Comparison of Hyper-dimensional LSA Spaces for Semantic Differences

John C. Martin

Dissertation Defense

20 May 2016

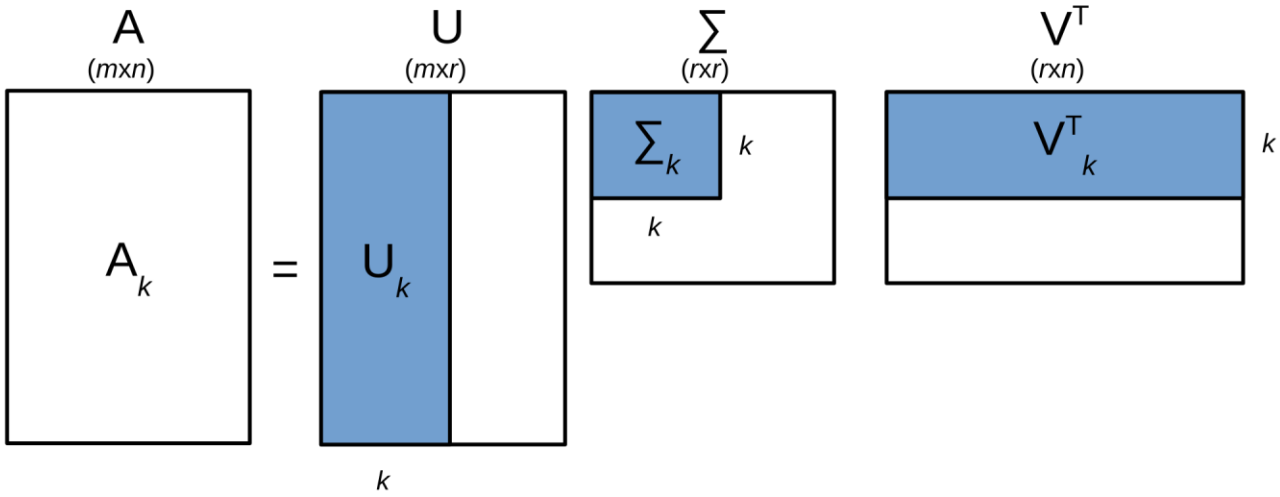
Overview

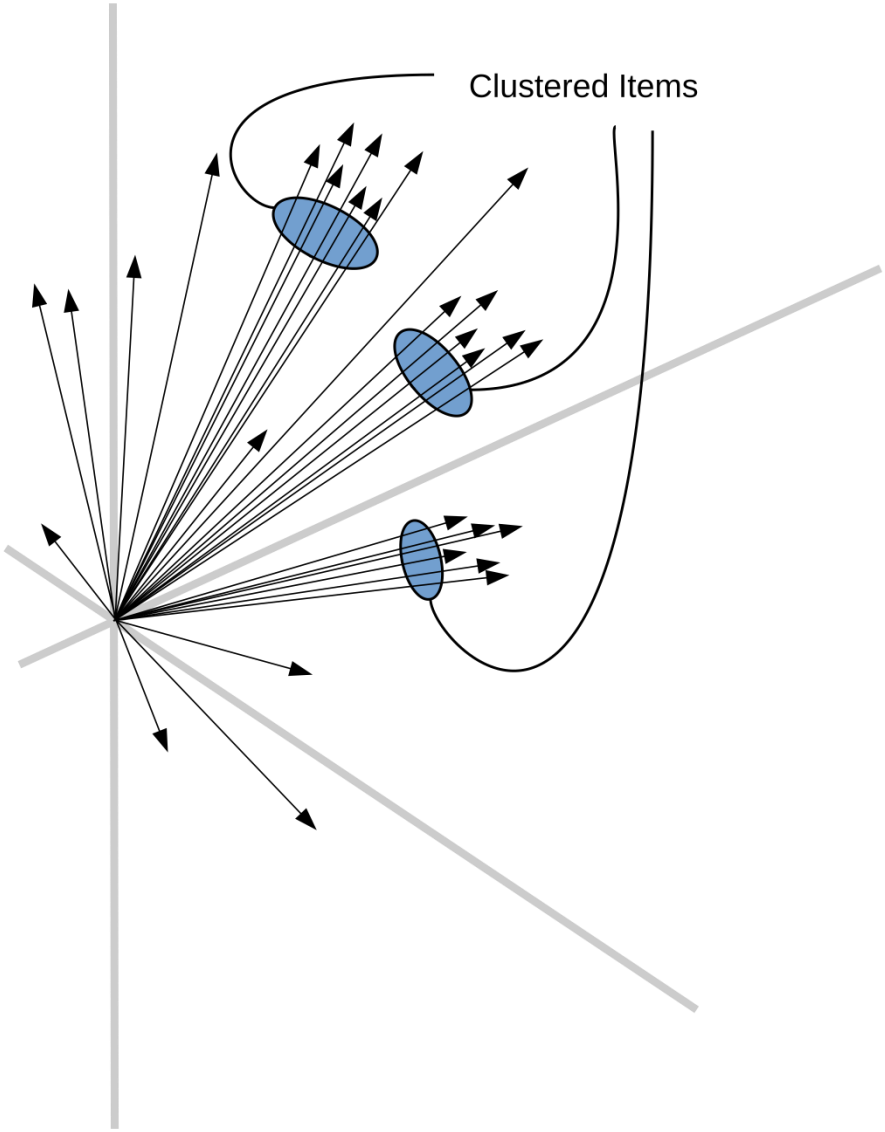
- Review LSA model of learning
  - What is meaning?
- Measures
- Experiments
- Semantic Measurement Model
- Q & A

