Clustering for Reserving Segmentation

John Avitabile

Jay Cooke

September 7, 2018

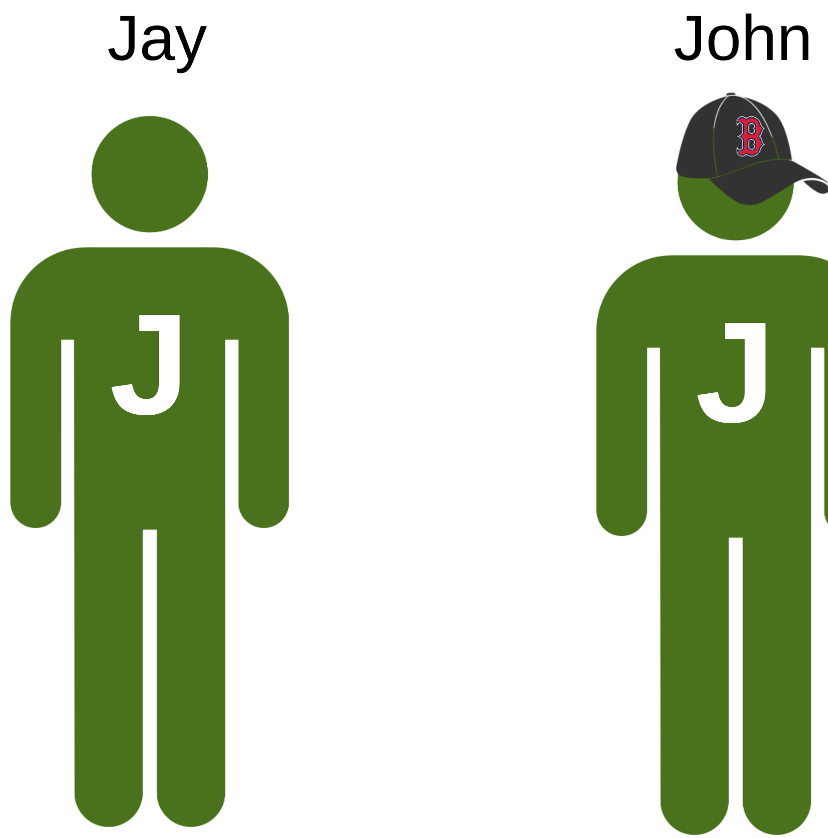
Agenda

- The Need for Segmentation
- Introduction to Clustering
- Clustering with Error Distributions
- Stability Measurement
- Summary and Implementation

Agenda

- The Need for Segmentation
- Introduction to Clustering
- Clustering with Error Distributions
- Stability Measurement
- Summary and Implementation

Team of two reserving actuaries



Each actuary has the capacity to do **one reserve analysis**. An analysis can include data from one segment or combined data from multiple segments.

Four segments of business

A
Bodily Injury

B
Property Damage

C
Collision

D
Comprehensive

Four segments of business

A Small Accounts	**B** Medium Accounts
C Large Accounts	**D** Construction Accounts

These four segments of business need to be combined into **two** groups…but how do we do it?

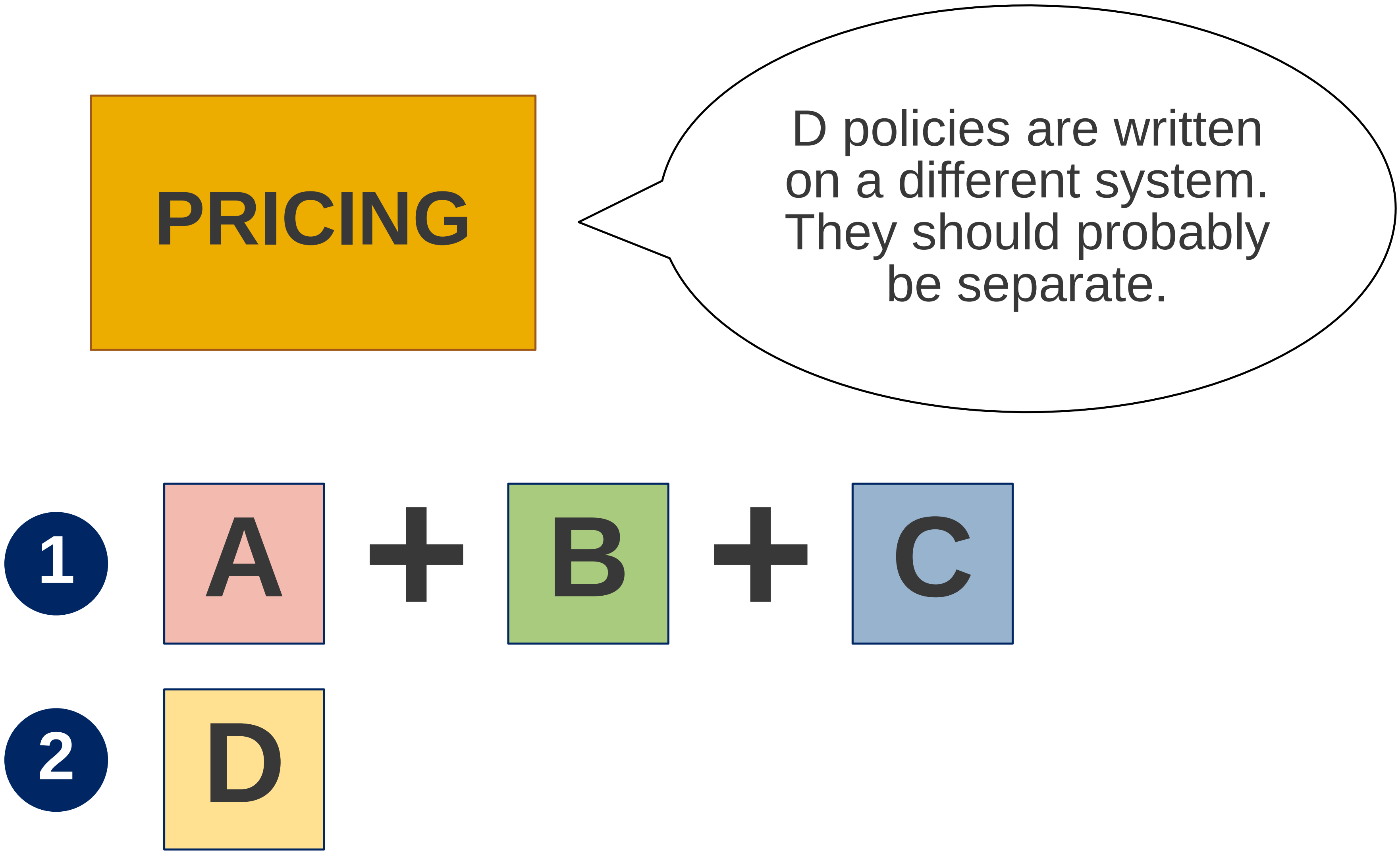
The Claims department says…

CLAIMS

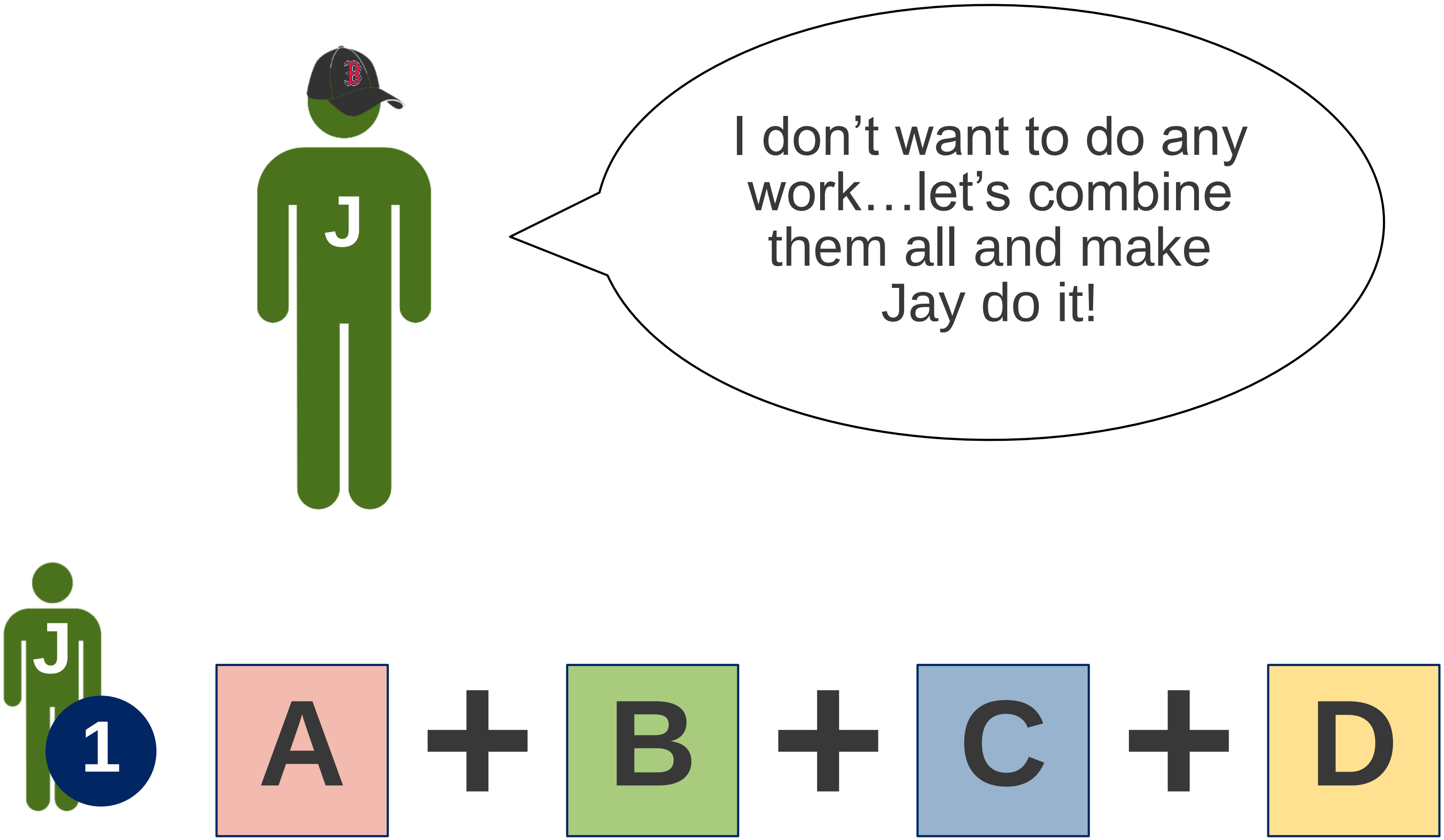
We have AD adjusters and BC adjusters. That probably impacts development.

1. **A + D**
2. **B + C**

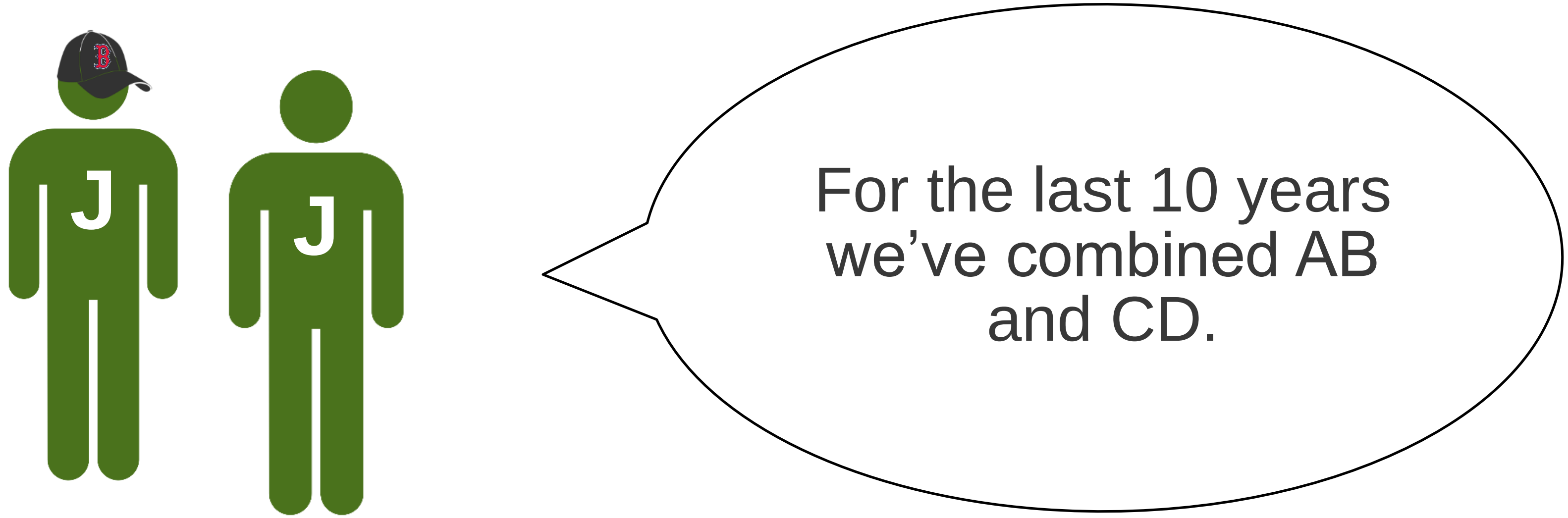
The Pricing department says…



John says…


J
I don’t want to do any work…let’s combine them all and make Jay do it!
J
1
A
B
C
D

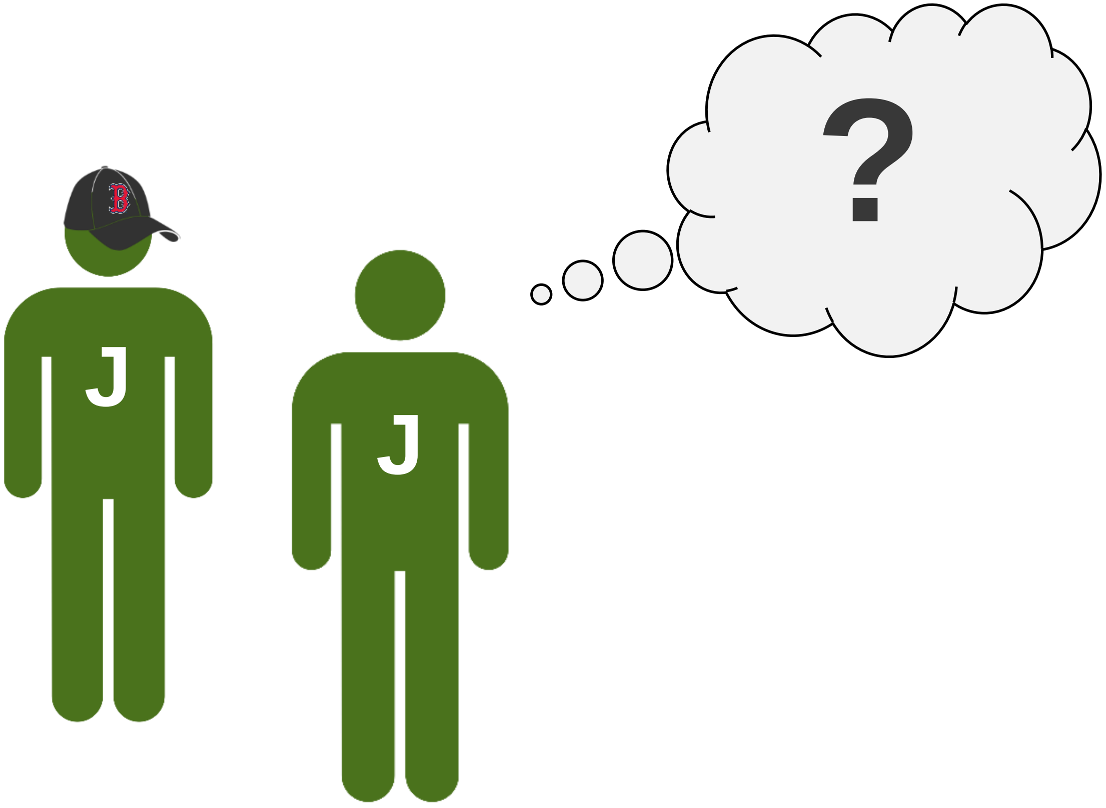
Reserving history says…



1 A + B

2 C + D

How should we group these segments??



We want a method that is…

Data Driven

Credible

Not prone to biases

CLUSTERING

How does this work? Let's use our four segments as an example.

