UNT
UNIVERSITY OF NORTH TEXAS®

Cluster Analysis Using K-Means

Rion McDonald
Senior Data Analyst

UNT
UNIVERSITY OF NORTH TEXAS®

Cluster Analysis

1. Used to subdivide observations into groups with similar characteristics

2. Used to perform *unsupervised* machine learning

3. Covers a variety of methods/algorithms

UNT
UNIVERSITY OF NORTH TEXAS®

K-Means Clustering

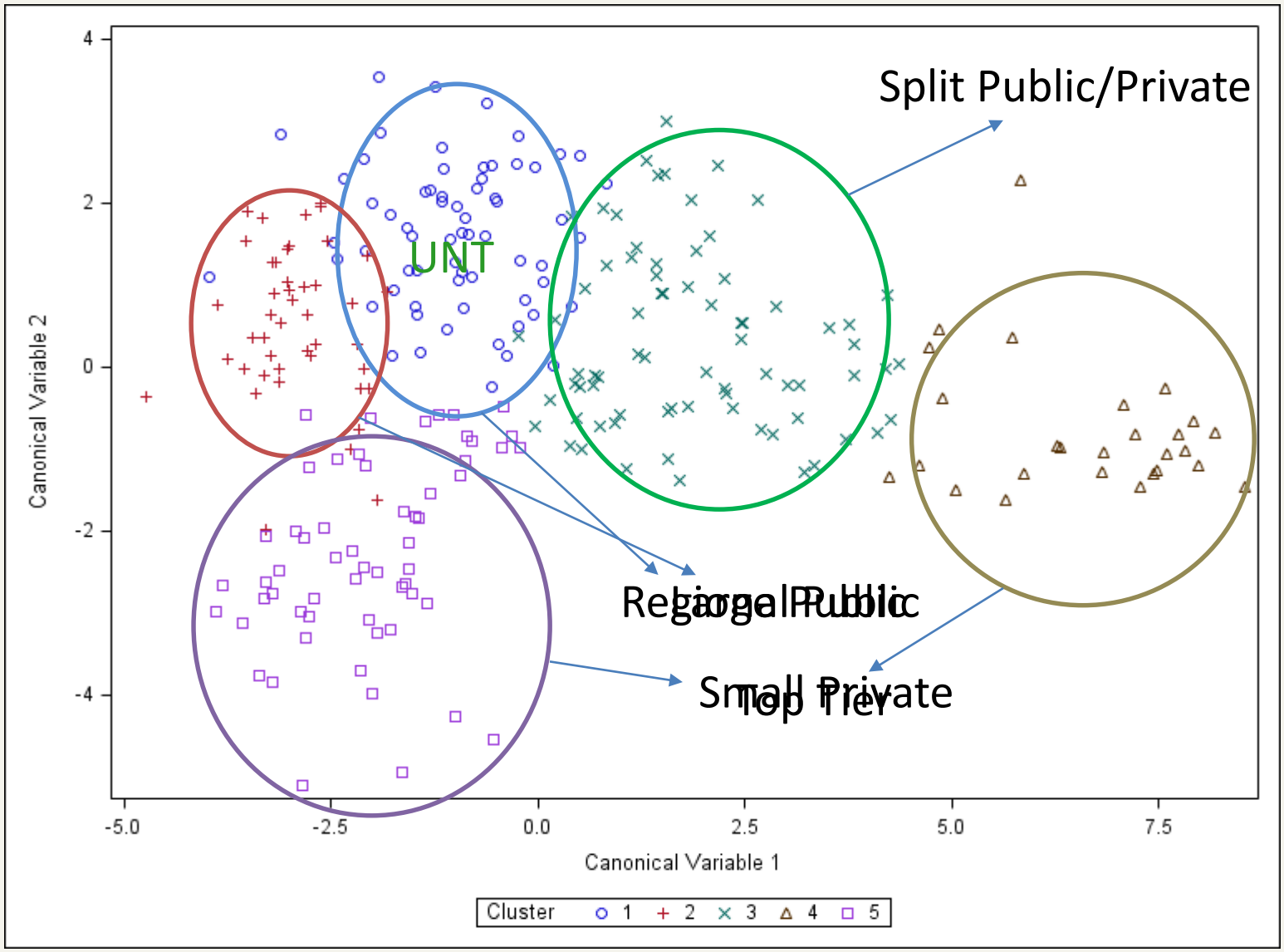
1. Separates data into *k* groups pre-defined by analyst