The LSA Model of Learning

- Orthogonal Axes
- Dimensionality Reduction
- Mapping System

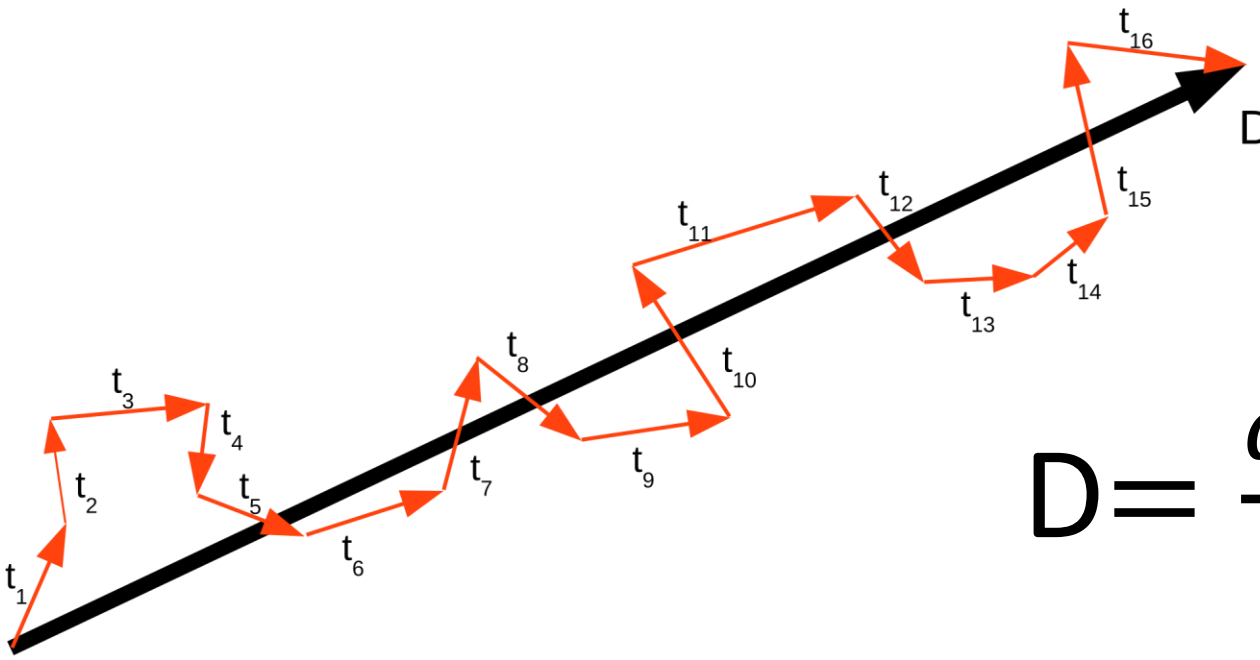
Meaning


A
(mxn)
U
(mxr)
Σ
(rxr)
V^T
(rxn)
A_k
=
U_k
Σ_k
V^T_k
k
k
k
k


Clustered Items

Compositionality Constraint

The meaning of a document is the sum of the meaning of its words



$$D = \frac{q^T U_k}{\sum_k}$$

Compositionality Constraint Corollary

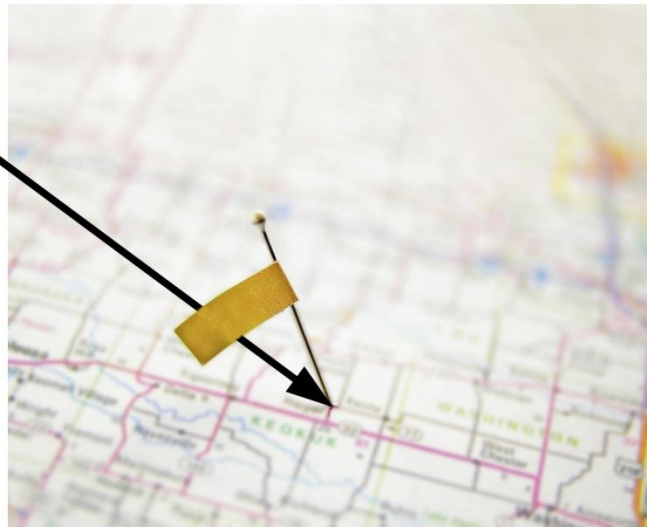
The meaning of a word is defined by the documents in which it appears

(and does not appear)