Agenda

- The Need for Segmentation
- Introduction to Clustering
- Clustering with Error Distributions
- Stability Measurement
- Summary and Implementation

Uses for Cluster Analysis

Market Research

Grouping customers for targeted marketing and product design

Computer Science

Partitioning a digital image into regions for image recognition

Social Science

Identifying patterns in unsolved crimes

Insurance

NCCI Hazard Group Mapping for workers' compensation (Robertson, 2007)

Let's take a closer look at our four segments

A	B
C	D

Segment A

A

Loss Year	Paid Loss $ by Age				
	12	24	36	48	60
2013	1,200	2,350	2,620	2,620	2,620
2014	1,400	2,550	2,810	2,810	
2015	1,000	2,020	2,210		
2016	1,300	2,320			
2017	1,200				

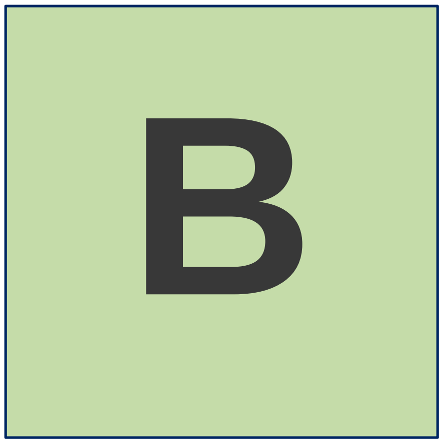
Two Important Development Factors

Loss Year	Age-to-Age			
	12-24	24-36	36-48	48-60
2013	1.96	1.11	1.00	1.00
2014	1.82	1.10	1.00	
2015	2.02	1.09		
2016	1.78			
Selected (Mean)	**1.90**	**1.10**	**1.00**	**1.00**

Selected simple all-year average for simplicity

No development beyond 36 months

Segment B



Loss Year	Paid Loss $ by Age				
	12	24	36	48	60
2013	1,560	2,210	2,740	2,740	2,740
2014	1,820	2,290	2,890	2,890	
2015	1,300	1,460	1,830		
2016	1,690	2,400			
2017	1,560				

Loss Year	Age-to-Age			
	12-24	24-36	36-48	48-60
2013	1.42	1.24	1.00	1.00
2014	1.26	1.26	1.00	
2015	1.12	1.25		
2016	1.42			
Selected (Mean)	**1.30**	**1.25**	**1.00**	**1.00**

Segment C

C

Loss Year	Paid Loss $ by Age				
	12	24	36	48	60
2013	1,120	1,760	2,270	2,270	2,270
2014	960	1,030	1,310	1,310	
2015	1,040	3,330	4,420		
2016	800	1,080			
2017	960				

Loss Year	Age-to-Age			
	12-24	24-36	36-48	48-60
2013	1.57	1.29	1.00	1.00
2014	1.07	1.27	1.00	
2015	3.20	1.33		
2016	1.35			
Selected (Mean)	**1.80**	**1.30**	**1.00**	**1.00**

Segment D

D

Loss Year	Paid Loss $ by Age				
	12	24	36	48	60
2013	1,120	1,350	1,420	1,420	1,420
2014	980	1,200	1,420	1,420	
2015	1,200	1,400	1,490		
2016	1,080	1,310			
2017	890				

Loss Year	Age-to-Age			
	12-24	24-36	36-48	48-60
2013	1.21	1.05	1.00	1.00
2014	1.22	1.18	1.00	
2015	1.17	1.06		
2016	1.21			
Selected (Mean)	**1.20**	**1.10**	**1.00**	**1.00**

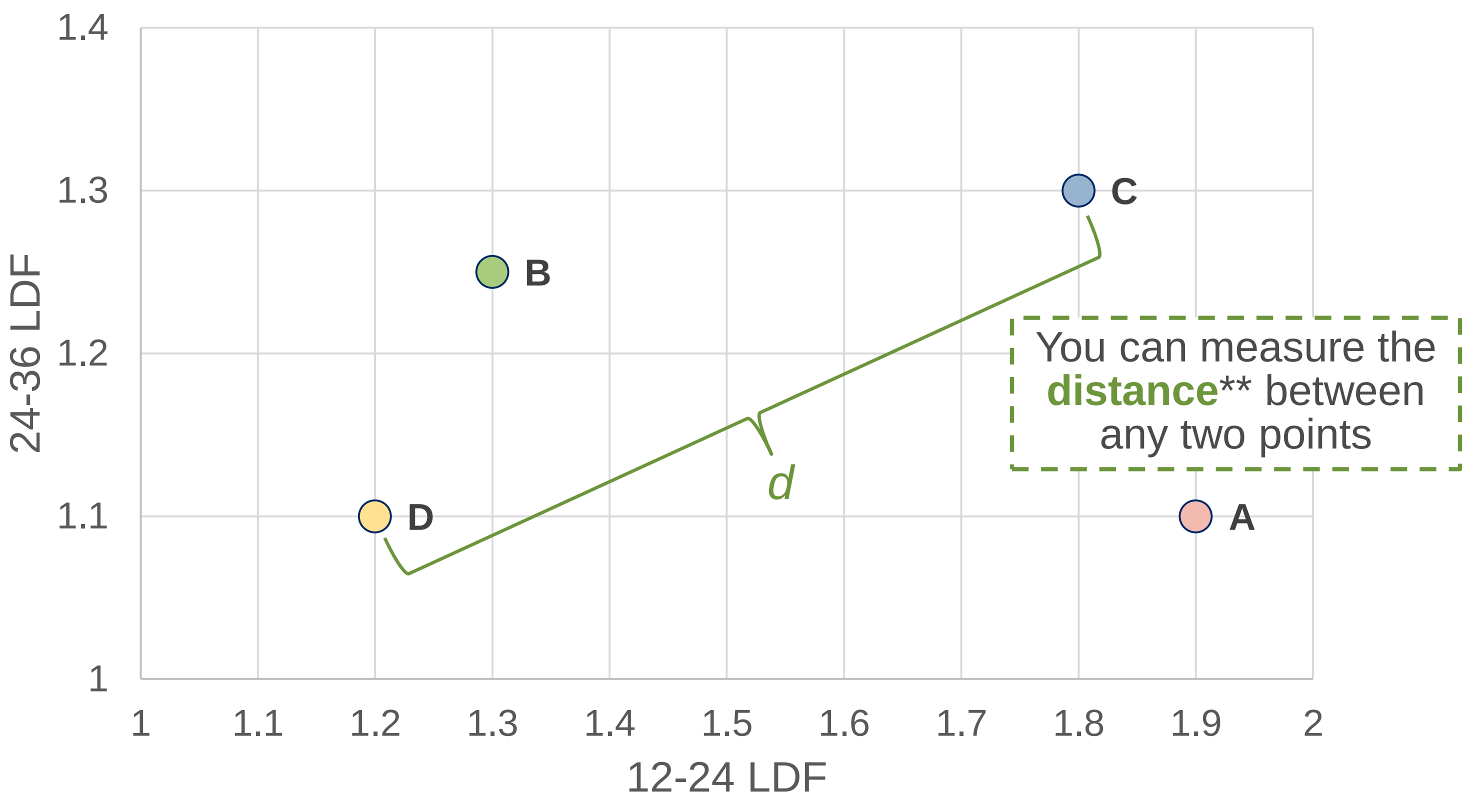
We have two LDFs for each segment

Segment	12-24 LDF	24-36 LDF
A	1.90	1.10
B	1.30	1.25
C	1.80	1.30
D	1.20	1.10

Now it's time for clustering!

Clustering Algorithm

- Let's plot the LDFs on a chart:

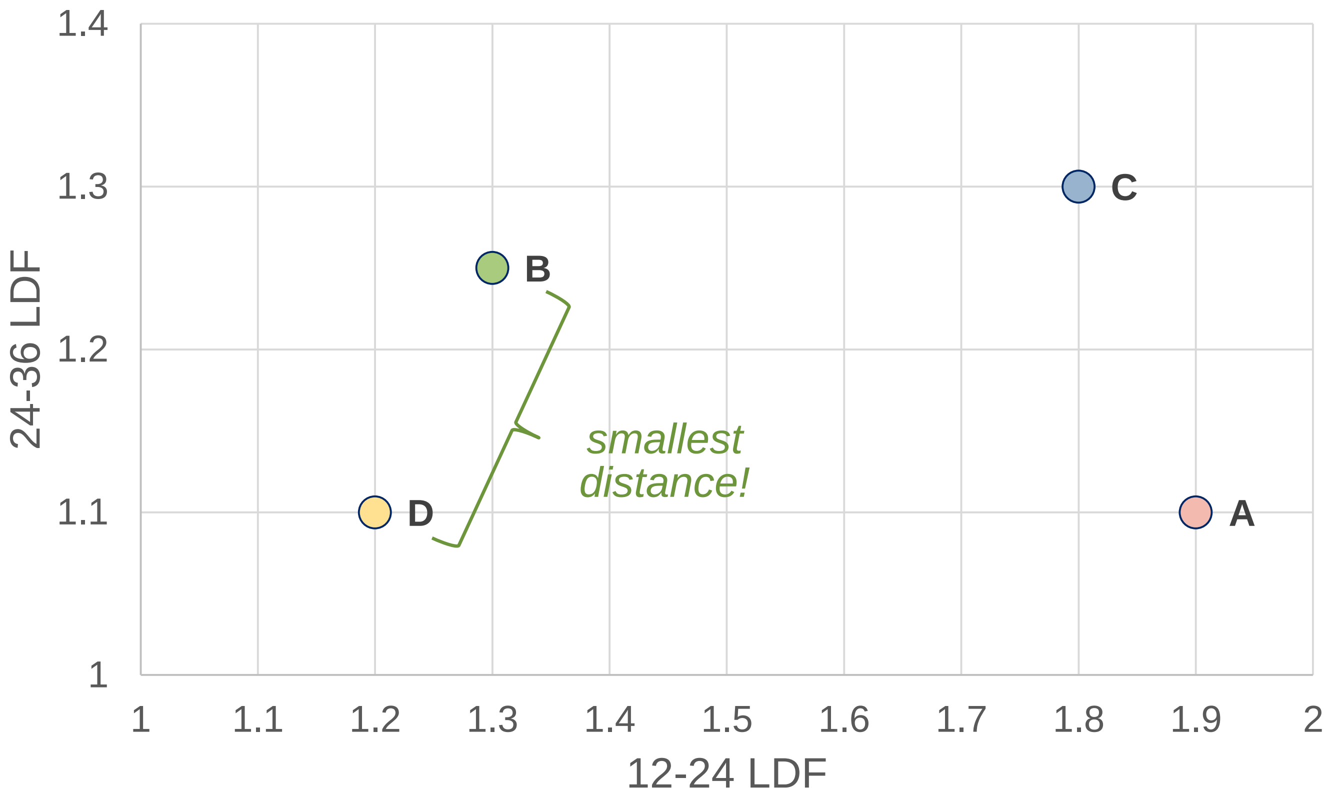


*Using hierarchical clustering throughout this presentation. There are other algorithms that could be used as well.

**We're using Euclidean distance, but there are other distance measure that could be used.

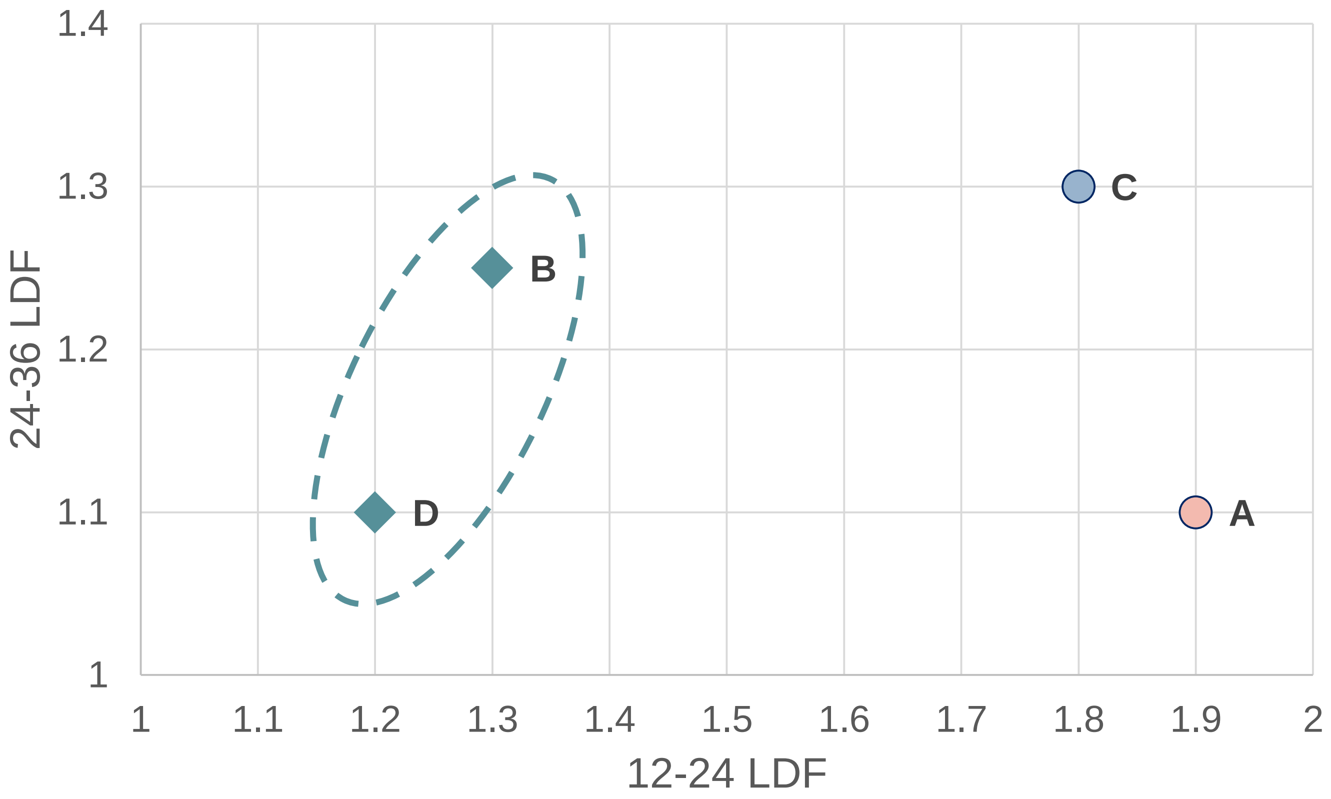
Clustering Algorithm

- Use the distance measure to find the closest two points:



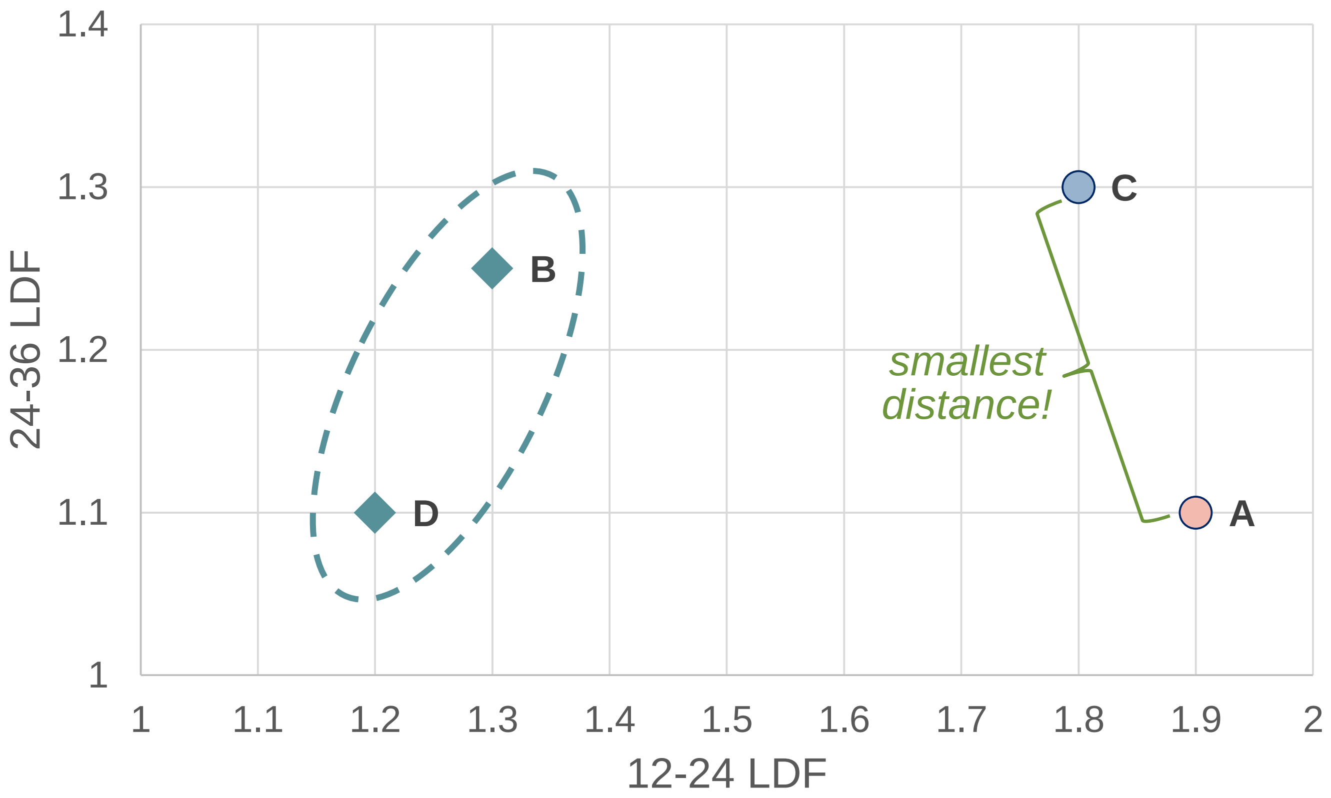
Clustering Algorithm

- The two closest points become one **cluster**!



Clustering Algorithm

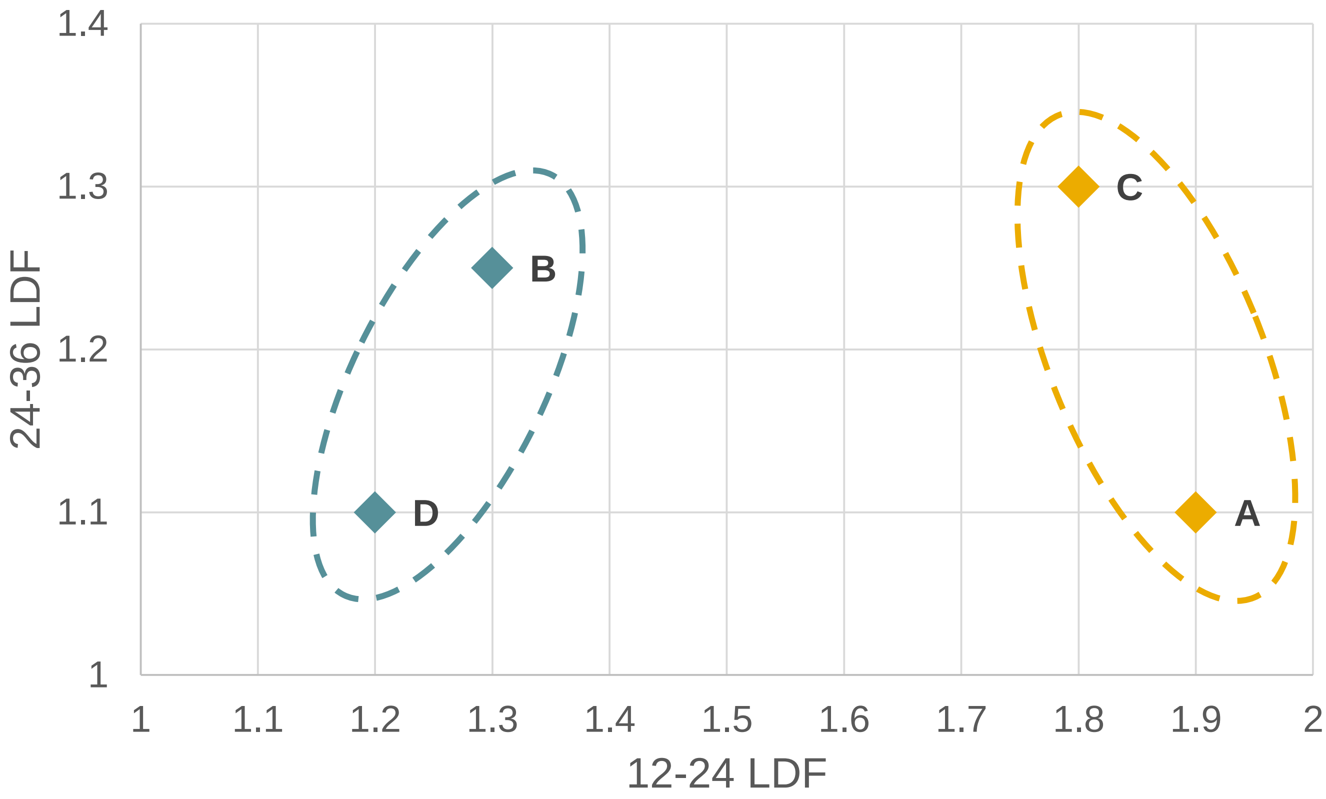
- Now use the distance measure again to find the next smallest distance among both the remaining points and the **cluster***.