2. Identifies groups such that…
   a) Group members as similar to other members as possible
   b) Group members as distinct from other groups as possible

3. Processes large datasets well

UNT
UNIVERSITY OF NORTH TEXAS®

U.S. News Rankings: School Clusters (Rank >= UNT)



UNT
UNIVERSITY OF NORTH TEXAS®

K-Means: Data Preparation

1. Use continuous numerical data

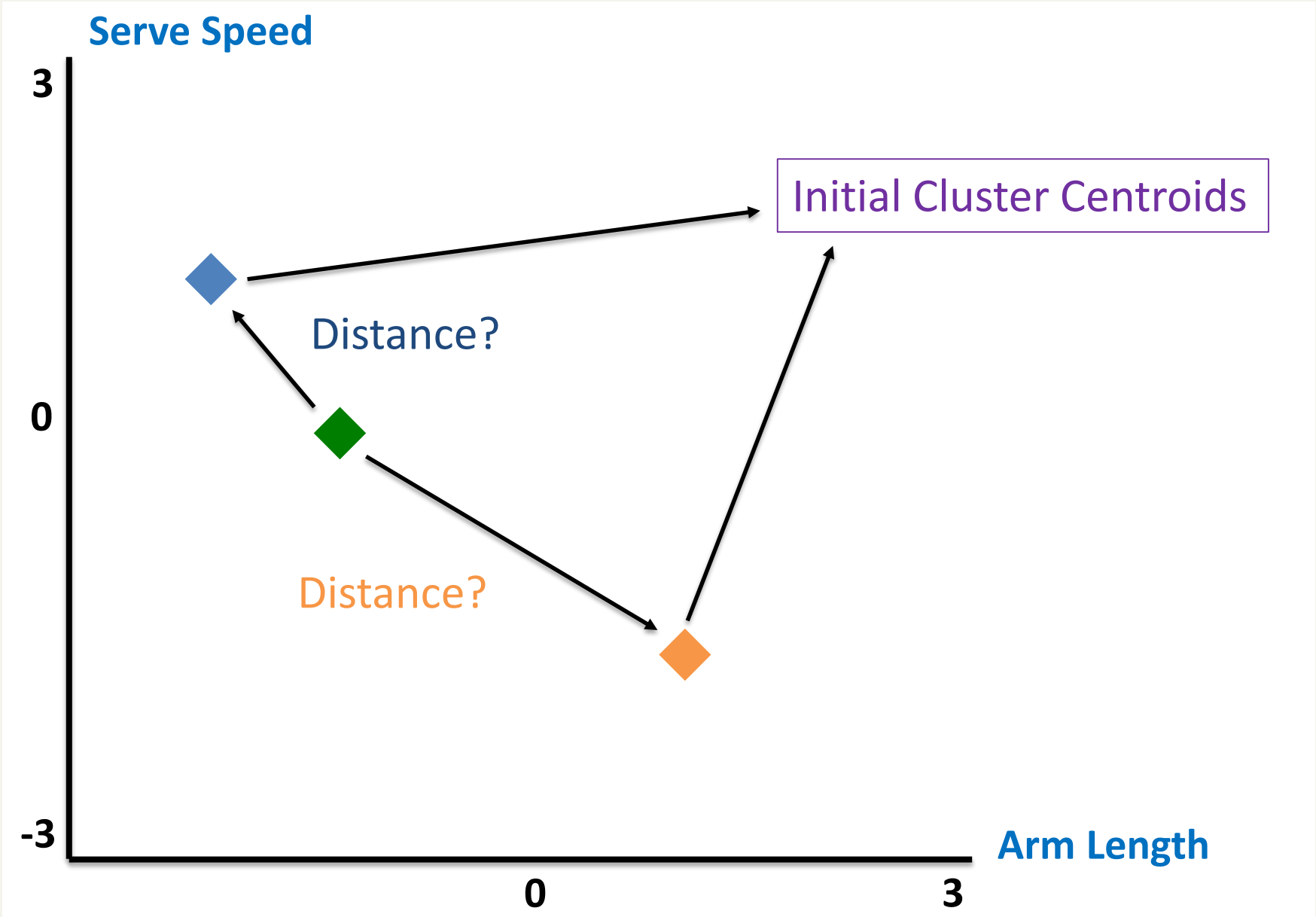
2. Standardize each variable before analysis

Standard Score =

$$\frac{\textit{Observation Value} - \textit{Variable Mean}}{\textit{Variable Standard Deviation}}$$