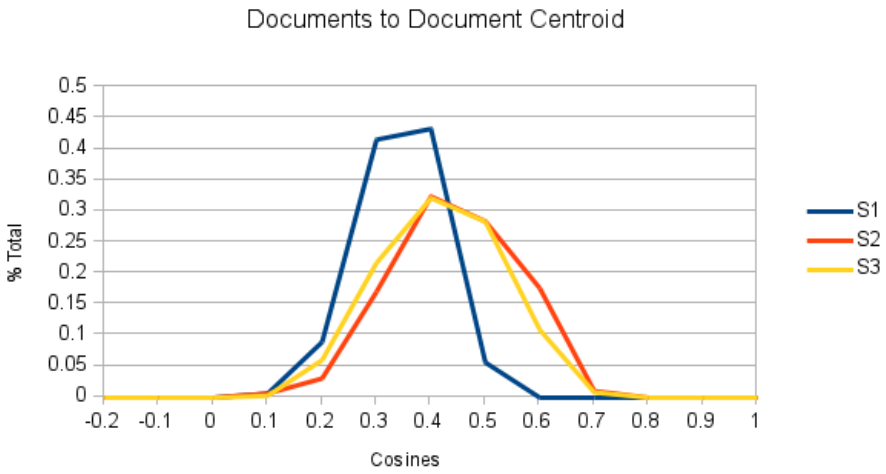
Meaning

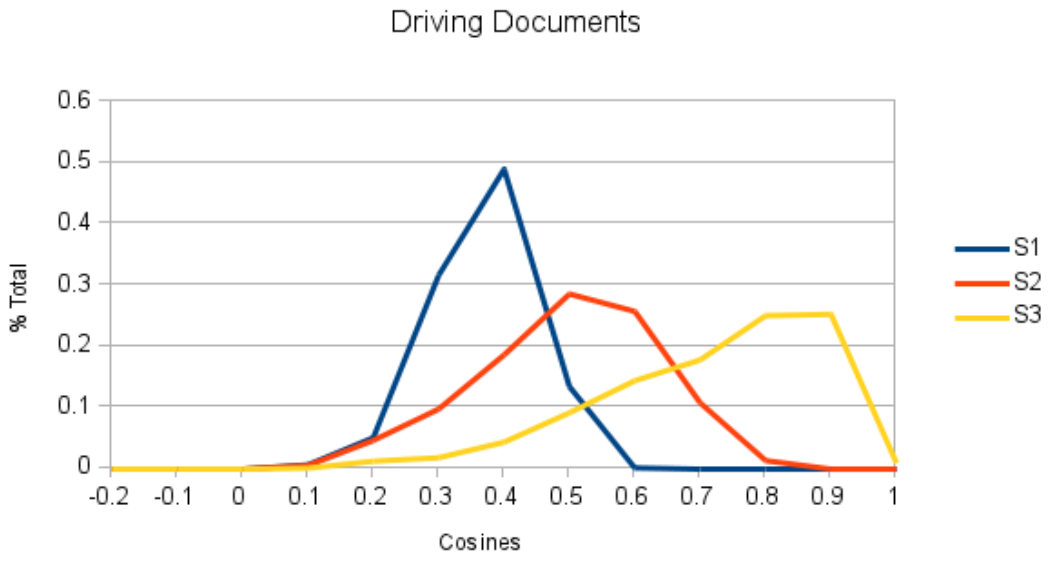
The Mapping system consists of:

- Term Vector Dictionary
- Singular Values



Motivation


Documents to Document Centroid
% Total
Cosines
S1
S2
S3


Driving Documents
% Total
Cosines
S1
S2
S3

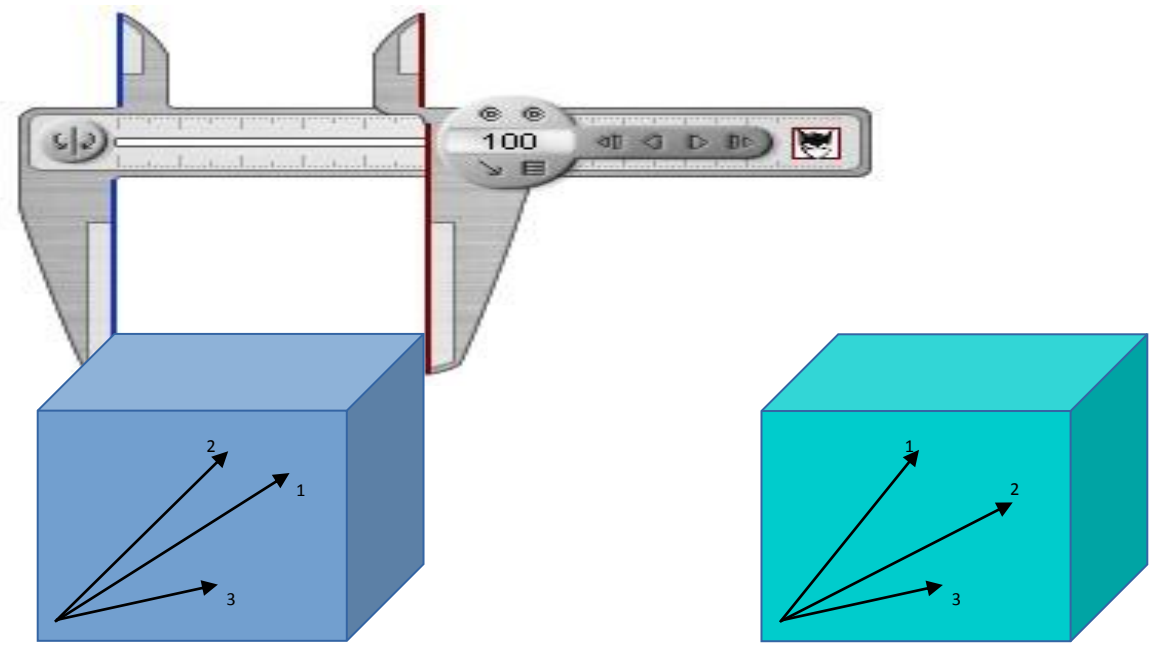
Objective

Find a measure or set of measures that can quantify the difference between two spaces

Measures

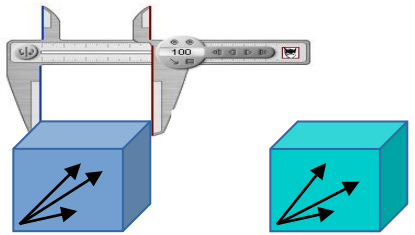
- Direct Comparison
- Projected Content Comparison
- Rotated Item Comparison