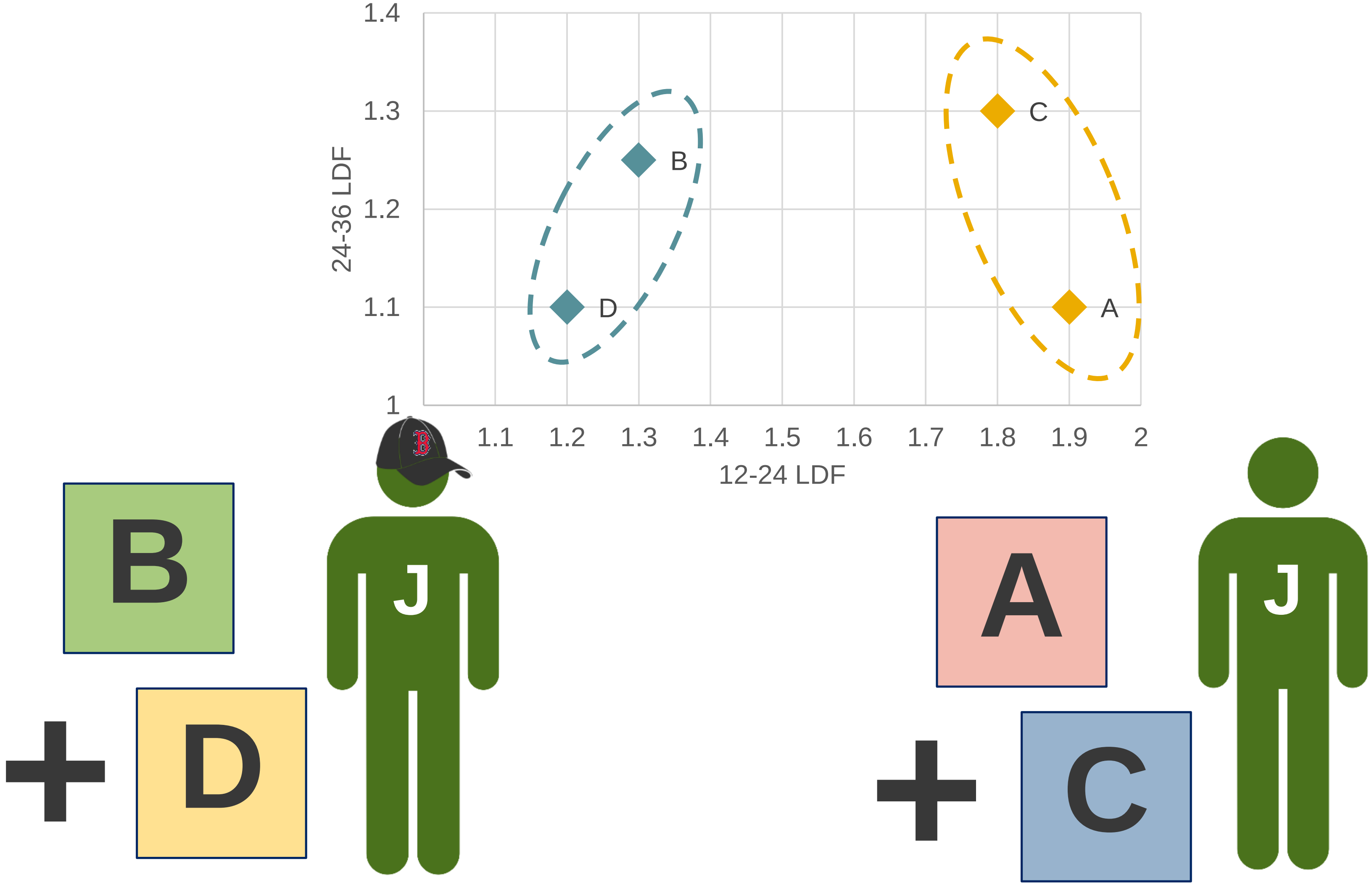
*There are different ways to measure the distance from a cluster. See appendix for details.

Clustering Algorithm

- Now we have **two clusters**!

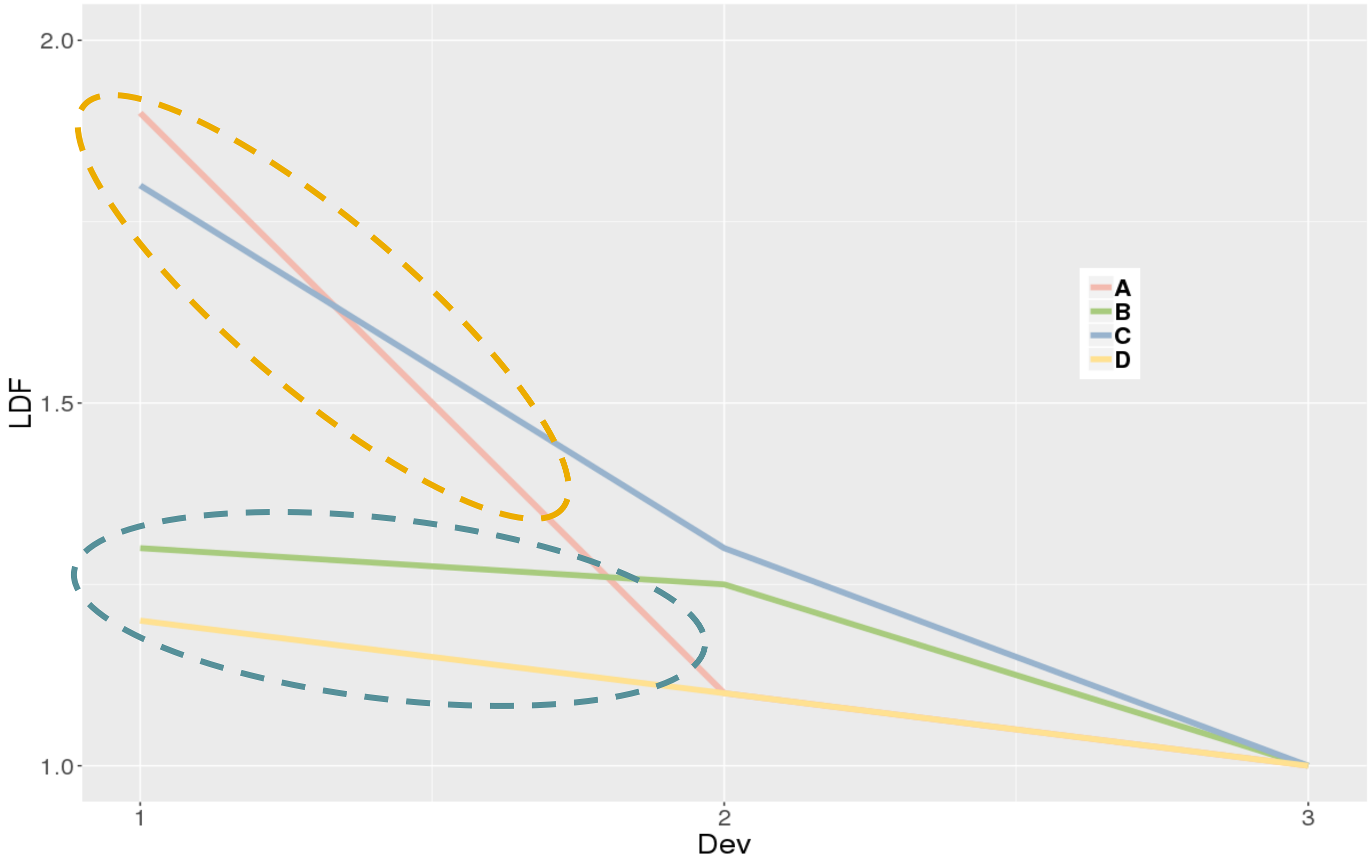


The clustering algorithm has a given us a recommended segmentation:

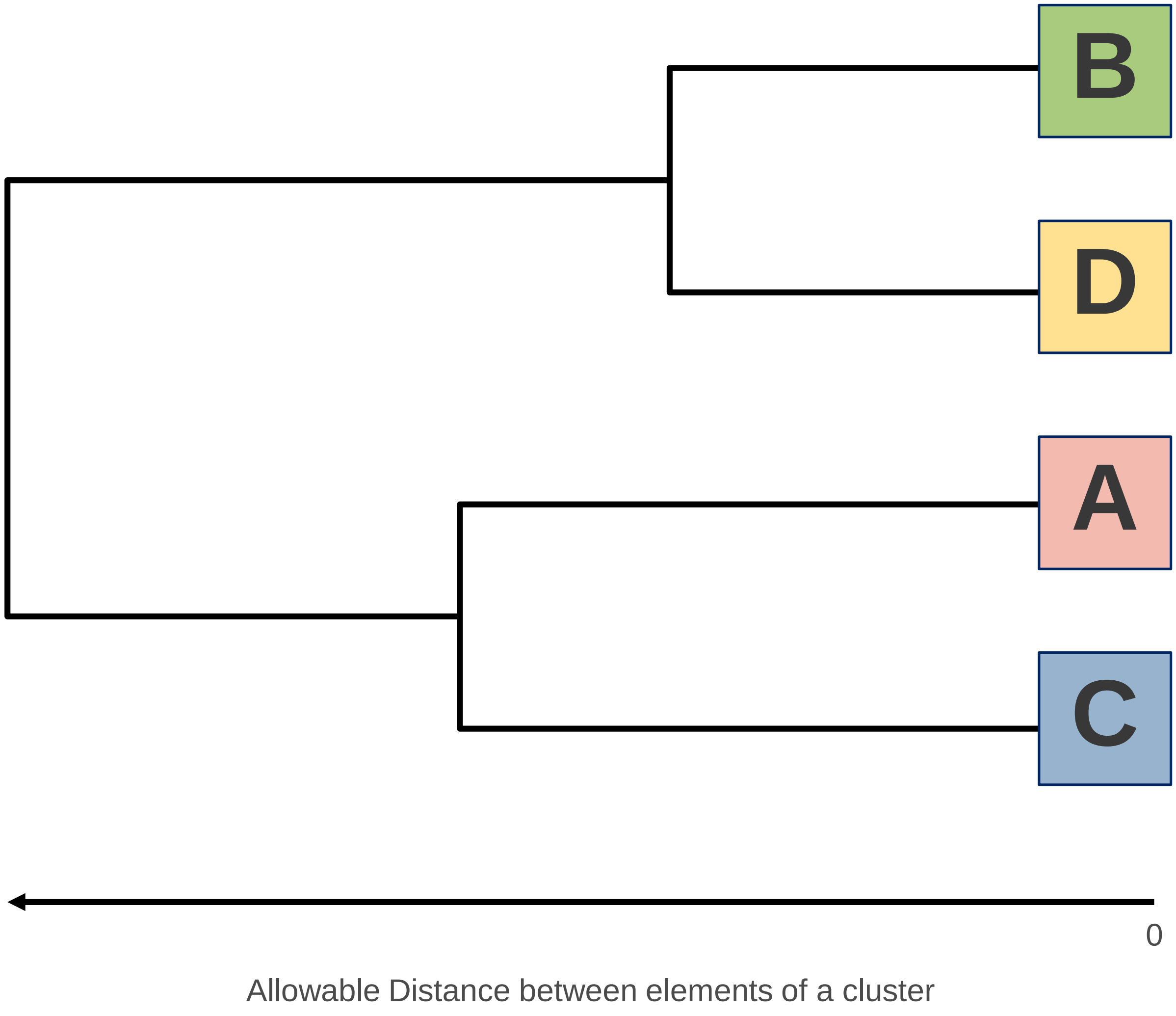


Graphing of LDFs

- Graphing the development factors by age shows that intuitively our result makes sense

