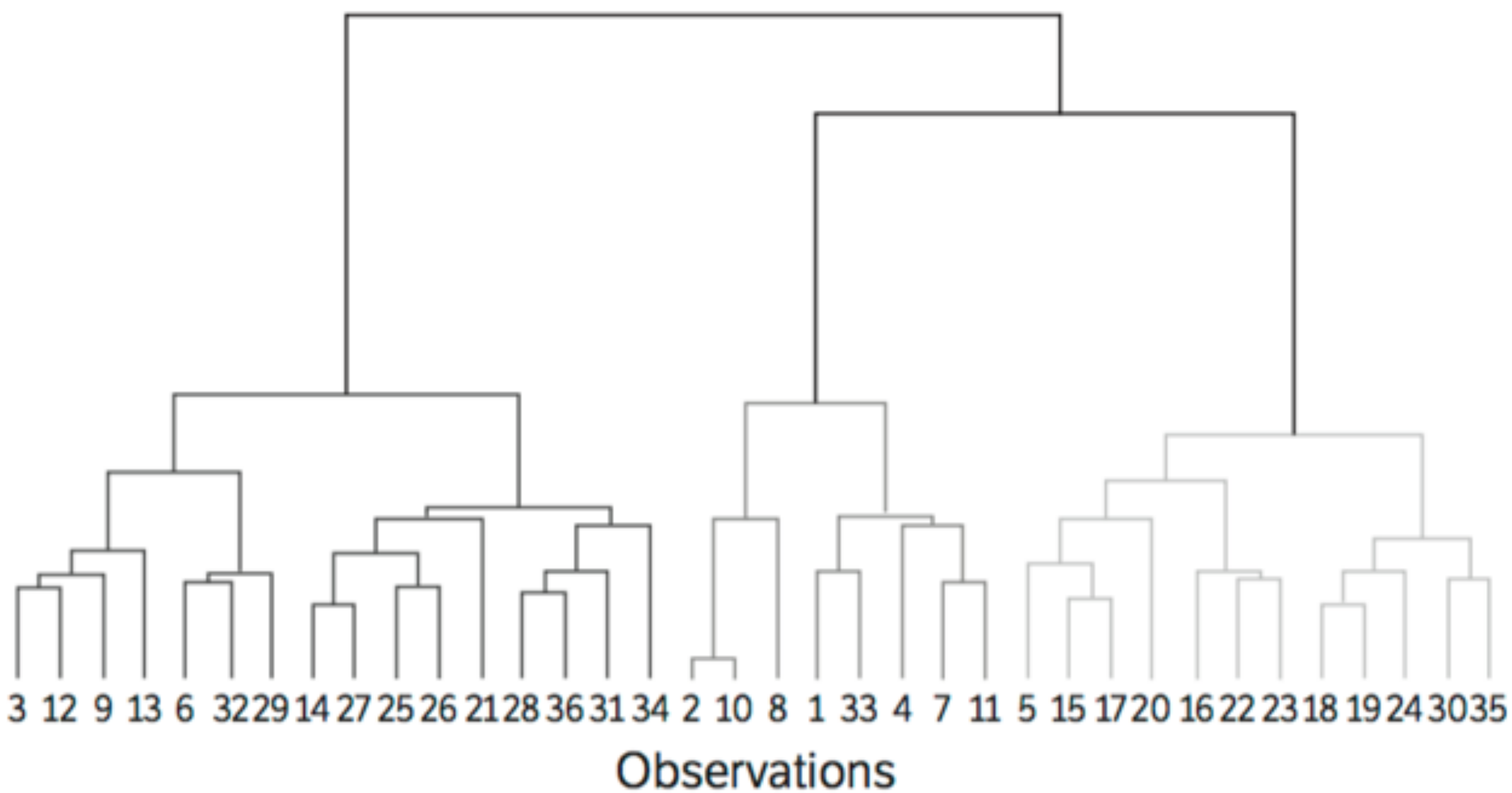
First Person	Numbers	Positivity
about me	six-wheeled	perpetual adorations
about my	275 degrees	mated with
am	three-card loo	hugging yourself
I	695	striking responsive cord
I'd	four-ply	wassailing
I'll	half-way	plucked up your spirits
I'm	three parts	offers ourselves
I for one	eight-member	promotive of
ich	third-world	enshrining
ich dien	3,5	devotes yourself
me	half-and-half measures	music lover
mea	8,3	delectated
meum	half-reclining	recharging my batteries
mine	26	recommends you for
my	634	shadow of your smile
myself	five-rater	regaining our composure

MFW

Only unigrams with relative frequency > 0.03

a	not
all	of
and	on
as	p_apos
at	p_comma
be	p_exlam
but	p_hyphen
by	p_period
for	p_ques
from	p_quote
had	p_semi
have	said
he	she
her	so
him	that
his	the
i	this
in	to
is	was
it	which
me	with
my	you

Hierarchical clustering



A Midsummer Night's Dream (3)
Twelfth Night (12)
Much Ado About Nothing (9)
Two Gentlemen (13)
Measure for Measure (6)
Othello (32)
Julius Caesar (29)