Direct Comparison Measures

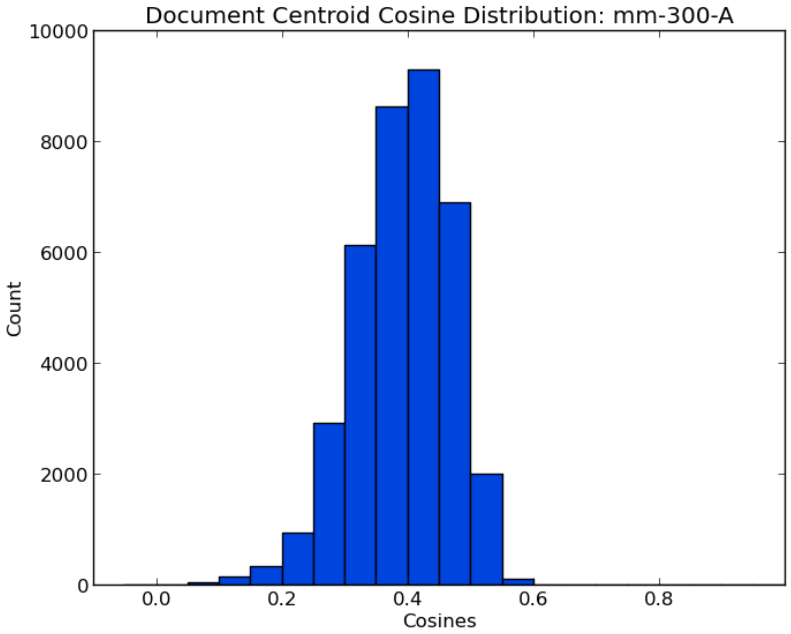


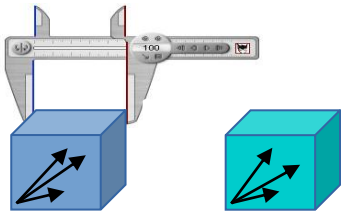
Individual Space Measures

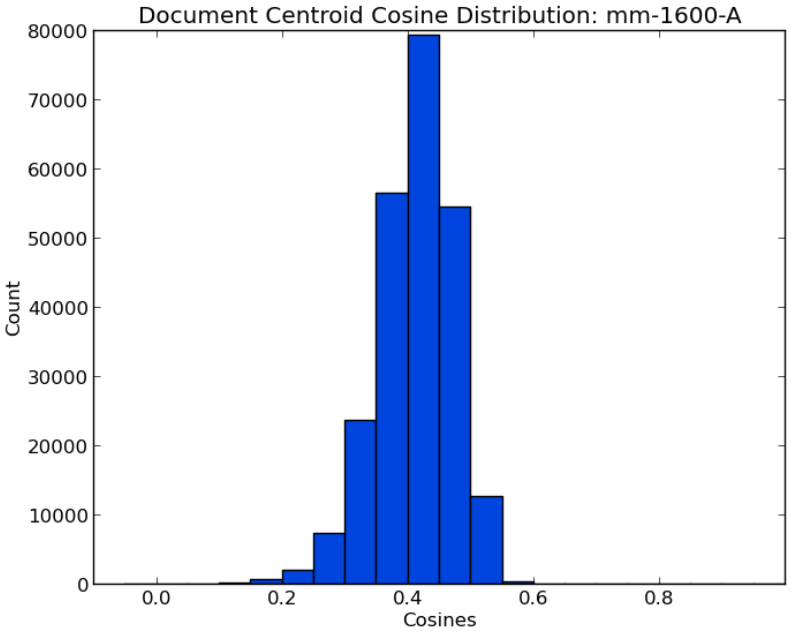
- Document Count
- Term Count
- Non-zeroes