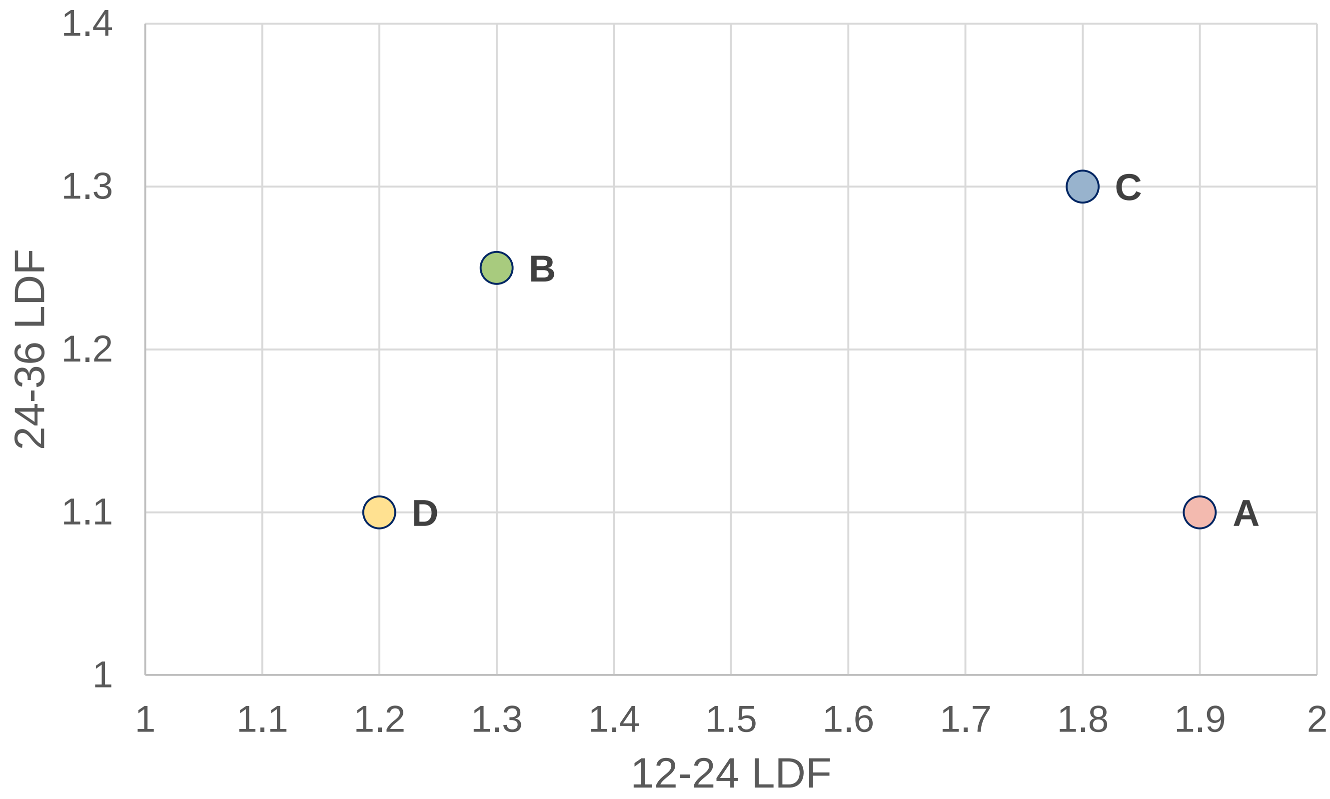


We can use a dendrogram to display the clustering results

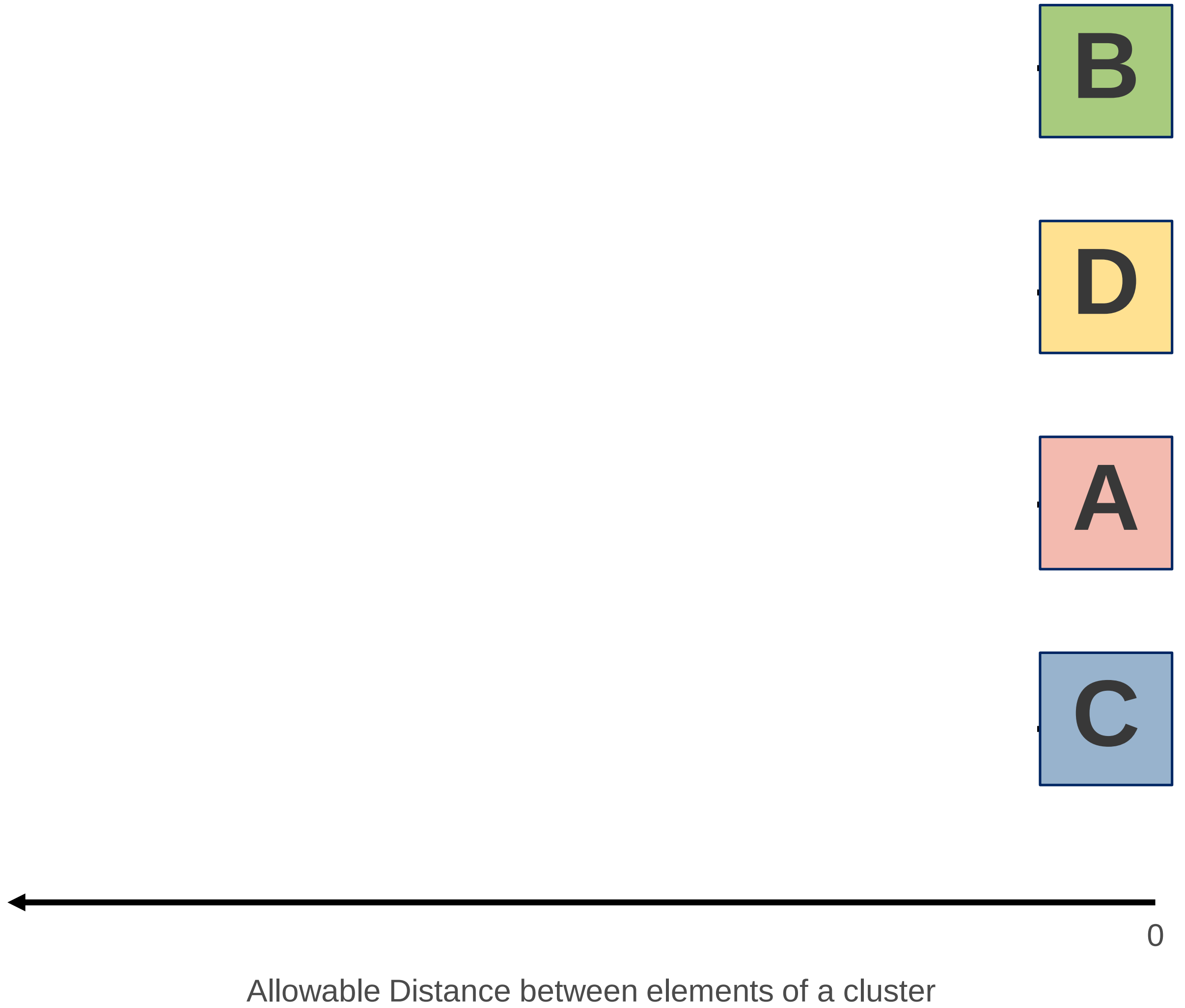


Creating a Dendrogram

- Start with our four points again

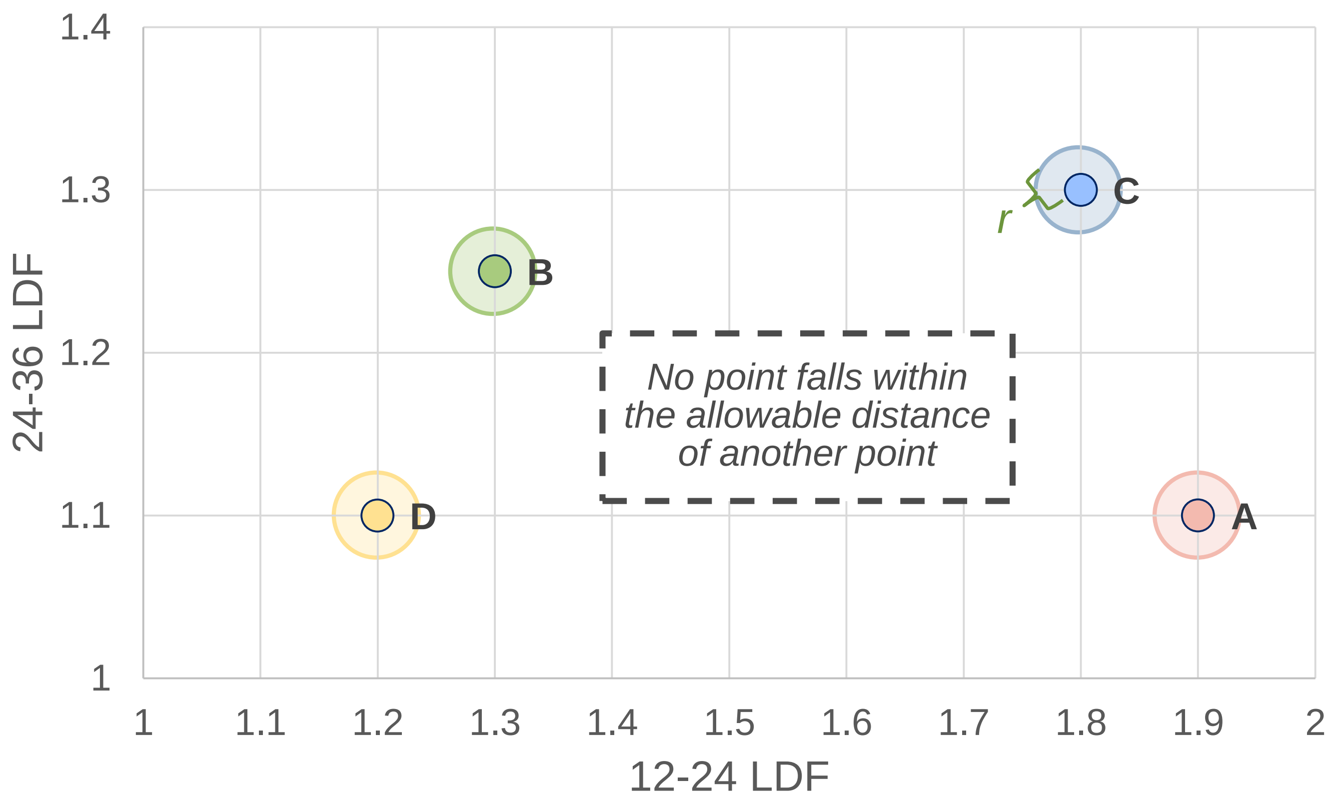


Creating a Dendrogram

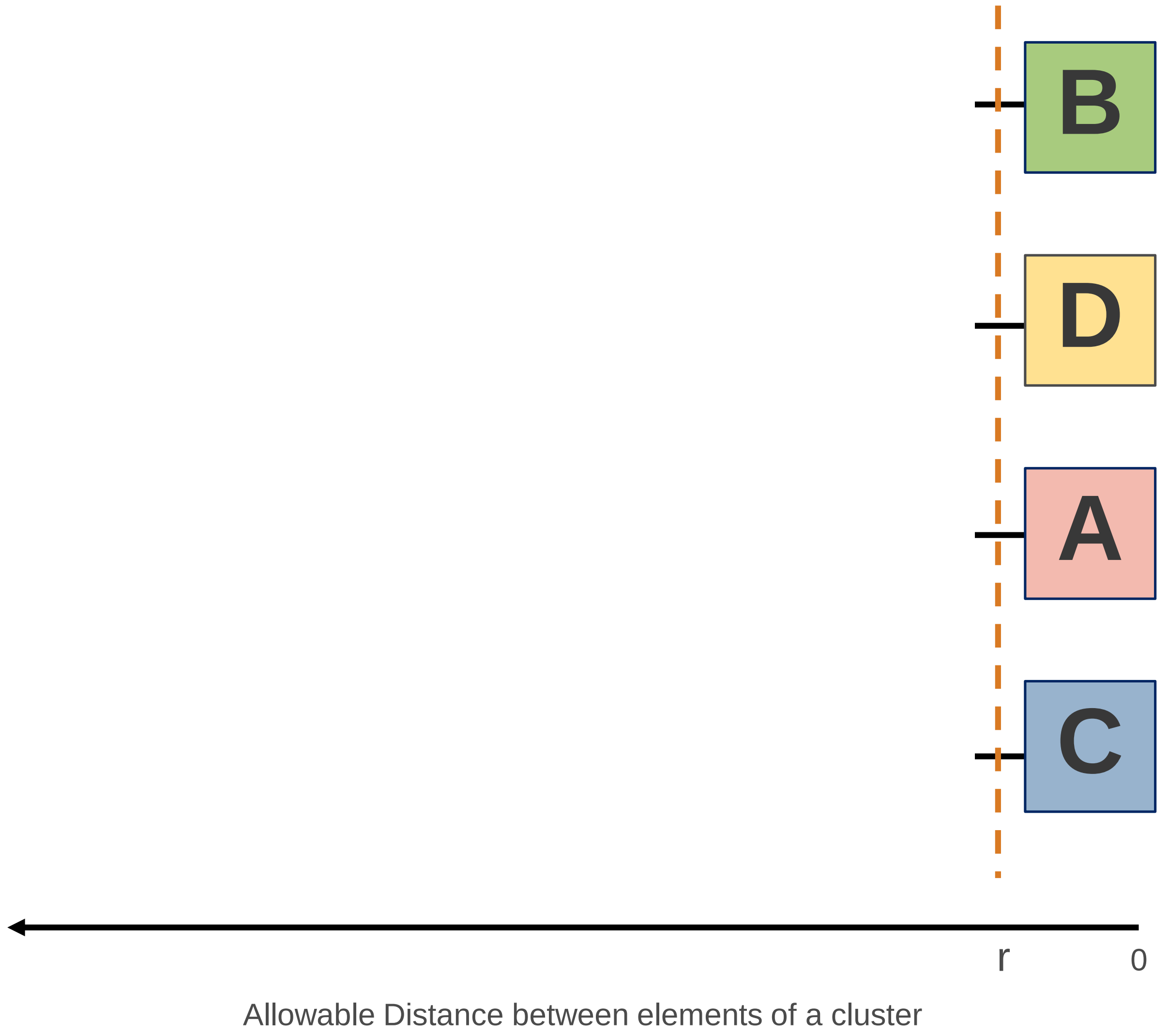


Creating a Dendrogram

- Draw a radius of **allowable distance** around each point

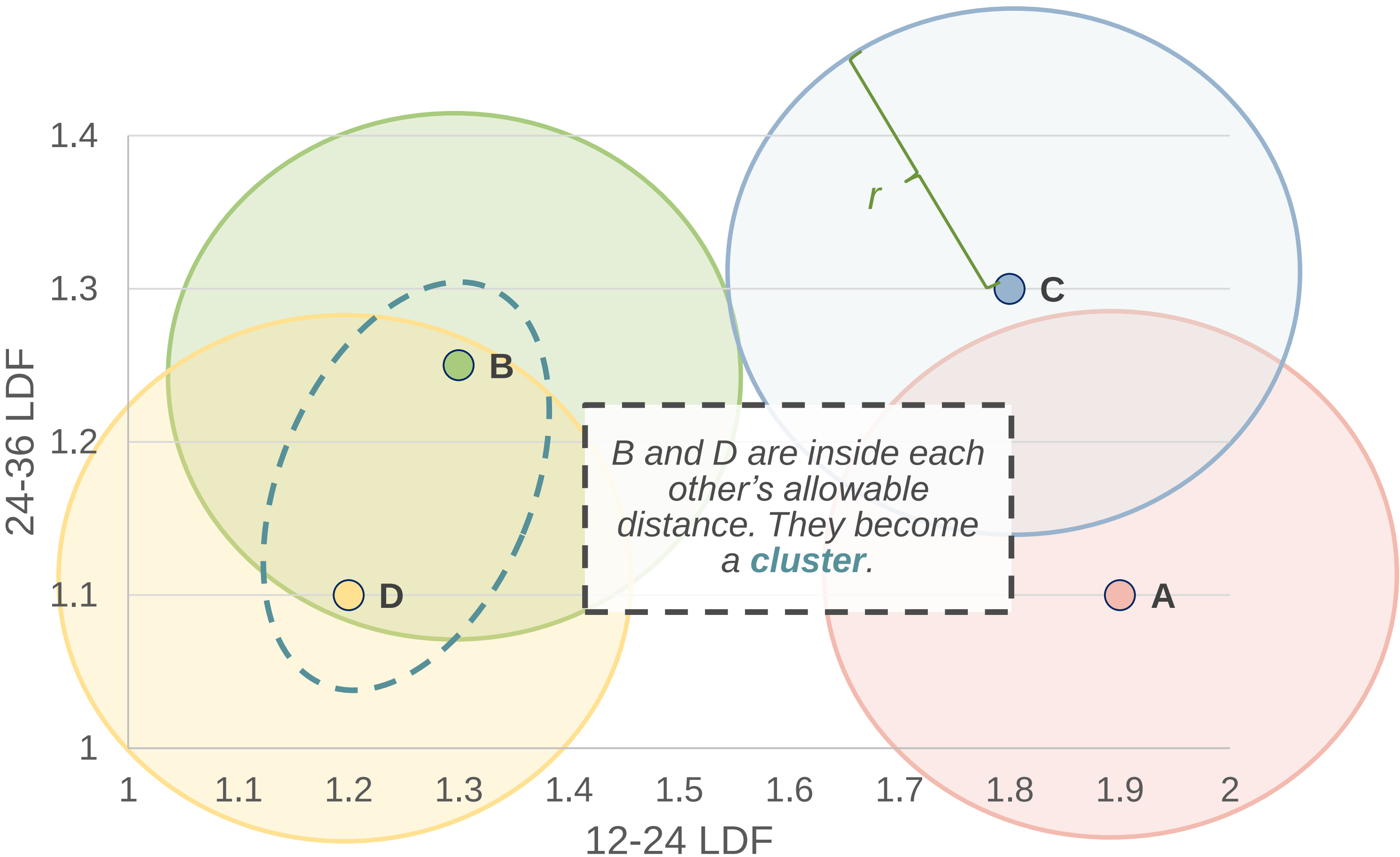


Creating a Dendrogram

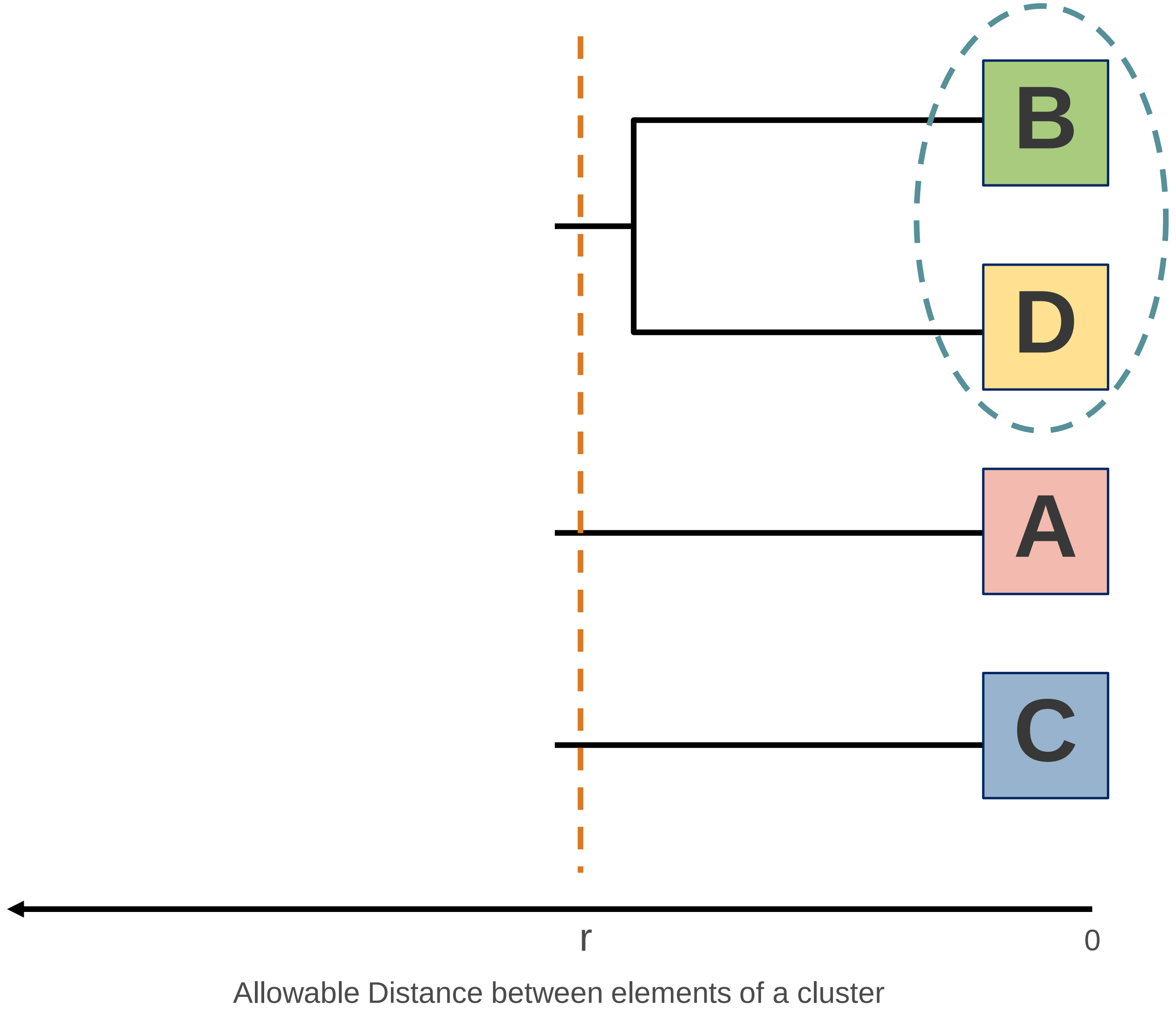


Creating a Dendrogram

- Now increase the allowable distance

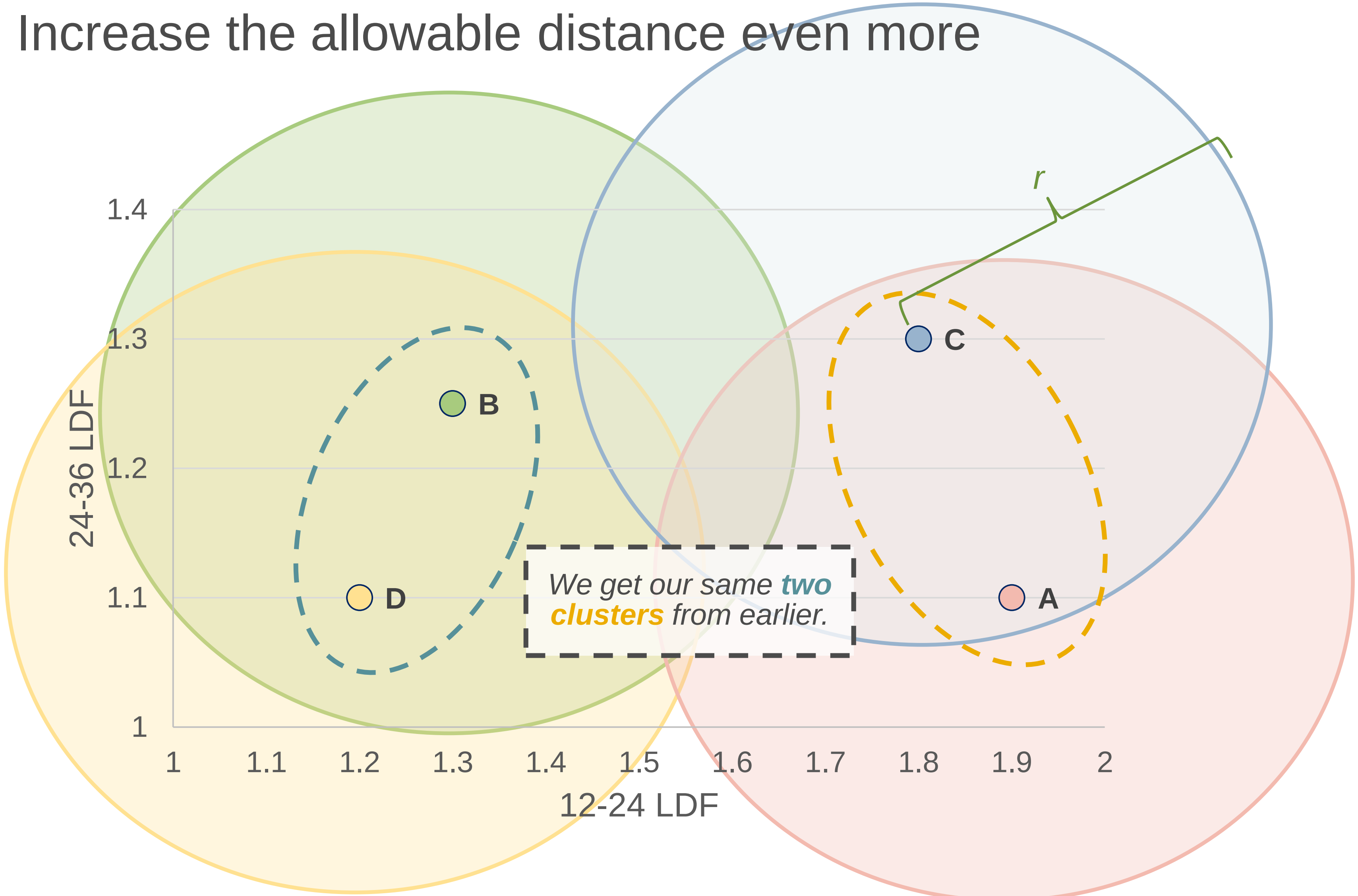


Creating a Dendrogram

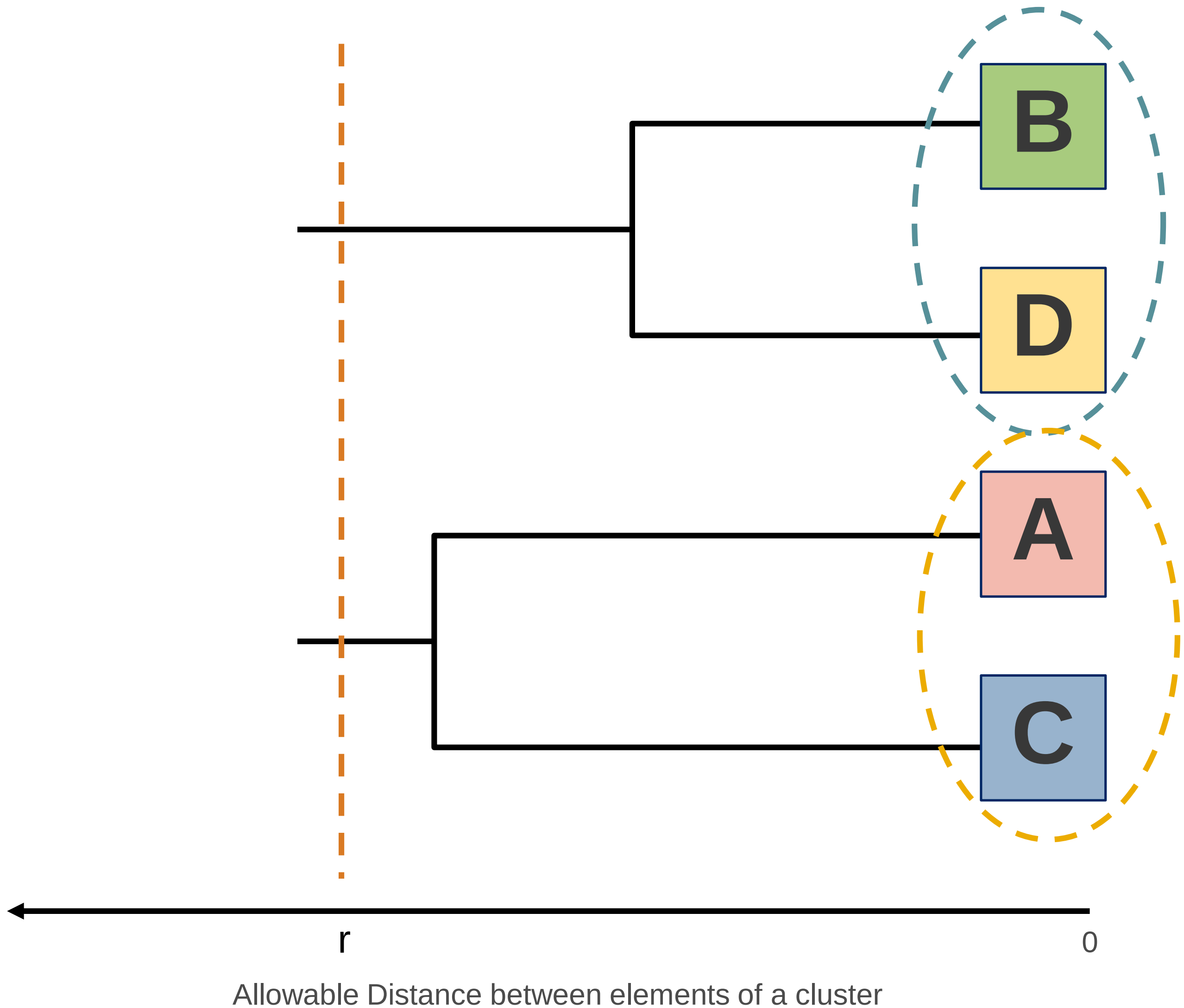


Creating a Dendrogram

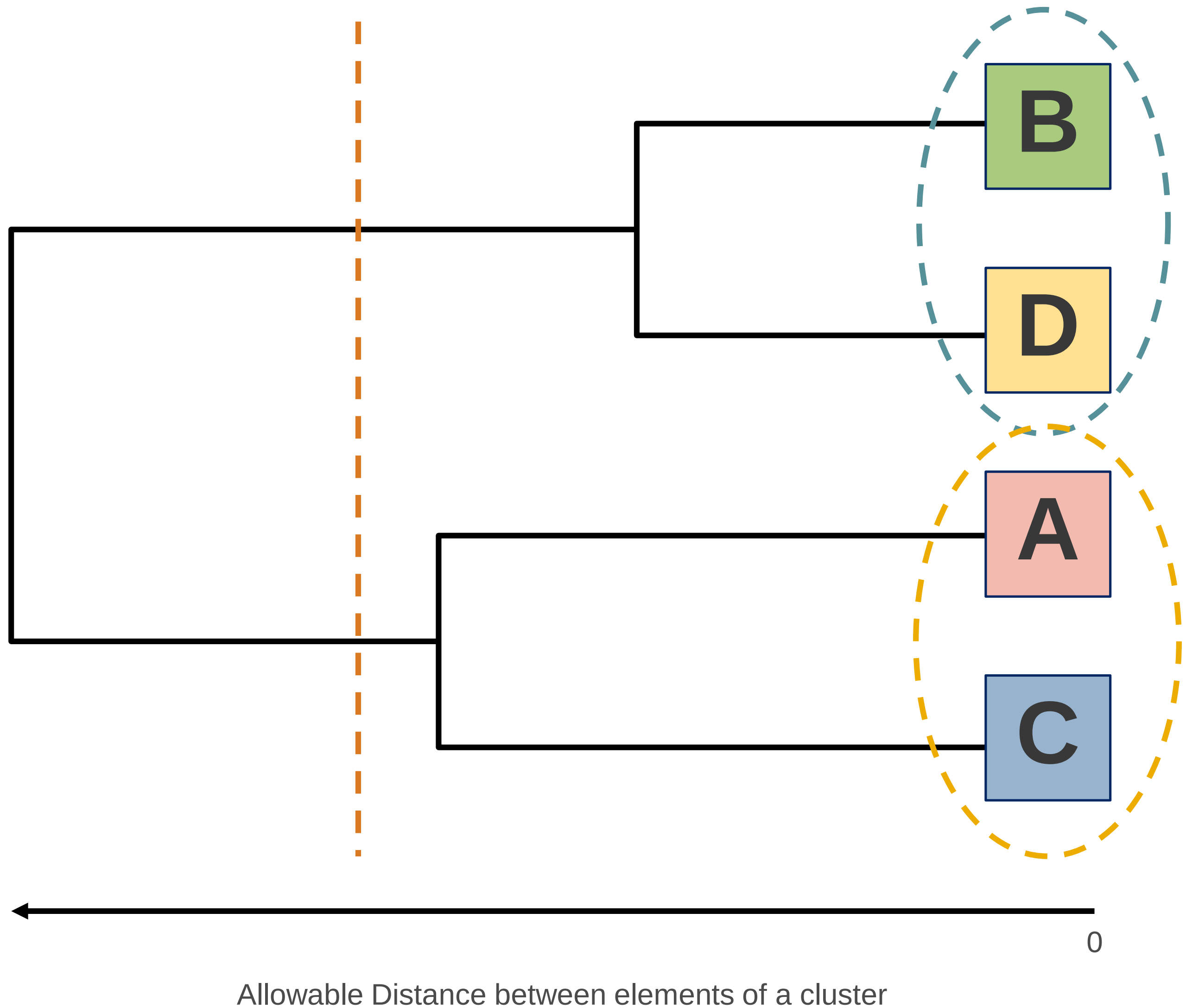