UNT
UNIVERSITY OF NORTH TEXAS®

K-Means Procedure (*K*=2)



UNT
UNIVERSITY OF NORTH TEXAS®

Euclidean Distance

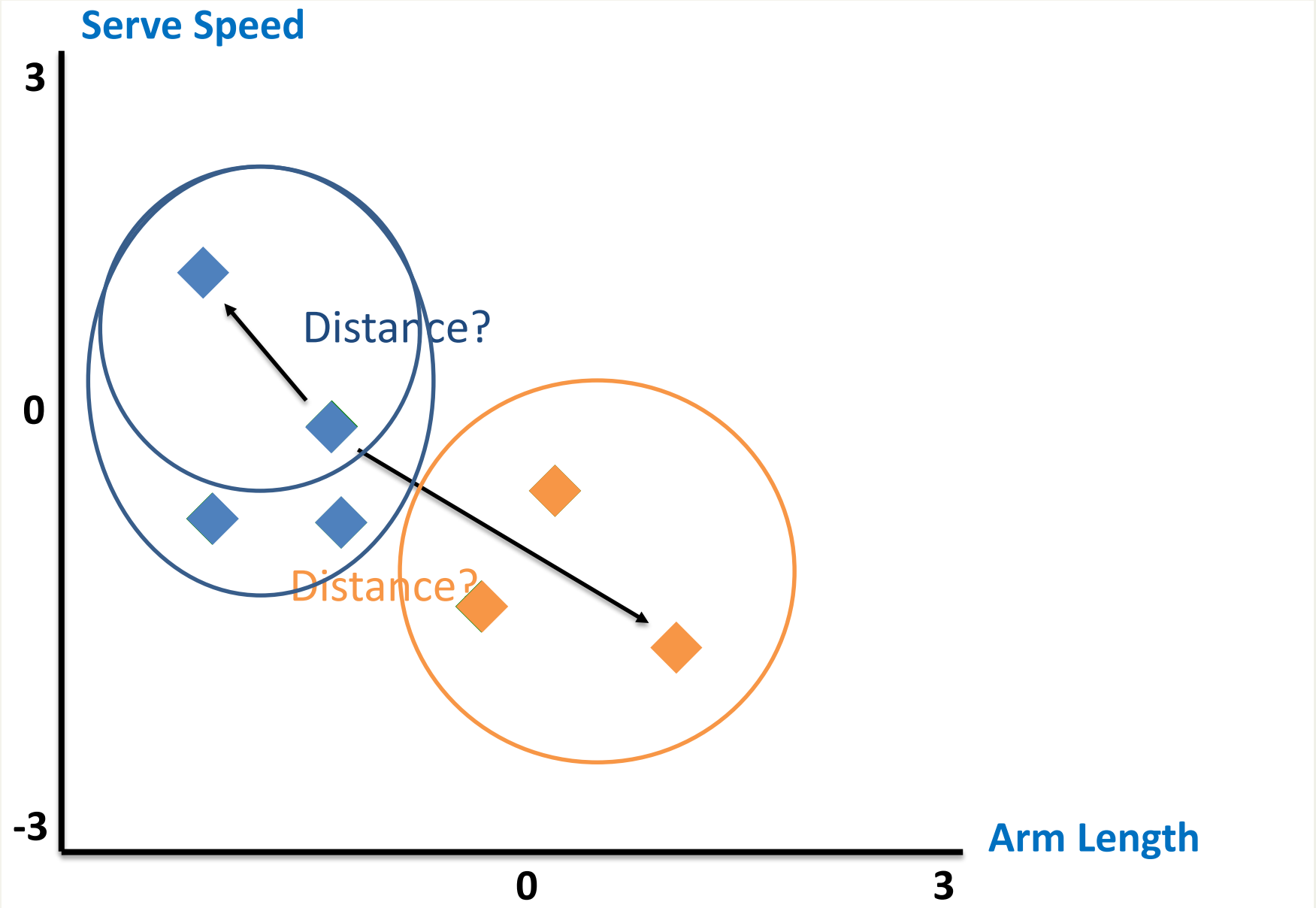
Distance =

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

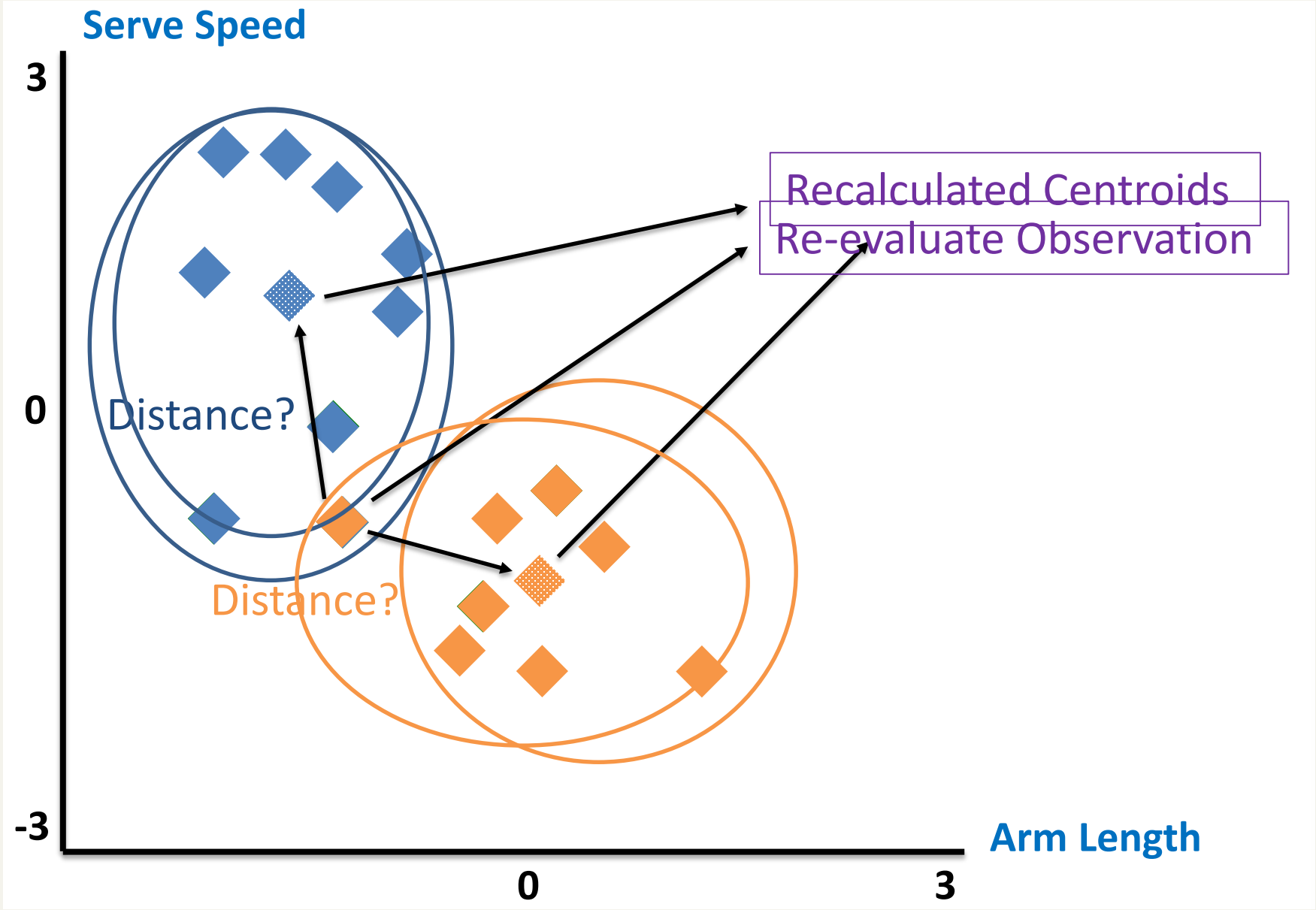
$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 \ldots + (n_1 - n_2)^2}$$

UNT
UNIVERSITY OF NORTH TEXAS®

K-Means Procedure (*K*=2)



K-Means Procedure (k=2)



UNT
UNIVERSITY OF NORTH TEXAS®

Finding the Correct Number of Clusters (*K*)