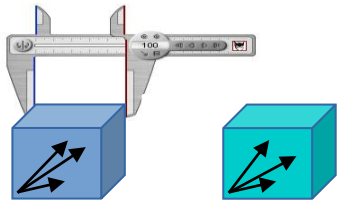
Distribution Analysis

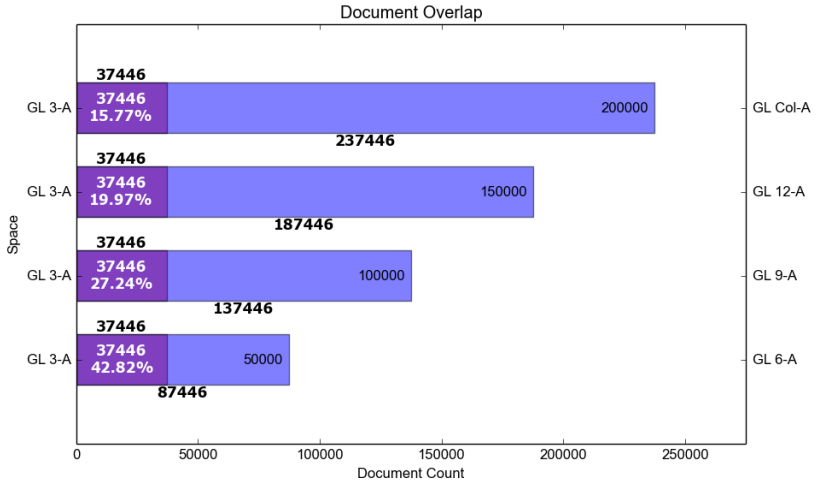


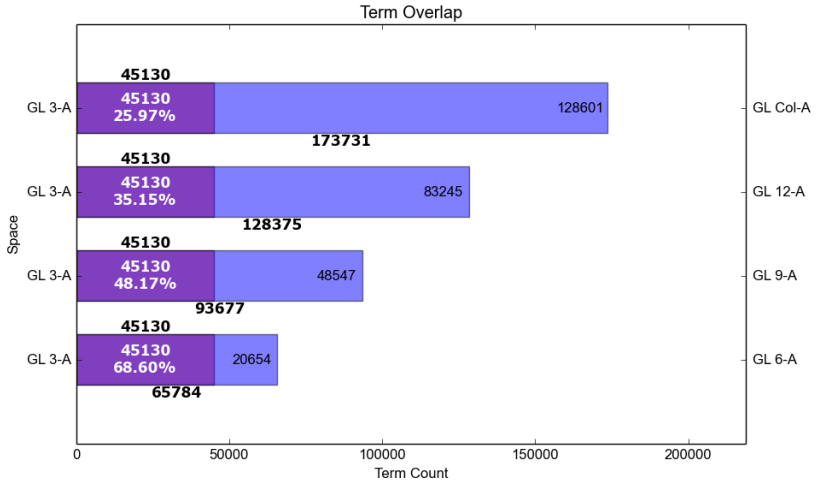




Term and Document Overlap

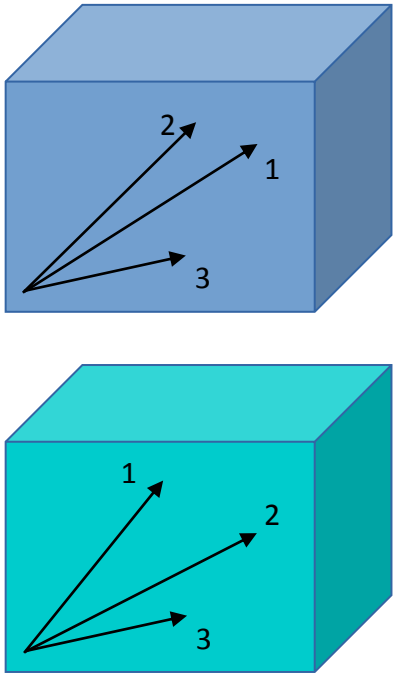




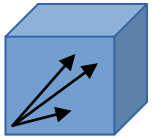


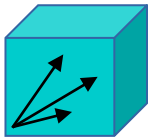
Projected Content Comparisons

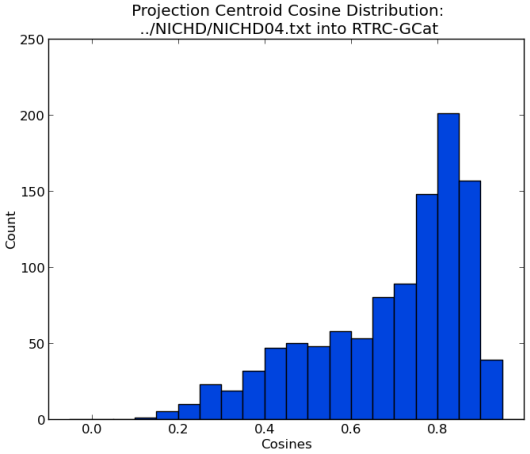
Matched items projected into each space

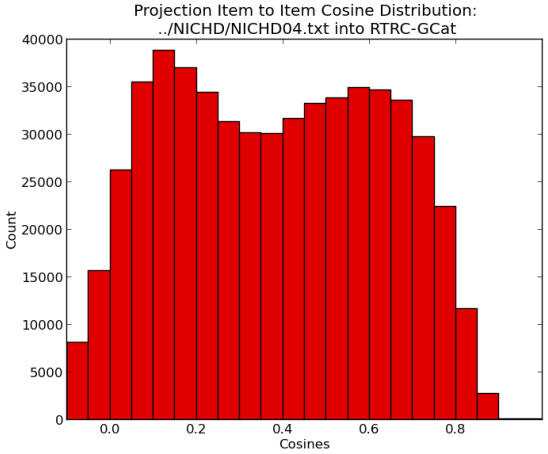