- Increase the allowable distance even more



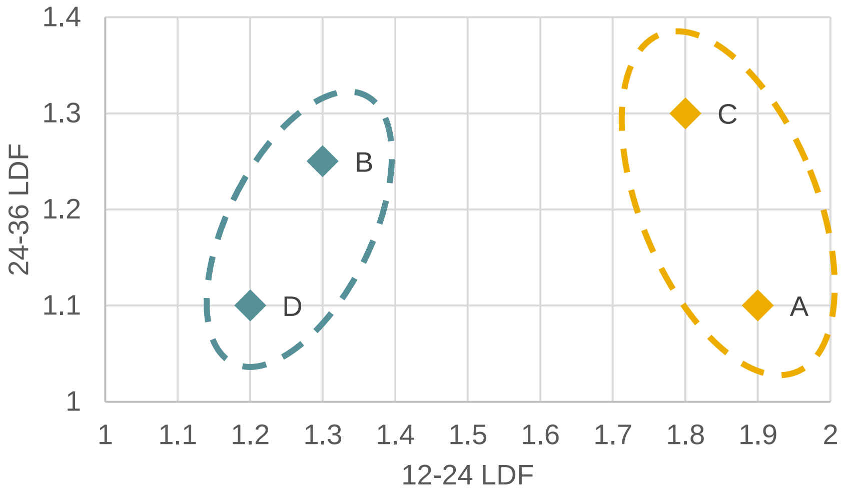
Creating a Dendrogram

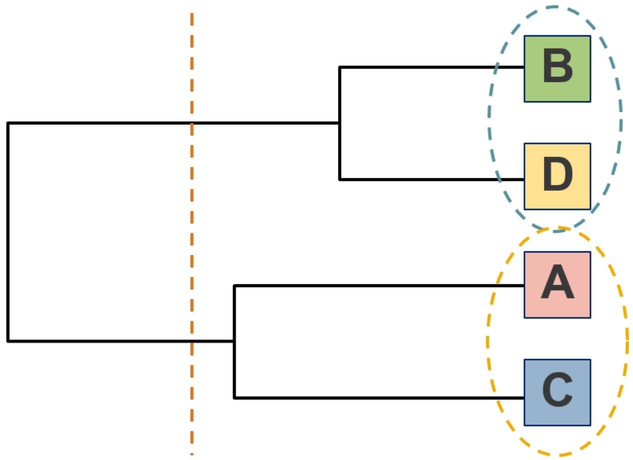


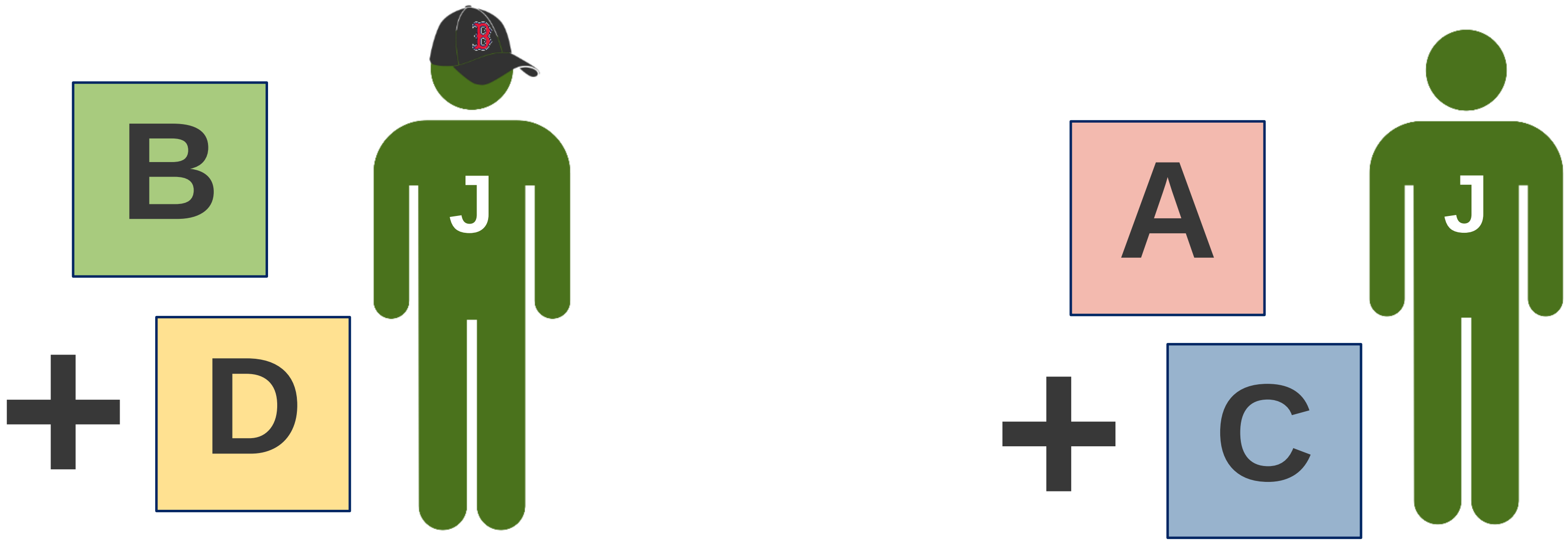
We can use a dendrogram to display the clustering results



Results – we're done now, right?





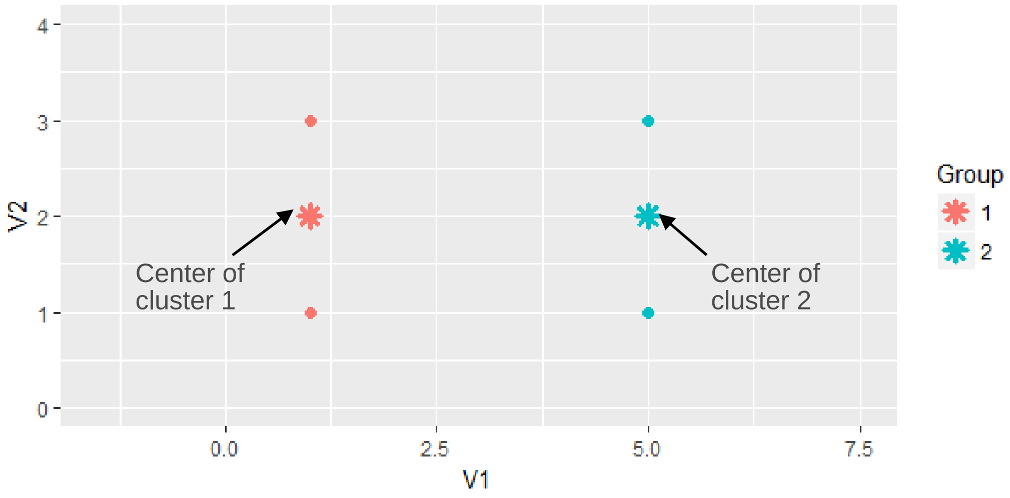


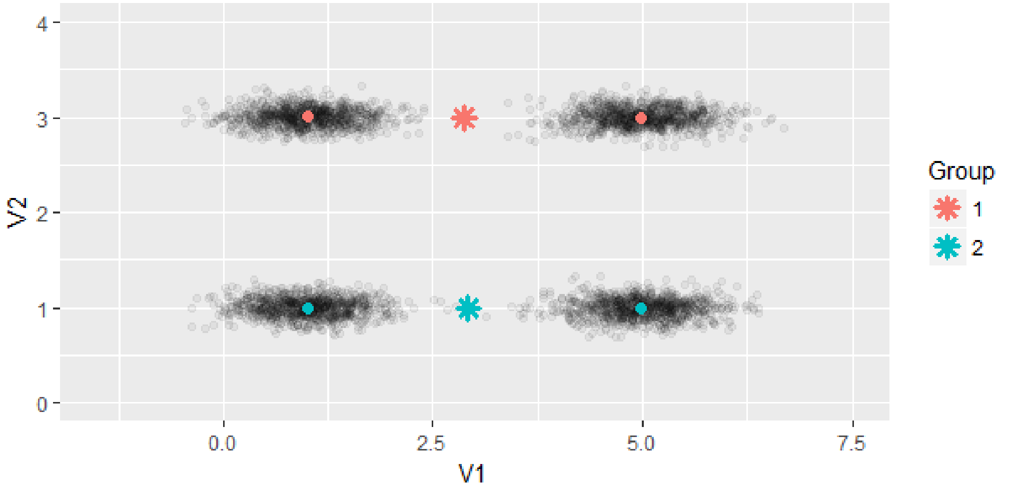
Agenda

- The Need for Segmentation
- Introduction to Clustering
- **Clustering with Error Distributions**
- Stability Measurement
- Summary and Implementation

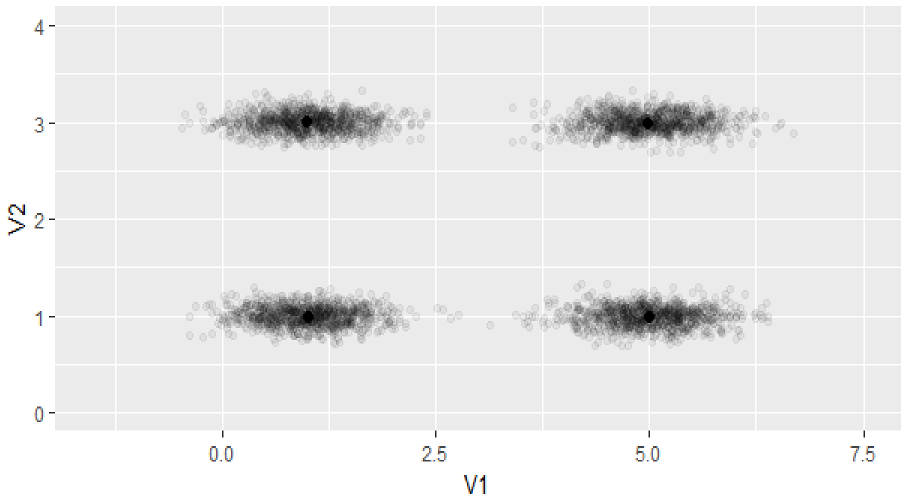
Clustering can be adapted to apply to development factors and the error distributions around them

Why is it important to consider error distributions?