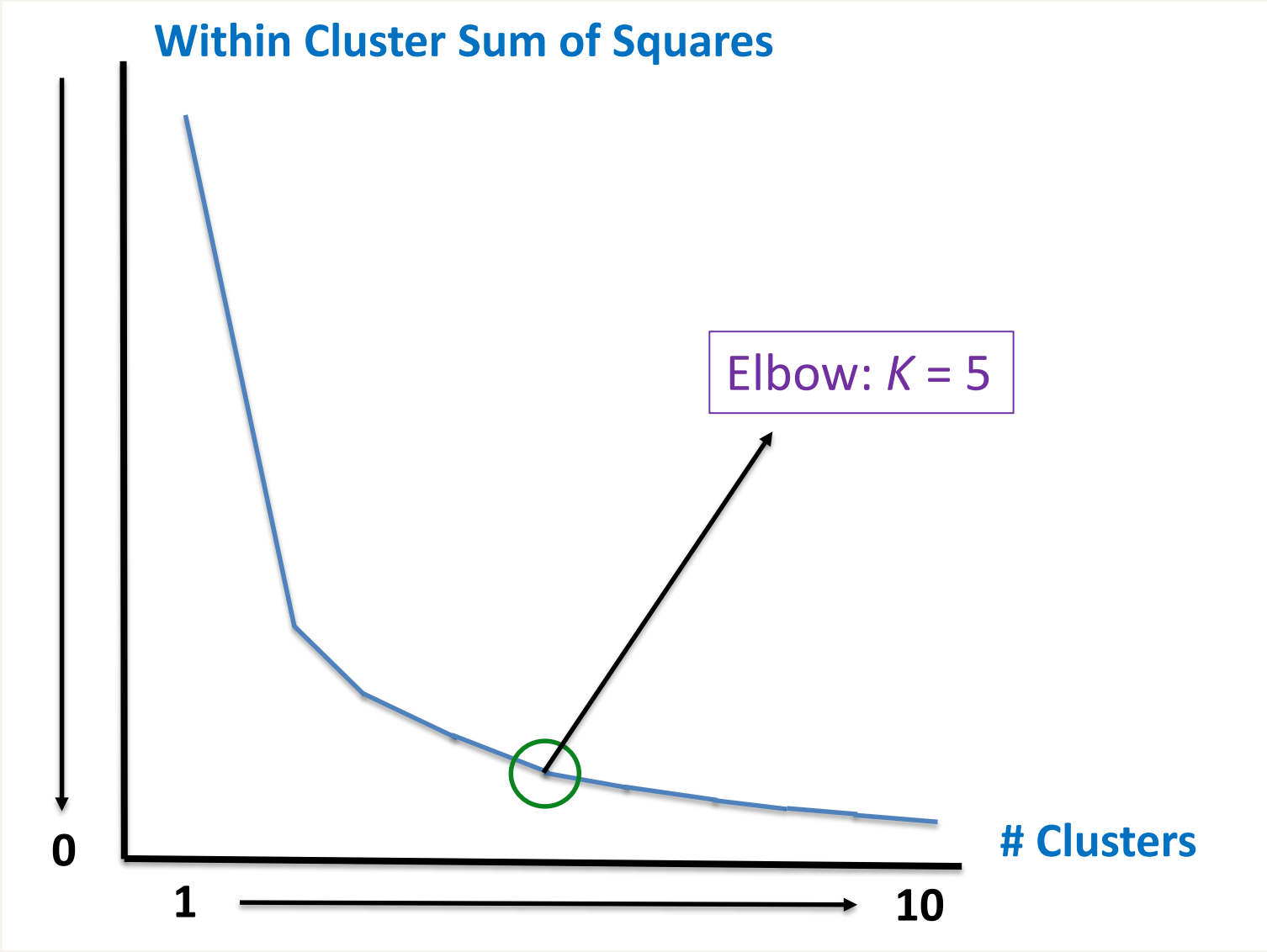
1. Try different # of clusters
2. Examine fit/performance measure(s)
3. Exercise judgment in regard to accuracy vs. complexity

UNT
UNIVERSITY OF NORTH TEXAS®

Example of Evaluation Statistics for a Given *K*

Variable	Within Cluster Standard Deviation	R-Squared
Retention Rate	0.47316	77.9%
Peer Assessment	0.48453	76.8%
Student-Faculty Ratio	0.70031	51.6%
Alumni Giving Rate	0.60083	64.4%
.....	.....	.....
Overall	0.63016	60.8%