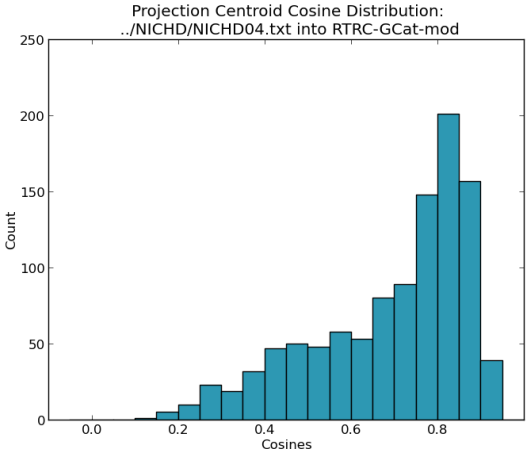
Projected Item Distribution

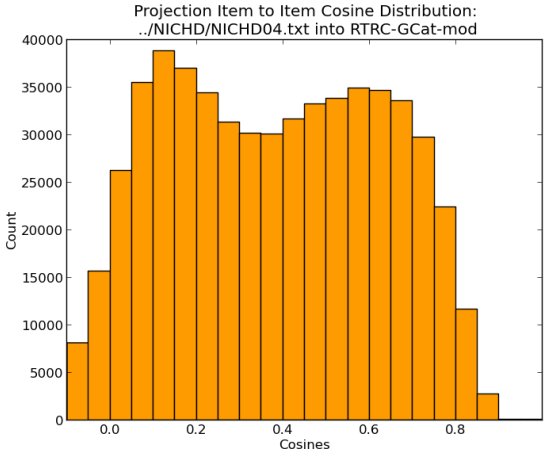




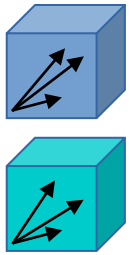


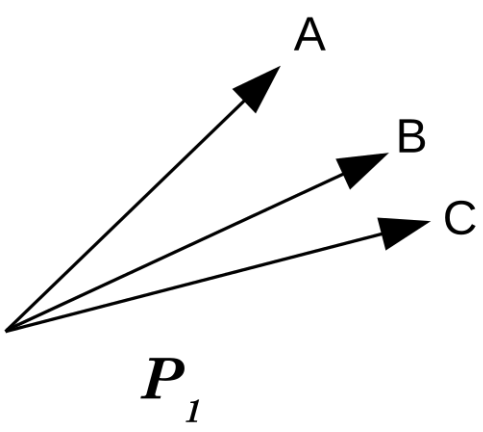


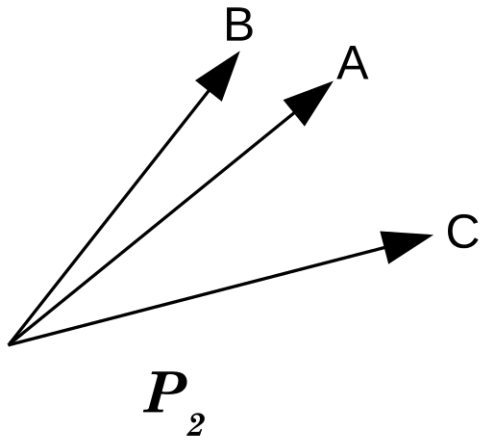




Three-Tuple Comparisons

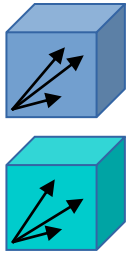


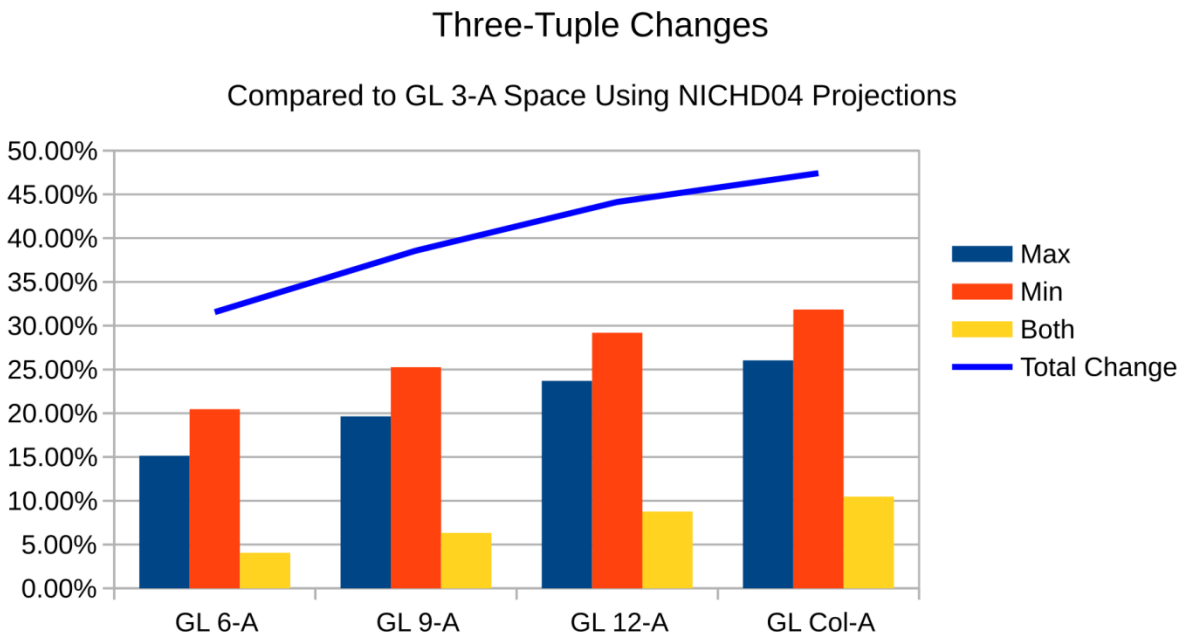




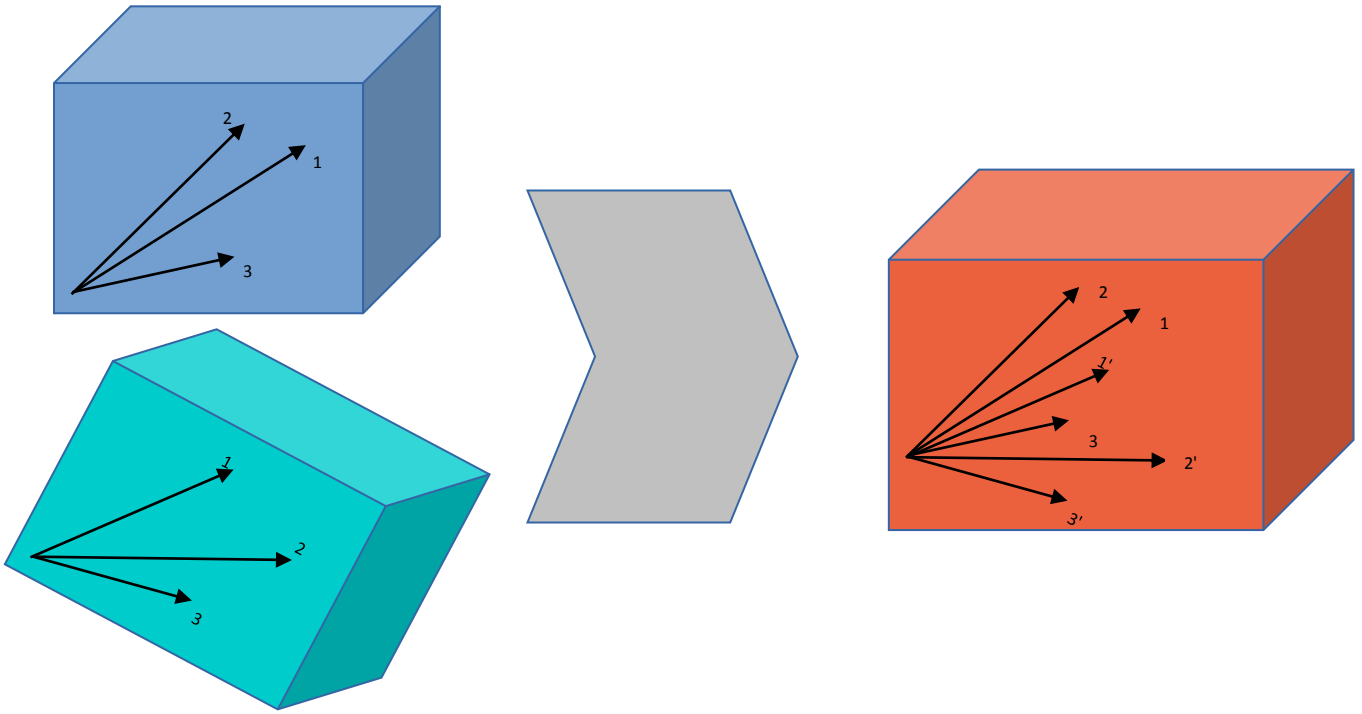
$$\{(A,B,C) \mid A = p_i, B = p_j, C = p_k, \text{where } i \neq j \neq k, \forall p \in P\}$$