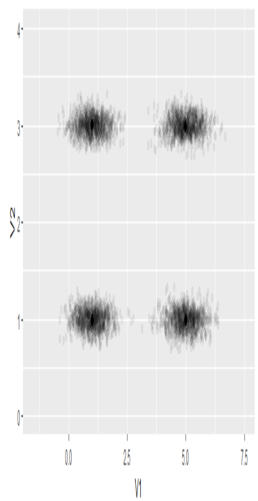




Error-based clustering extends the clustering algorithm to use the error associated with the data



Scale each measurement according to its standard error, perform a new clustering



In practice, the error structure can be more complex than shown in this example

Let's take an even closer look at our four segments

A

B

C

D

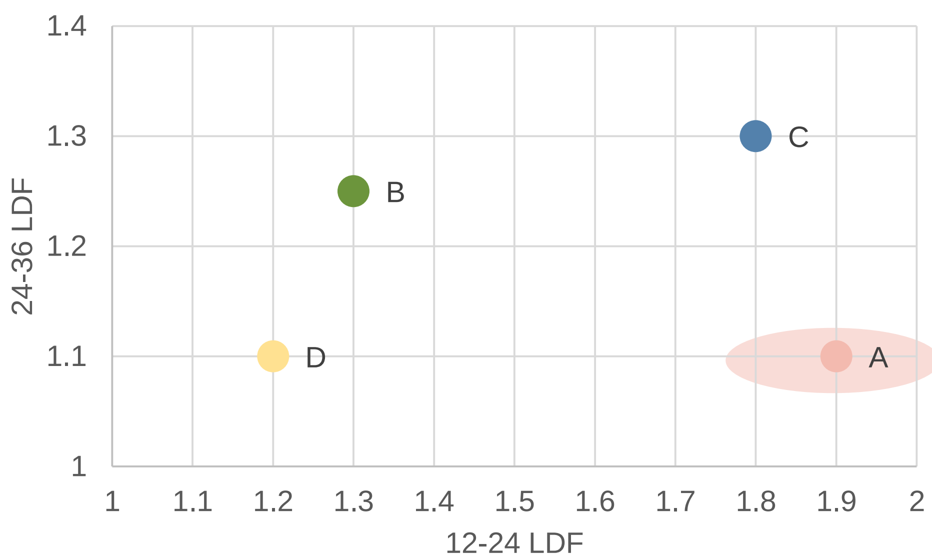
Segment A

A

Loss Year	Age-to-Age			
	12-24	24-36	36-48	48-60
2013	1.96	1.11	1.00	1.00
2014	1.82	1.10	1.00	
2015	2.02	1.09		
2016	1.78			
Selected (Mean)	**1.90**	**1.10**	**1.00**	**1.00**

Variation in first development period

More steady in second period

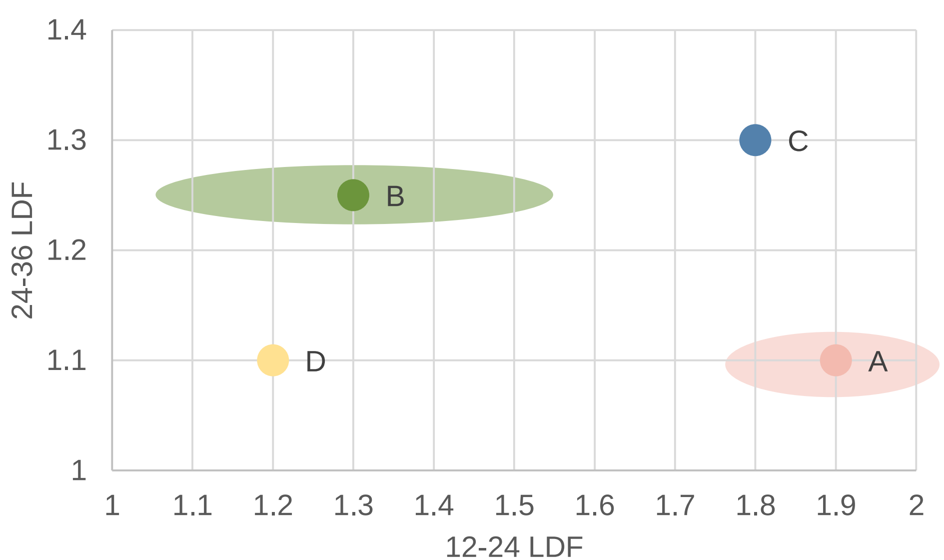


Lots of ways to calculate range around LDFs (bootstrap, Mack)…see other CLRS presentations ☺

Segment B

B

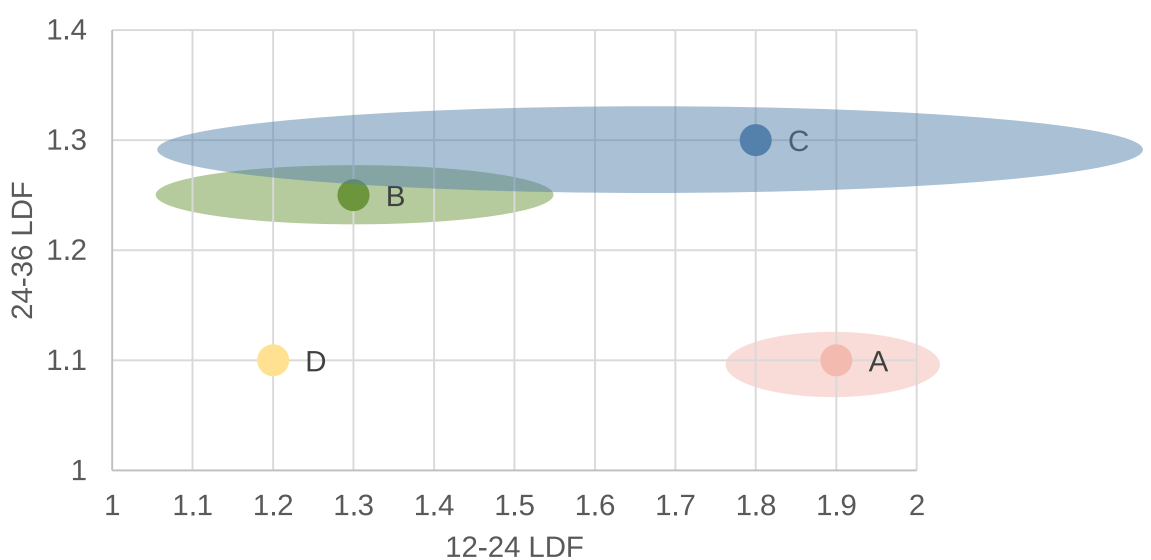
Loss Year	Age-to-Age			
	12-24	24-36	36-48	48-60
2013	1.42	1.24	1.00	1.00
2014	1.26	1.26	1.00	
2015	1.12	1.25		
2016	1.42			
Selected (Mean)	**1.30**	**1.25**	**1.00**	**1.00**



Segment C

C

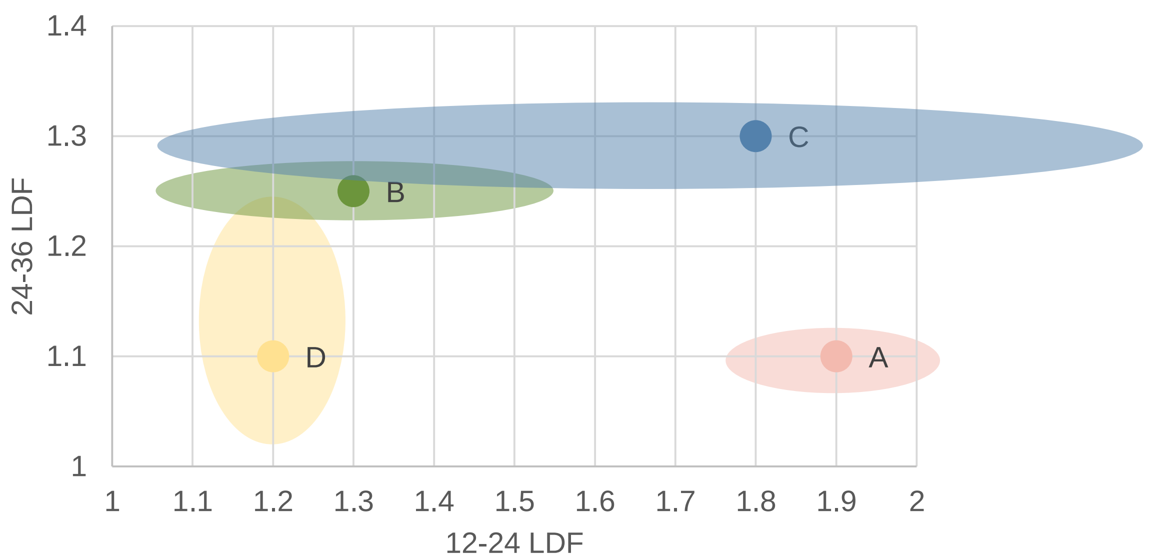
Loss Year	Age-to-Age			
	12-24	24-36	36-48	48-60
2013	1.57	1.29	1.00	1.00
2014	1.07	1.27	1.00	
2015	3.20	1.33		
2016	1.35			
Selected (Mean)	**1.80**	**1.30**	**1.00**	**1.00**



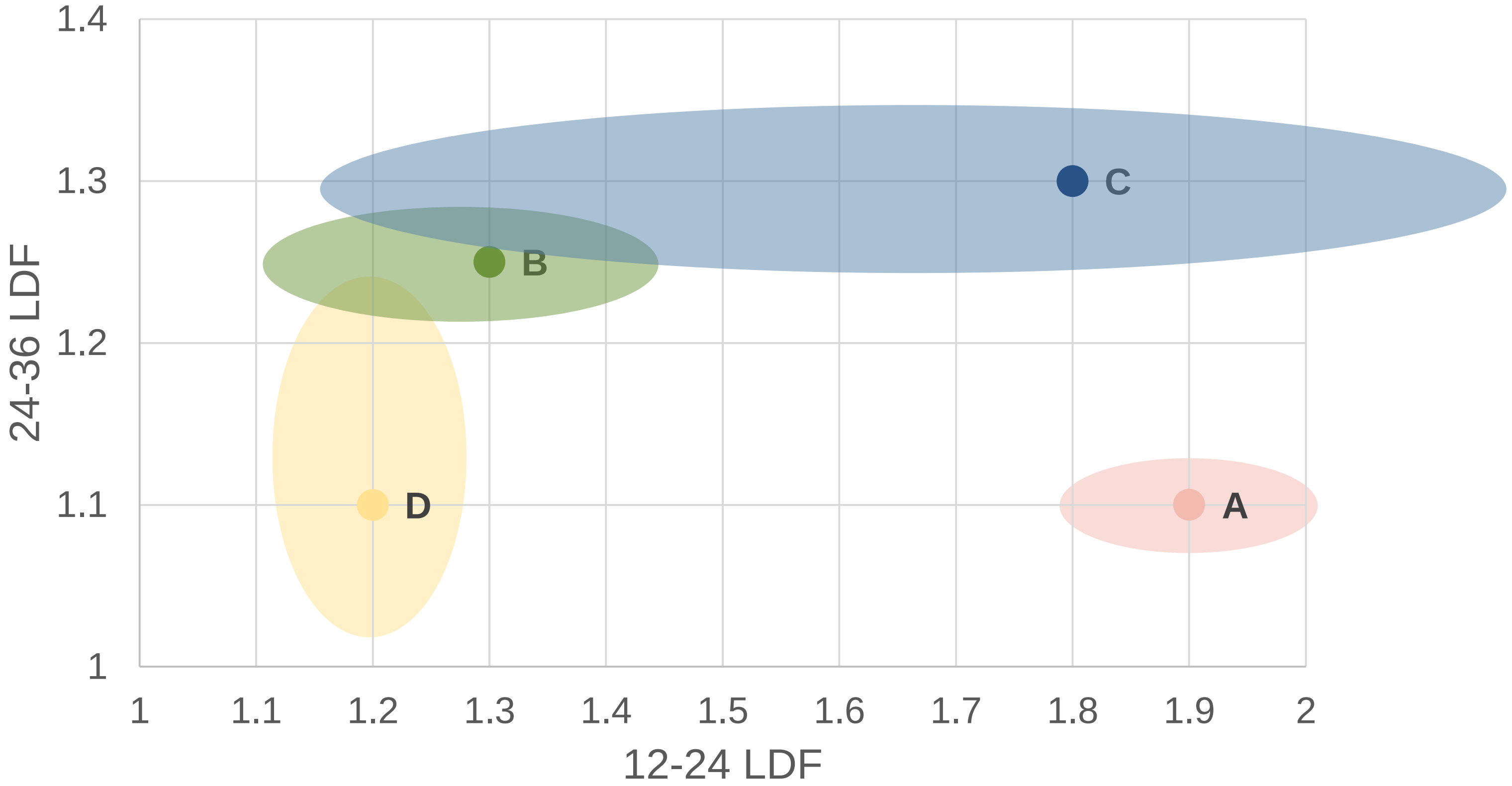
Segment D

D

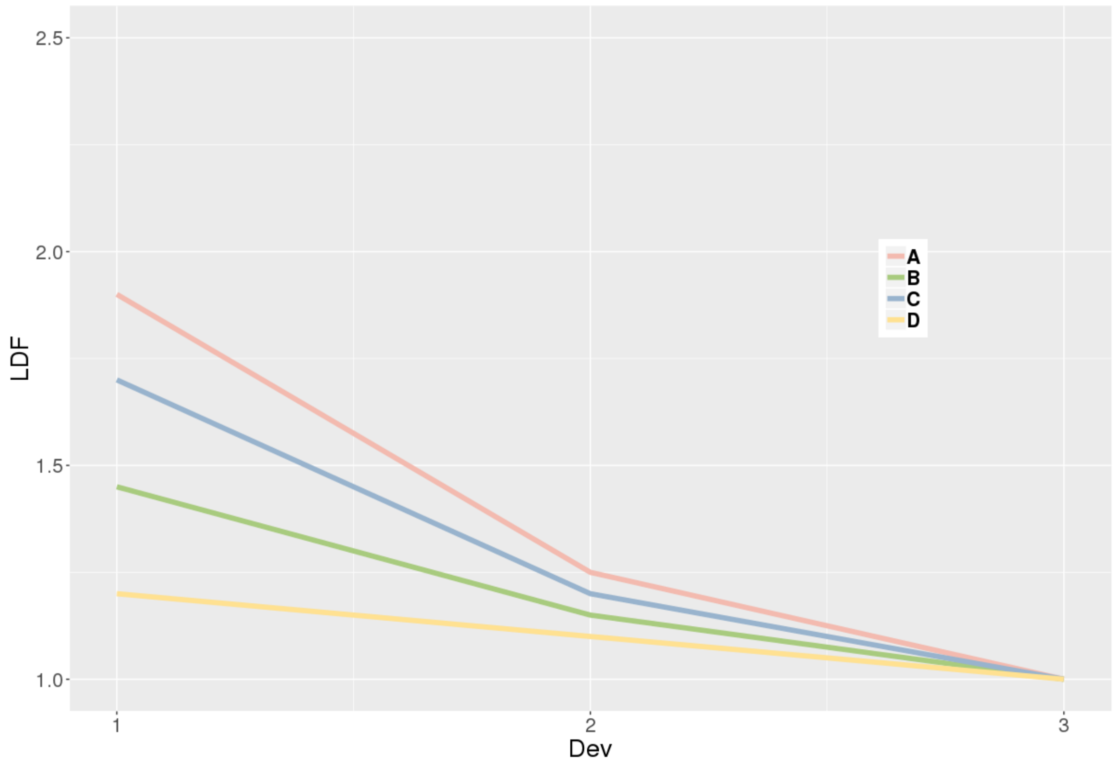
Loss Year	Age-to-Age			
	12-24	24-36	36-48	48-60
2013	1.21	1.05	1.00	1.00
2014	1.22	1.18	1.00	
2015	1.17	1.06		
2016	1.21			
Selected (Mean)	**1.20**	**1.10**	**1.00**	**1.00**

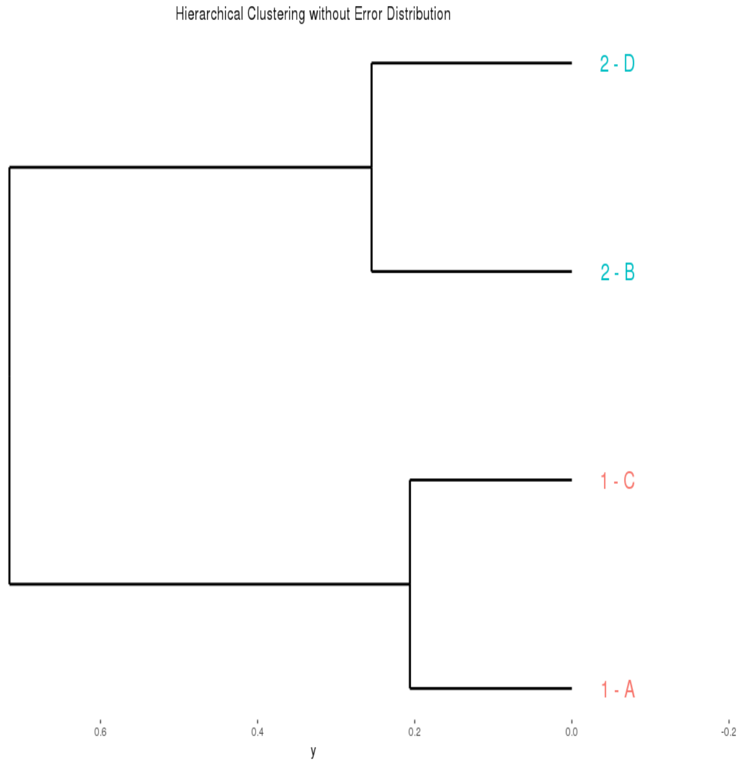


Is it still obvious how we should group these segments?