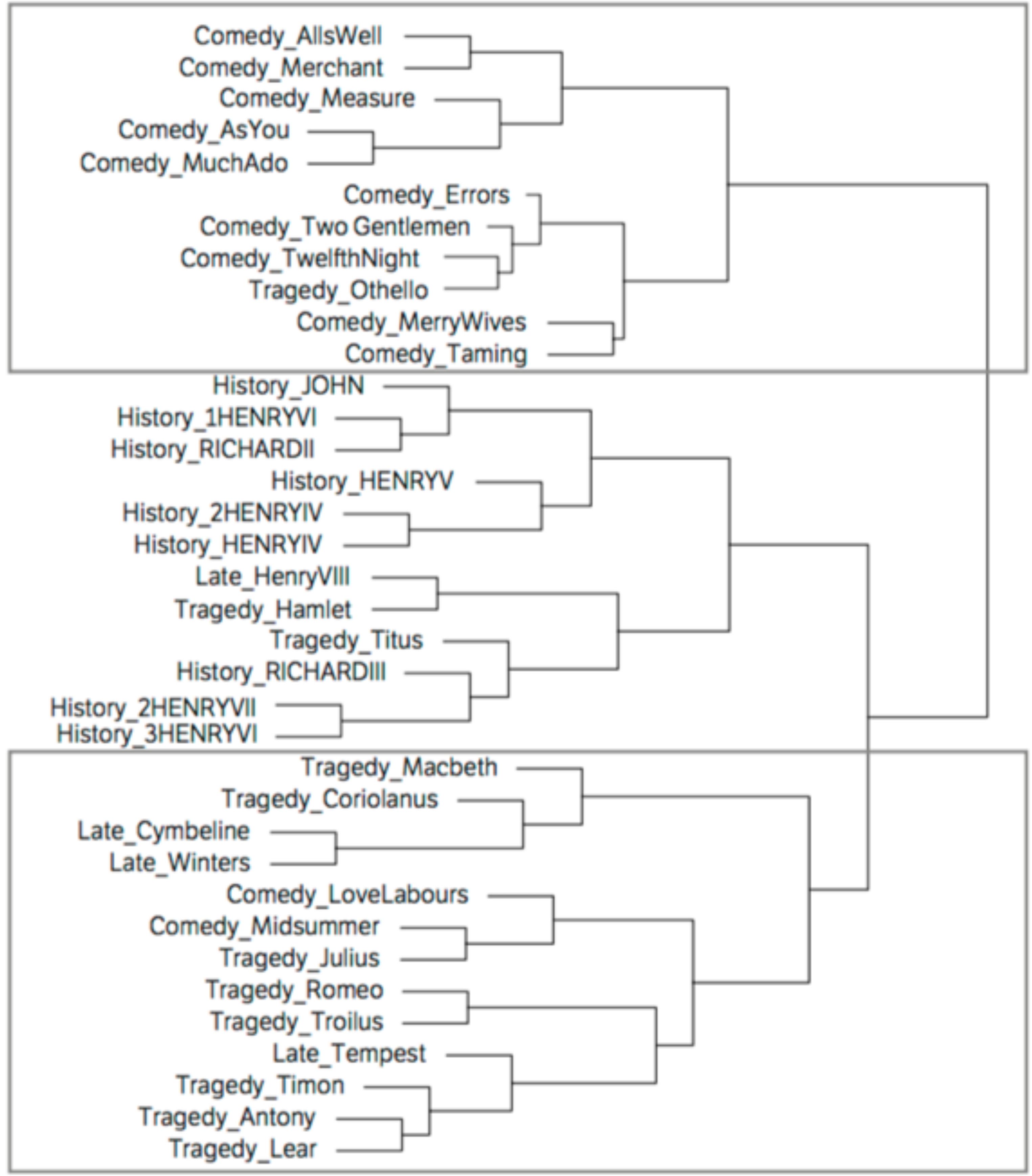
The Winter's Tale (14)
Cymbeline (27)
Antony and Cleopatra (25)
Coriolanus (26)
Henry VIII (21)
Hamlet (28)
Troilus and Cressida (36)
Macbeth (31)
Timon of Athens (34)

All's Well That Ends Well (2)
Taming of the Shrew (10)
Merry Wives of Windsor (8)
A Midsummer Night's Dream (1)
Romeo and Juliet (33)
Comedy of Errors (4)
Merchant of Venice (7)
The Tempest (11)

Love's Labours' Lost (5)
1 Henry IV (15)
2 Henry IV (17)
Henry V (20)
1 Henry VI (16)
King John (22)
Richard II (23)

2 Henry VI (18)
2 Henry VI (19)
Richard III (24)
King Lear (30)
Titus Andronicus (35)

Allison et al. 2009



Allison et al. 2009

“But there is also a simpler explanation: namely, that these features which are so effective at differentiating genres, and so entwined with their overall texture – these features cannot offer new insights into structure, because they aren't independent traits, but mere consequences of higher-order choices. Do you want to write a story where each and every room may be full of surprises? Then locative prepositions, articles and verbs in the past tense are bound to follow. They are the effects of the chosen narrative structure.”

Project presentation

Monday April 25 (6) + Wednesday April 27 (5)

10 min presentation +
3-5 min questions

http://www.phdcomics.com/comics.php?f=1553

Final report

- 8 pages, single spaced.
- Complete description of work undertaken
    - Data collection
    - Methods
    - Experimental details
    - Comparison with past work
    - Analysis
- See many of the papers we've read this semester for examples.

Final report

- Clarity. For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?
- Originality. How original is the approach or problem presented in this paper? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?
- Soundness. Is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by proper experiments, proofs, or other argumentation?
- Substance. Does this paper have enough substance, or would it benefit from more ideas or results? Do the authors identify potential limitations of their work?
- Evaluation. To what extent has the application or tool been tested and evaluated? Does this paper present a compelling argument for
- Meaningful comparison. Do the authors make clear where the presented system sits with respect to existing literature? Are the references adequate? Are the benefits of the system/application well-supported and are the limitations identified?
- Impact. How significant is the work described? Will novel aspects of the system result in other researchers adopting the approach in their own work?

http://mybinder.org/repo/dbamman/dds