Three-Tuple Relationship Changes

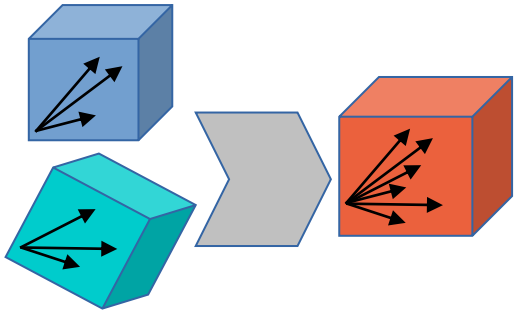




Rotations and Transform Comparisons



The Transform



$$A_1 = Project(A, S_1)$$

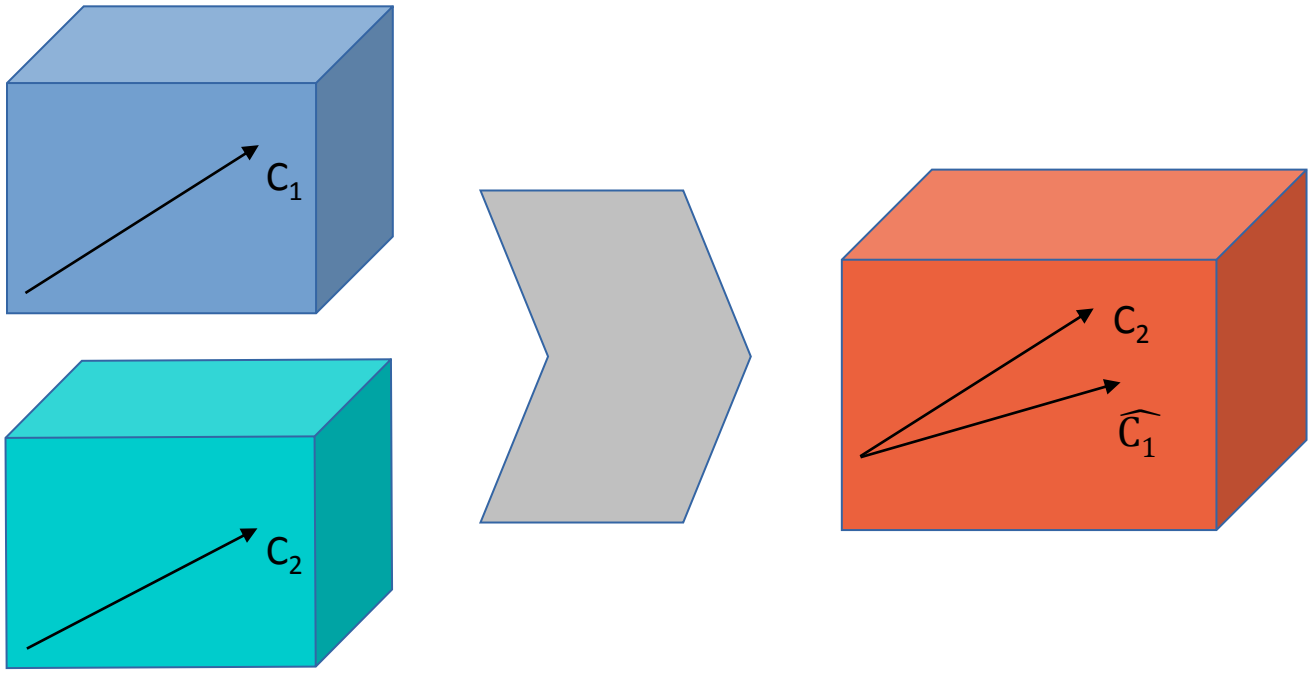
$$A_2 = Project(A, S_2)$$

$$A_1{}^T A_2 = U \Sigma V^T$$

$$Q = UV^T$$

$$\|A_1 Q - A_2\|_F$$

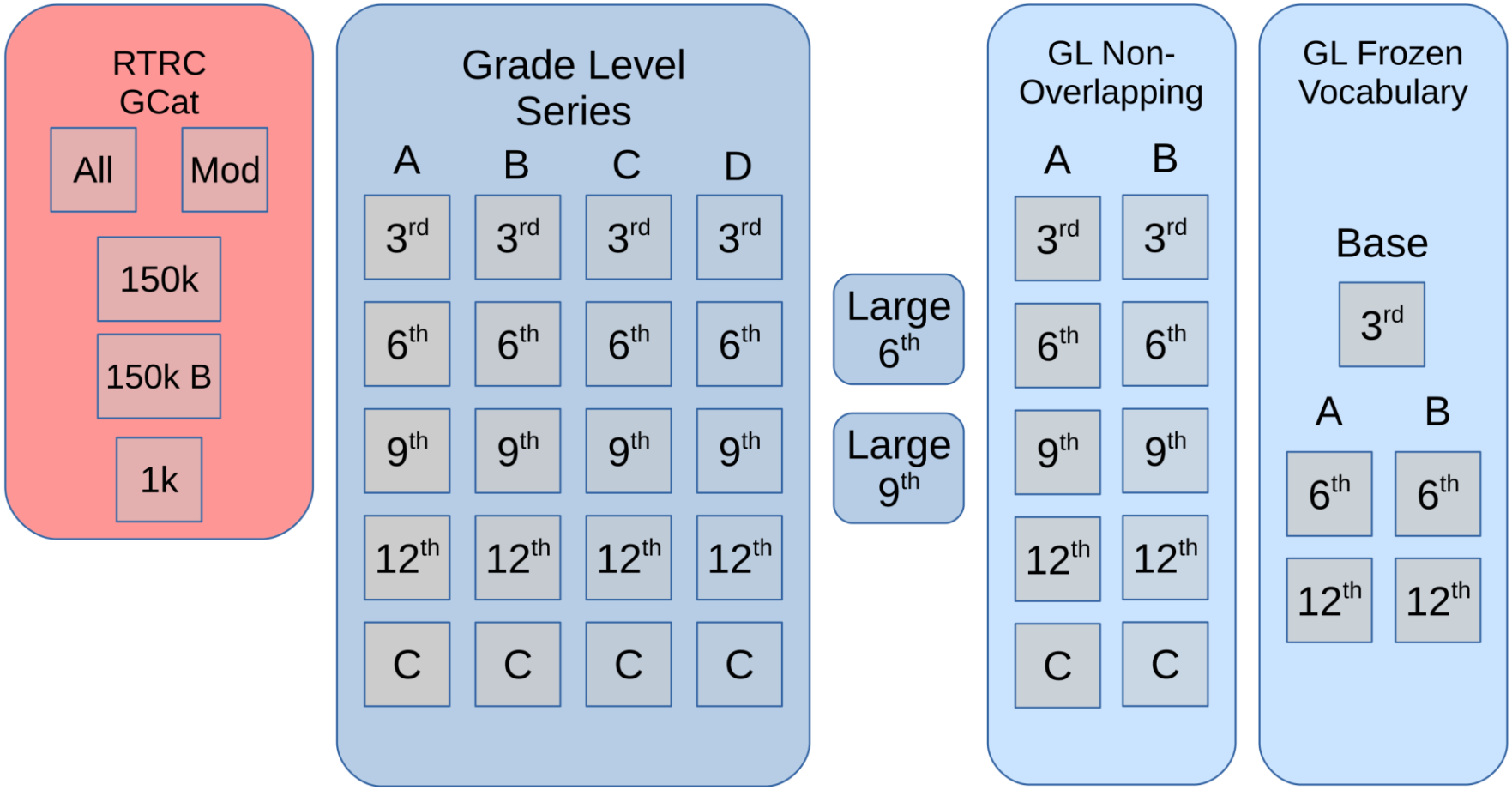
Comparative Space Centroid Analysis



Overlapping Term Vector Norm