Standard Clustering

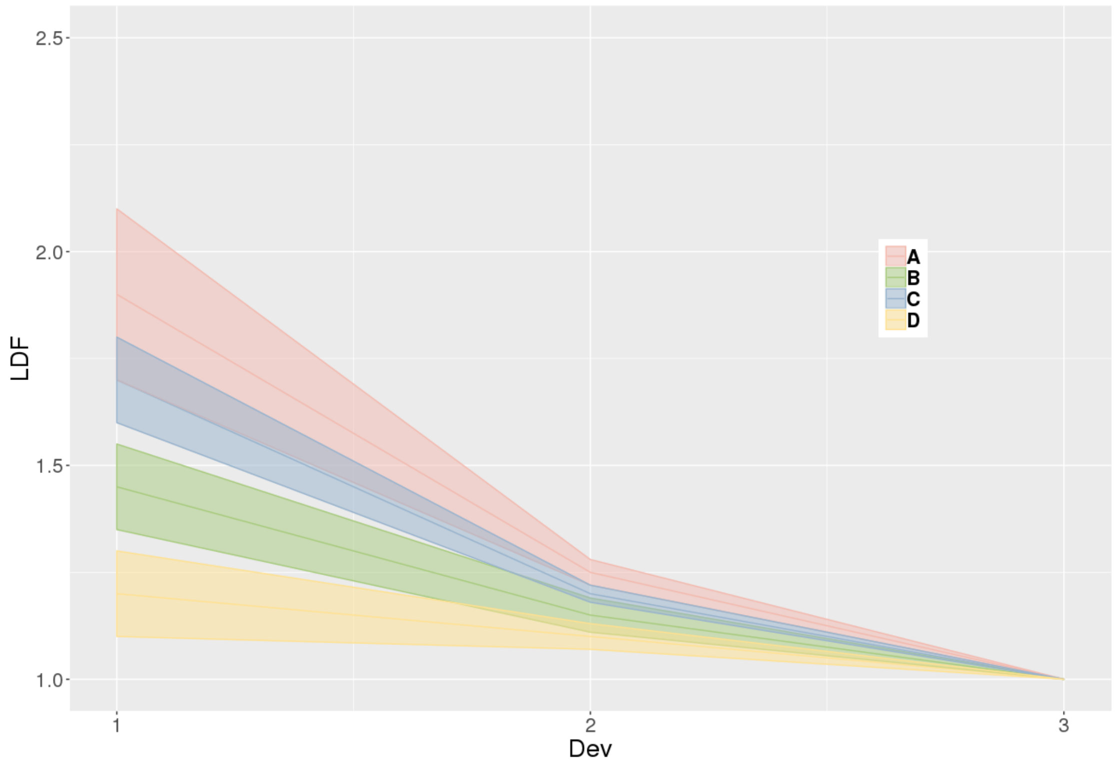


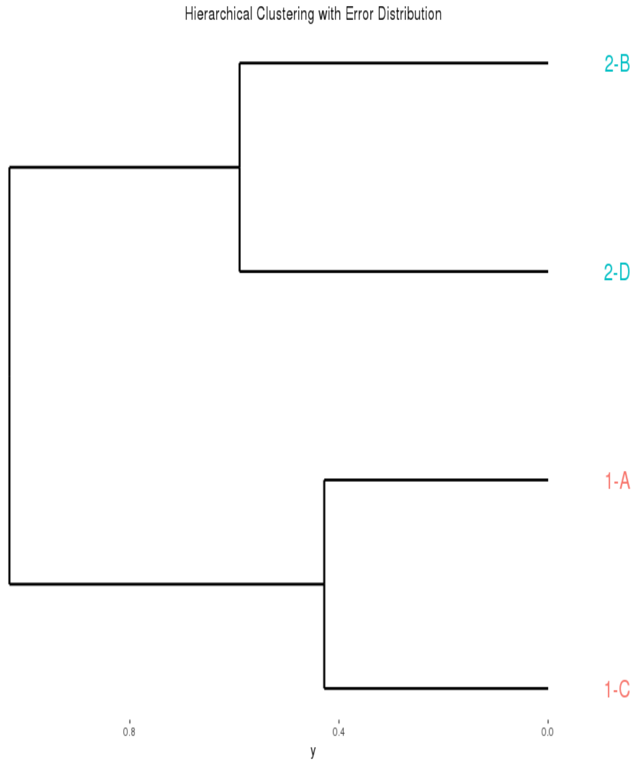


Under our standard clustering algorithm, the development patterns cluster into two bins:

(A C) and (B D)

Error-based Clustering – Scenario 1





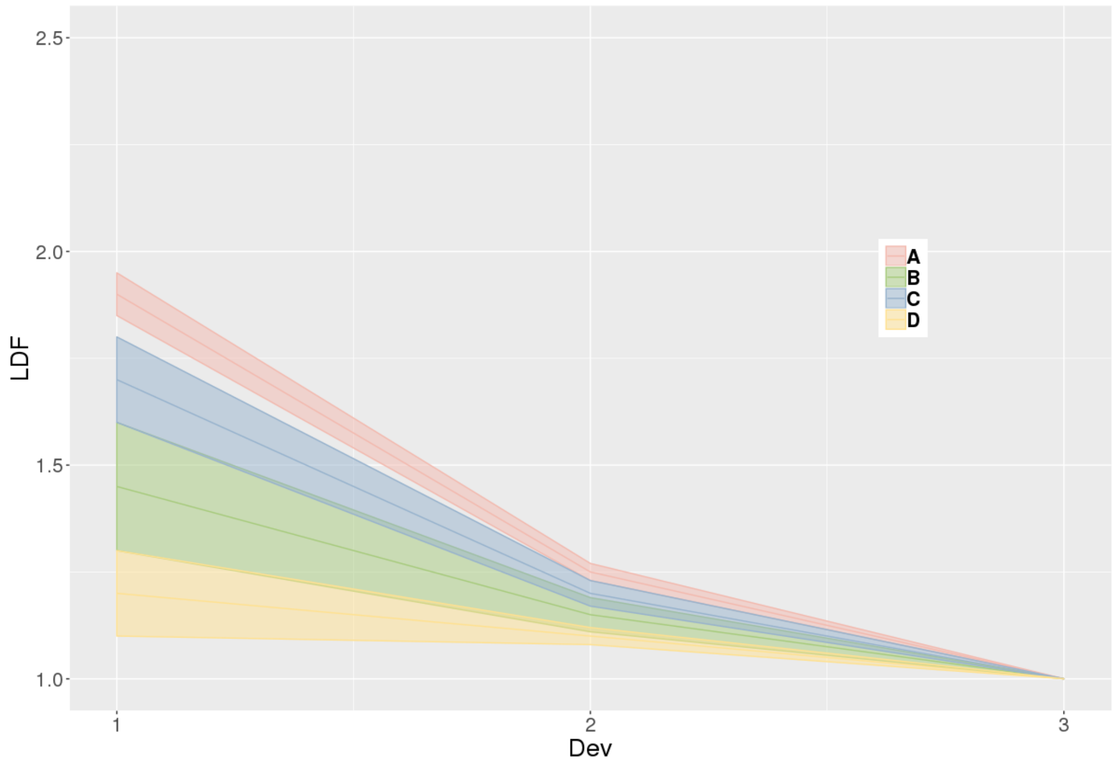
Error-based clustering

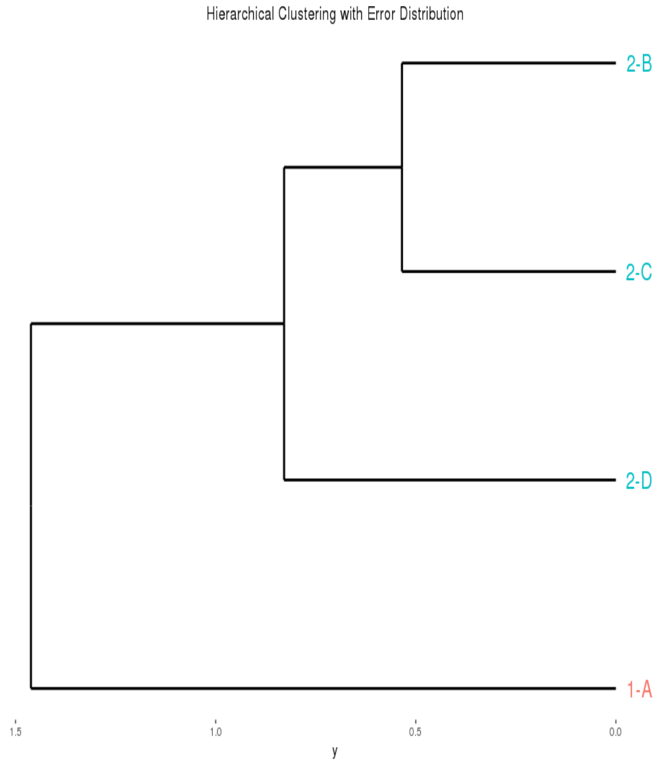
Same means

The patterns cluster into the same two bins:

(B D), (A, C)

Error-based Clustering – Scenario 2





Error-based clustering

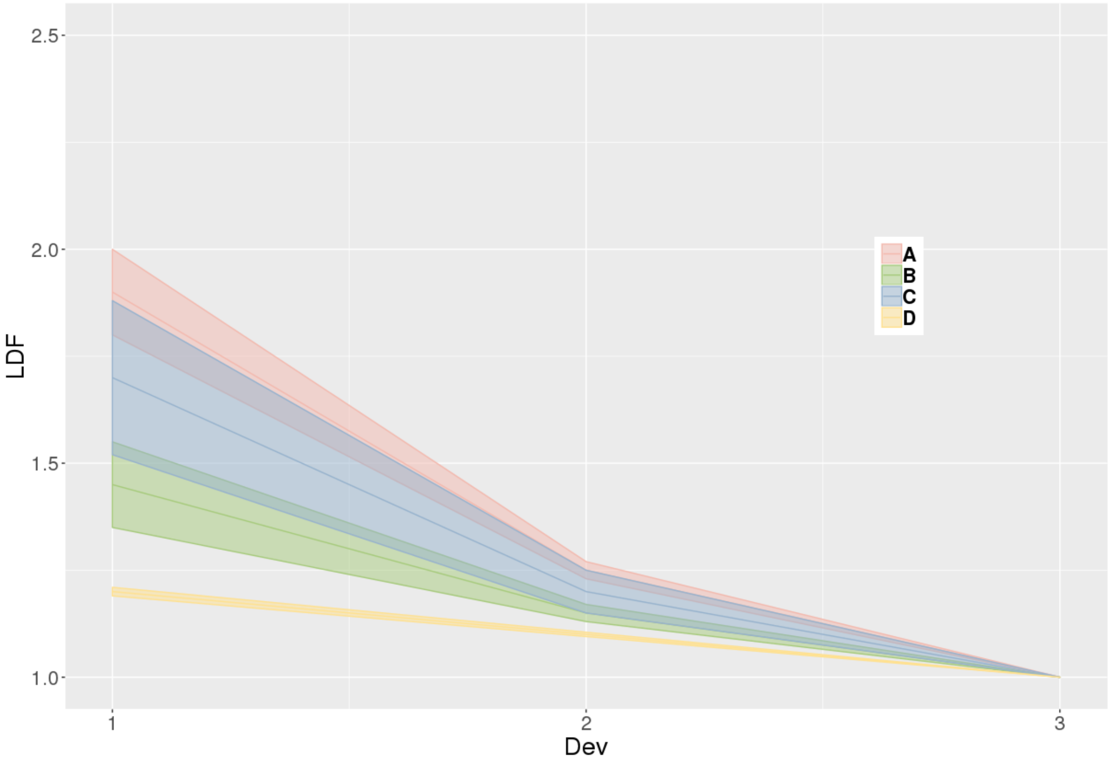
Same means

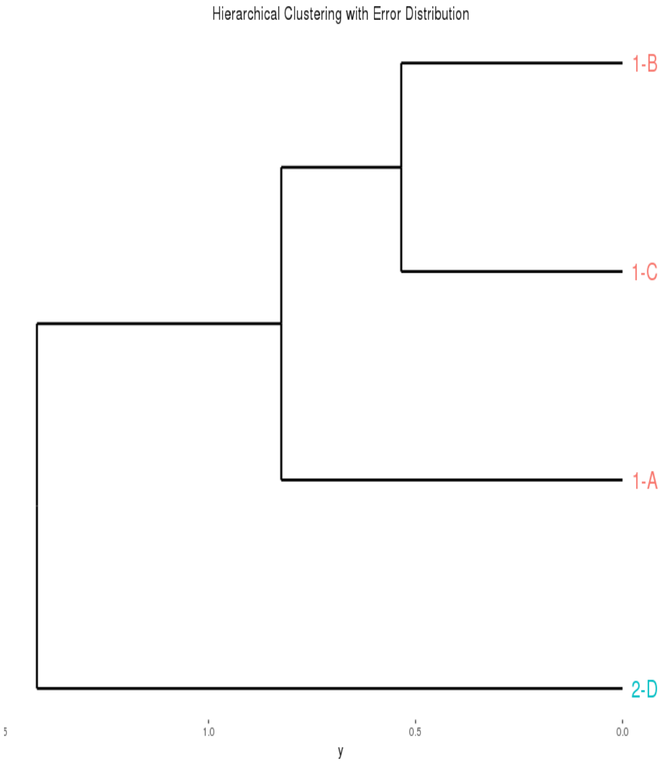
Increased error on B, less error on A

the patterns cluster into a new binning:

(A), (B C D)

Error-based Clustering – Scenario 3





Error-based clustering

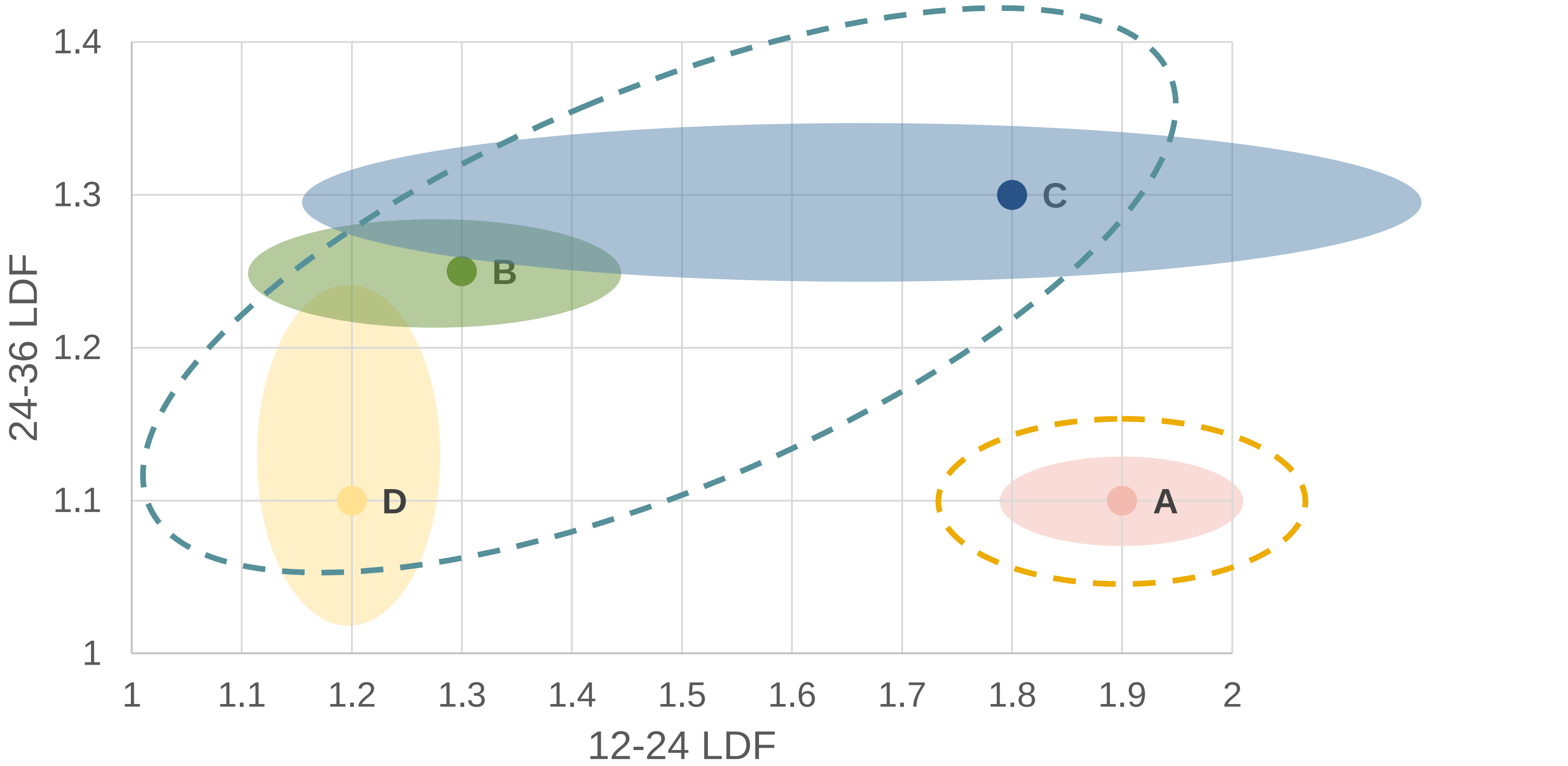
Same means

Increased error on C, less error on D

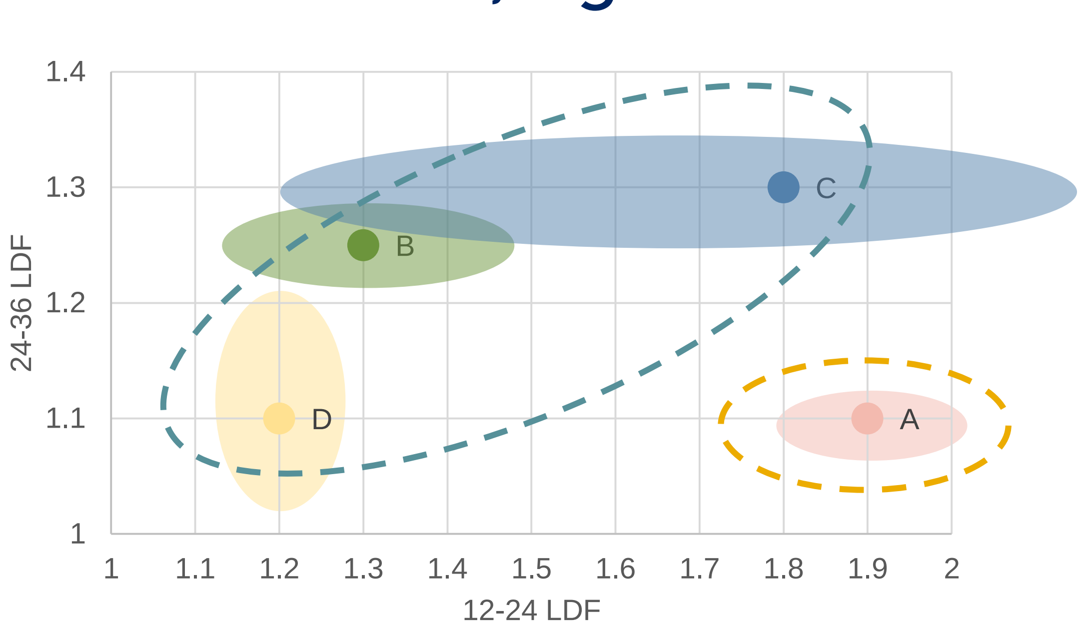
the patterns cluster into a new binning:

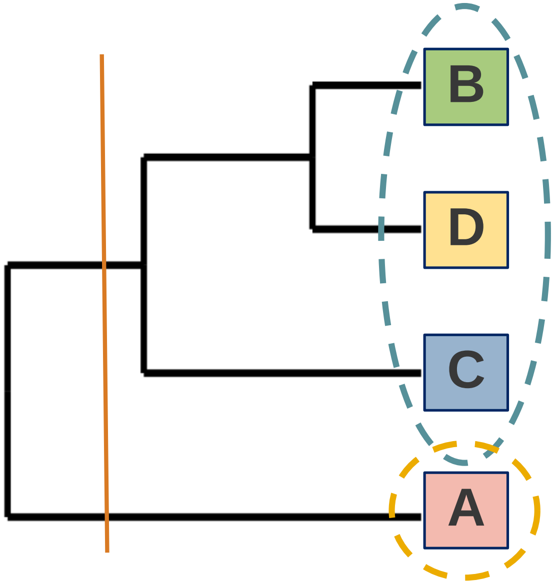
(A, B, C) (D)

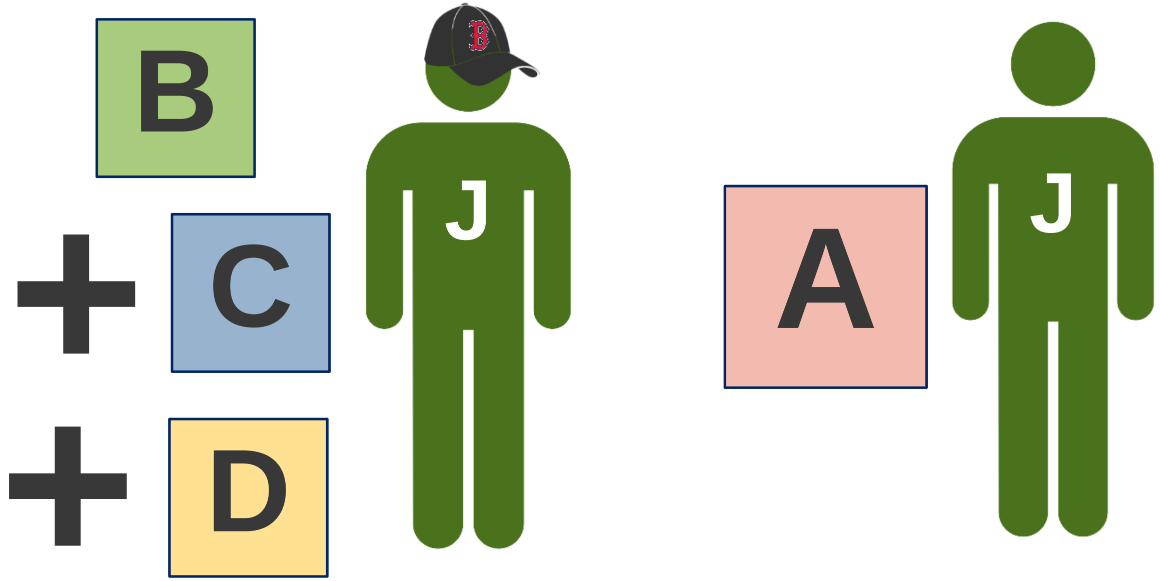
Considering the error distribution may give us a new recommendation



We’re done now, right?