Sum of Squared Distance Graph



UNT
UNIVERSITY OF NORTH TEXAS®

Delaware Cost Study Data

1. National study began in 1992

2. Productivity/Cost measures tied to degree programs

3. External benchmarking by classification, program characteristics, and peer group

4. Internal benchmarking opportunities available

UNT
UNIVERSITY OF NORTH TEXAS®

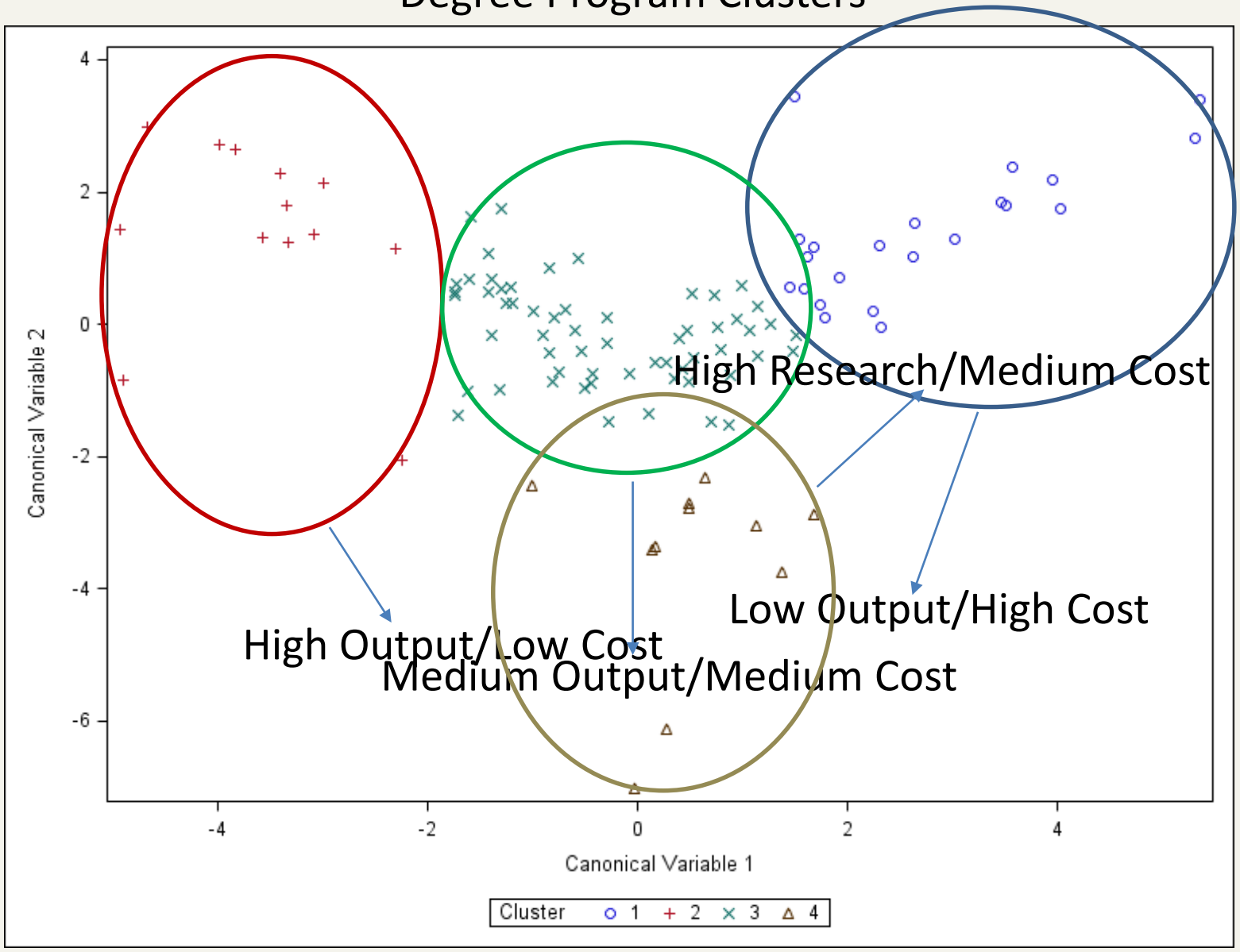
Cost Study Cluster Data

By program…

1. Total yearly degrees awarded (three-year average)
2. Total yearly SCH
3. Annual instructional cost per SCH
4. Annual research expenditures per tenure/tenure track faculty FTE

UNT
UNIVERSITY OF NORTH TEXAS®

Degree Program Clusters



UNT
UNIVERSITY OF NORTH TEXAS®

Program Cluster Averages

Cluster	Degrees	SCH	Instructional Expense	Research Expenditures
Low Output/ High Cost *(n=22)*	25	2,391	$554	$7,466
High Output/ Low Cost *(n=13)*	199	28,848	$148	$23,822
Medium Output/ Medium Cost *(n=60)*	62	8,030	$206	$11,770
High Research/ Medium Cost *(n=12)*	21	3,699	$237	$87,947

UNT
UNIVERSITY OF NORTH TEXAS®

Thank You

rion.mcdonald@unt.edu