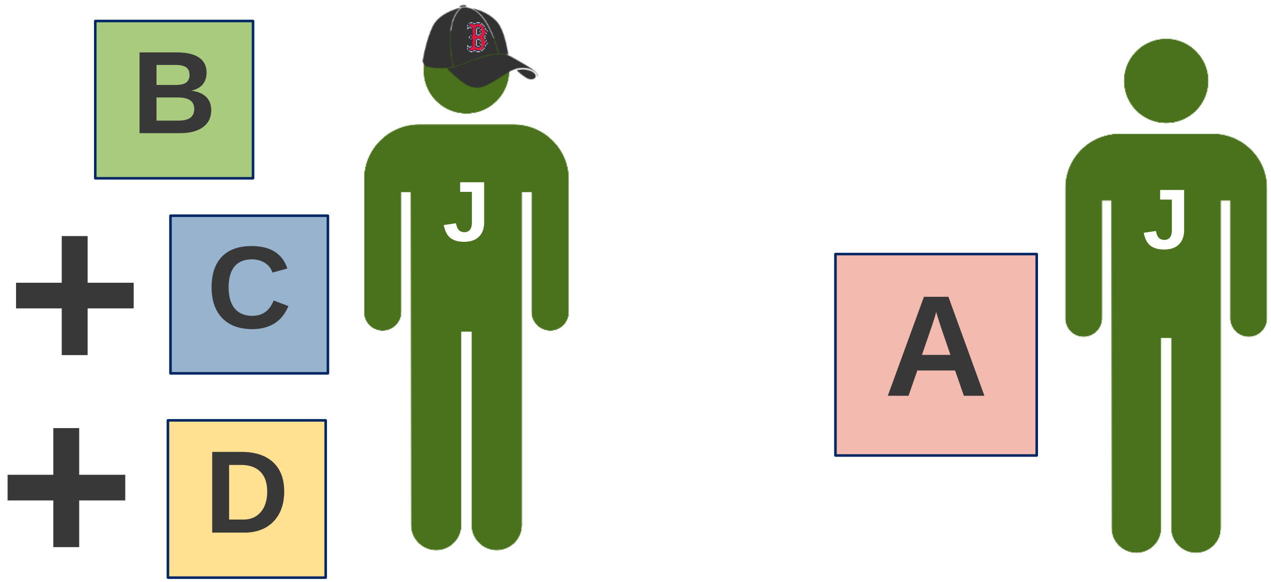


Agenda

- The Need for Segmentation
- Introduction to Clustering
- Clustering with Error Distributions
- **Stability Measurement**
- Summary and Implementation

Are we confident in the recommended segmentation?

- Do we trust the results of our clustering approach?
- Can we incorporate stochastic simulations?



Let's look at our segments one more time

A

B

C

D

Segment A

A

Loss Year	Age-to-Age			
	12-24	24-36	36-48	48-60
2013	1.96	1.11	1.00	1.00
2014	1.82	1.10	1.00	
2015	2.02	1.09		
2016	1.78			
Selected (Mean)	**1.90**	**1.10**	**1.00**	**1.00**

We can use **bootstrapping** to simulate more triangles for segment A

Iteration 1

Loss Year
2013
2014
2015
2016

Iteration 2

Loss Year
2013
2014
2015
2016

Iteration 3

Loss Year
2013
2014
2015
2016

Iteration 4

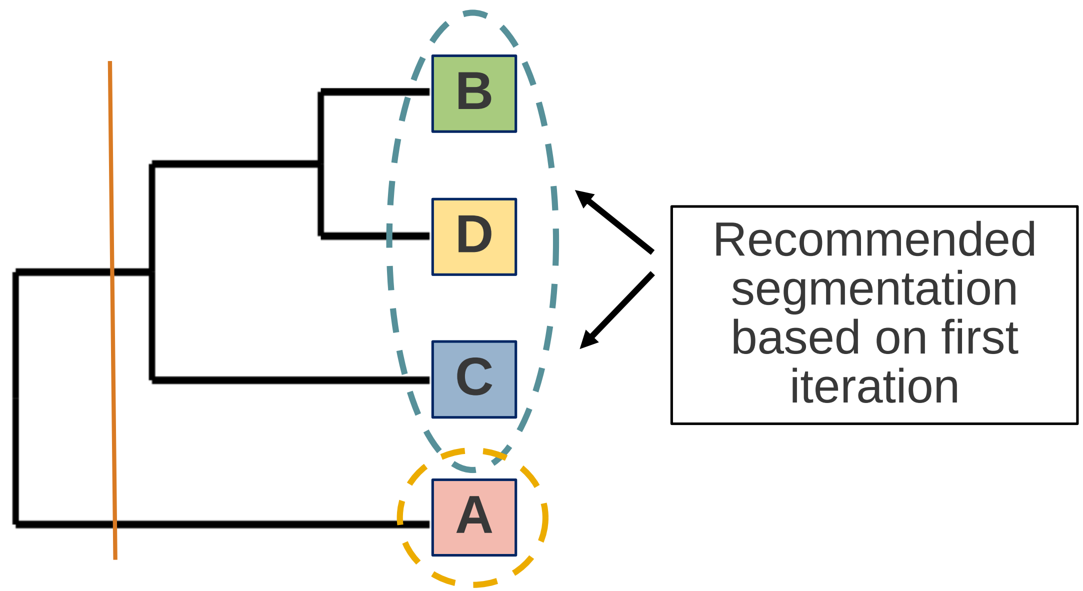
Loss Year	12-24	24-36	36-48	48-60
2013	1.93	1.13	1.00	1.00
2014	1.92	1.07	1.00	
2015	2.12	1.08		
2016	1.67			

Repeat for segments B-D

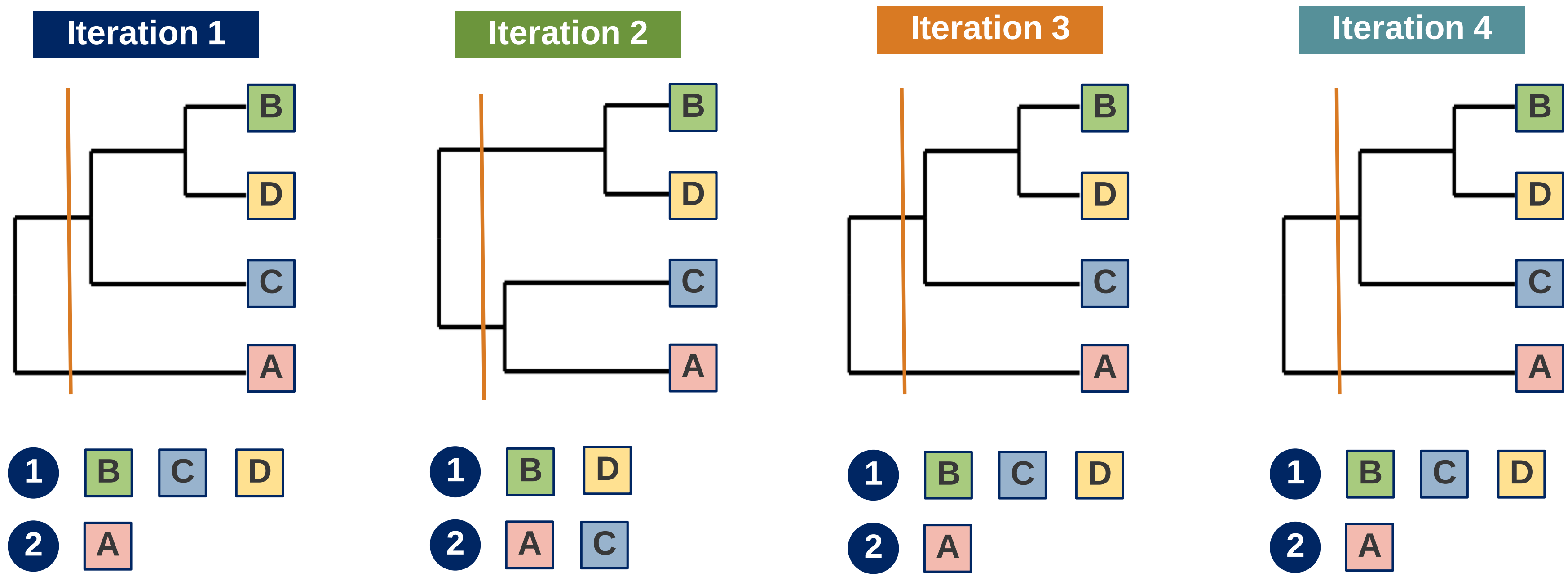
<table>
<tr><th>Iteration</th><th>A</th><th>B</th><th>C</th><th>D</th></tr>
<tr><td>1</td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
</tr>
<tr><td>2</td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
</tr>
<tr><td>3</td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
</tr>
<tr><td>4</td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
</tr>
</table>

Run the clustering algorithm on the first simulated group of triangles

<table>
<tr><th>Iteration</th><th>A</th><th>B</th><th>C</th><th>D</th></tr>
<tr><td>1</td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
<td><table><tr><th colspan="4">Age-to-Age</th></tr><tr><th>12-24</th><th>24-36</th><th>36-48</th><th>48-60</th></tr><tr><td>1.96</td><td>1.11</td><td>1.00</td><td>1.00</td></tr><tr><td>1.82</td><td>1.10</td><td>1.00</td><td></td></tr><tr><td>2.02</td><td>1.09</td><td></td><td></td></tr><tr><td>1.78</td><td></td><td></td><td></td></tr><tr><td>1.90</td><td>1.10</td><td>1.00</td><td>1.00</td></tr></table></td>
</tr>
</table>



Repeat for other iterations



Proposed segmentation

Cluster	Stability Score
B C D	0.75
A	0.75
B D	0.25
A C	0.25

Stability Score = probability of a cluster appearing

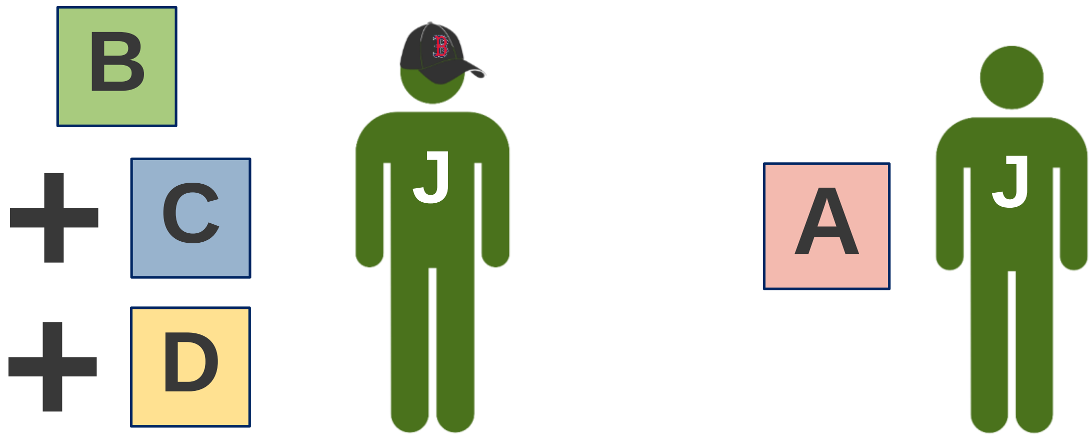
After many hundred iterations…

Cluster	Stability Score
B C D	0.81
A	0.81
B D	0.17
A C	0.17
other clusters	small

Rule of thumb:

- Stability score < 0.6: Unstable
- Stability score > 0.85: Highly stable

Fairly confident in our proposed segmentation, and very confident that B & D belong together



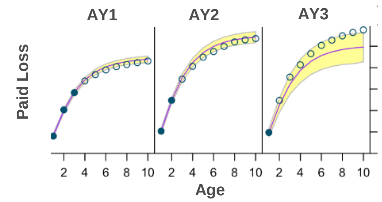
Agenda

- The Need for Segmentation
- Introduction to Clustering
- Clustering with Error Distributions
- Stability Measurement
- **Summary and Implementation**

Project Summary

Development Factors

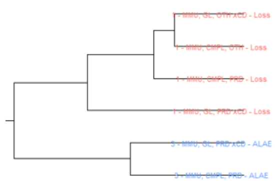
- Development factors and error ranges computed using actuarial techniques



Error distributions quantify uncertainty in development

Clustering Algorithm

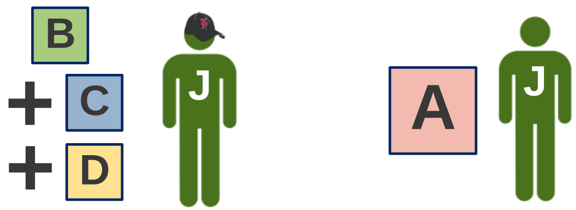
- Clustering with consideration for error distribution
- Stability score indicates confidence in recommendation and guidance into optimal number of clusters



Dendrogram displays clustering output

Recommendation

- Purely statistical recommendation for optimal segmentation



Implementation

- What we discussed today was a very simplified example of how this process works…in reality there are <u>**a lot of moving parts**</u>:

Segment definitions (market, coverage, policy & claims characteristics)

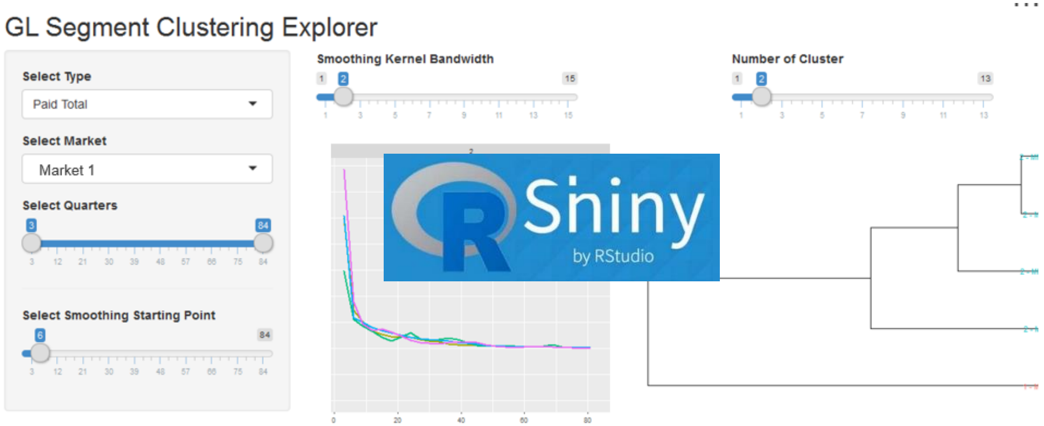
Accident years and development ages to incorporate

Methods to calculate LDFs and error ranges

Smoothing of LDFs

Clustering inputs (hierarchical vs. k-means, distance, linkage)

- We built a flexible **R Shiny** tool to accommodate all these varying inputs …



Segmentation recommendation based on desired inputs

References

- Kumar, Mahesh and Patel, Nitin R., “Clustering Data With Measurement Error”. Rutcor Research Report 12-2005

- Zumel, Nina and Mount, John. Practical Data Science with R. March 2014.

Thank You!

- Yue Hu
- Dan Stering
- Emily Allen
- Michaela Porter

Questions?

Appendix

Clustering – Distance Measurement and Linkage Criteria

- Distance measures – Metric used to determine pairwise distances
  - Euclidean (almost always)
  - Manhattan or City Block
  - Mahalanobis

- Linkage – Criterion used to determine distance between sets of observations

Linkage example

Distance between points

Using Euclidean metric

	1	2	3
2	11.5		
3	16.7	18.2	
4	9.5	20.5	16

- Clearly points 1 and 4 form the first cluster, since they are the closest two points
- Now we consider various options for finding distance between these points (1 and 4) and the remaining ones (2 and 3)

Linkage example – Complete (max)

Distance between points

	1	2	3
2	11.5		
3	16.7	18.2	
4	9.5	20.5	16

- For the Complete linkage we find the maximum distance between the cluster and the candidate point
- Point 3 is selected, since it is closest under this criteria

	Distance from 1	Distance from 4	
Point 2	11.5	20.5	**20.5**
Point 3	16.7	16	**16.7**

Linkage example – Single (min)

Distance between points

	1	2	3
2	11.5		
3	16.7	18.2	
4	9.5	20.5	16

- For the Single linkage we find the minumum distance between the cluster and the candidate point
- Point 2 is selected, since it is closest under this criteria

	Distance from 1	Distance from 4	
Point 2	11.5	20.5	11.5
Point 3	16.7	16	16

Linkage example – Average

Distance between points

	1	2	3
2	11.5		
3	16.7	18.2	
4	9.5	20.5	16

- For the Average linkage we find the mean distance between the cluster and the candidate point
- Point 2 is selected, since it is closest under this criteria

	Distance from 1	Distance from 4	
Point 2	11.5	20.5	**16**
Point 3	16.7	16	**16.35**

Dendograms

- Depending on the metric used to measure distance between clusters (Euclidean, etc.) and a linkage metric used (simple, etc.), the height of a branch indicates those observations linked by the branch have a distance from each other less than or equal to the height of the branch.

- Euclidean distance between members no greater than height of branch, using complete linkage
- Can ‘cut’ the tree at any given height based on how close we require the members to be.
- If we have our clustering based on Euclidean distance, using a complete (max) linkage, and we cut at a given height, we can say that the maximum Euclidean distance to all other members of the cluster is less than the cut point.
- The longer this distance is between changes in clusters, the more distinct the groupings

Error-based clustering

- Error-based clustering is equivalent to standard clustering (Euclidean) if the errors associated with all data points are the same, and if the errors of variables of a data point are the same and uncorrelated.

- Mahalanobis distance
  - Used to find distance from a distribution to a point
    - Transform variables into uncorrelated variables
    - Set their variances equal to 1
    - Calculate simple Euclidean distance
    - Tie in with Principal Components
  - Can be extended to distance between two clusters – sum of Sigmas