$$\|T_1 Q - T_2\|_F = \left\|\hat{T}\right\|_F = \sqrt{\sum_{i=1}^{|\hat{T}|}\sum_{j=1}^{k}\left|\hat{t}_{i,j}\right|^2}$$

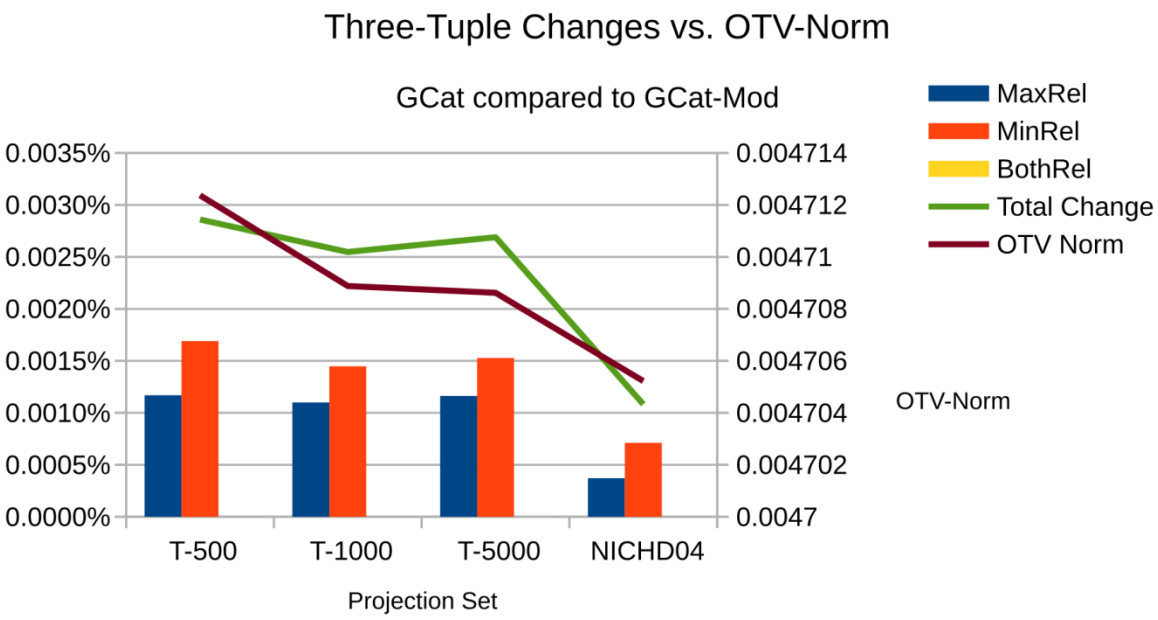
6


RTRC GCat
All
Mod
150k
150k B
1k
Grade Level Series
A
B
C
D
3rd
3rd
3rd
3rd
6th
6th
6th
6th
9th
9th
9th
9th
12th
12th
12th
12th
C
C
C
C
Large 6th
Large 9th
GL Non-Overlapping
A
B
3rd
3rd
6th
6th
9th
9th
12th
12th
C
C
GL Frozen Vocabulary
Base
3rd
A
B
6th
6th
12th
12th

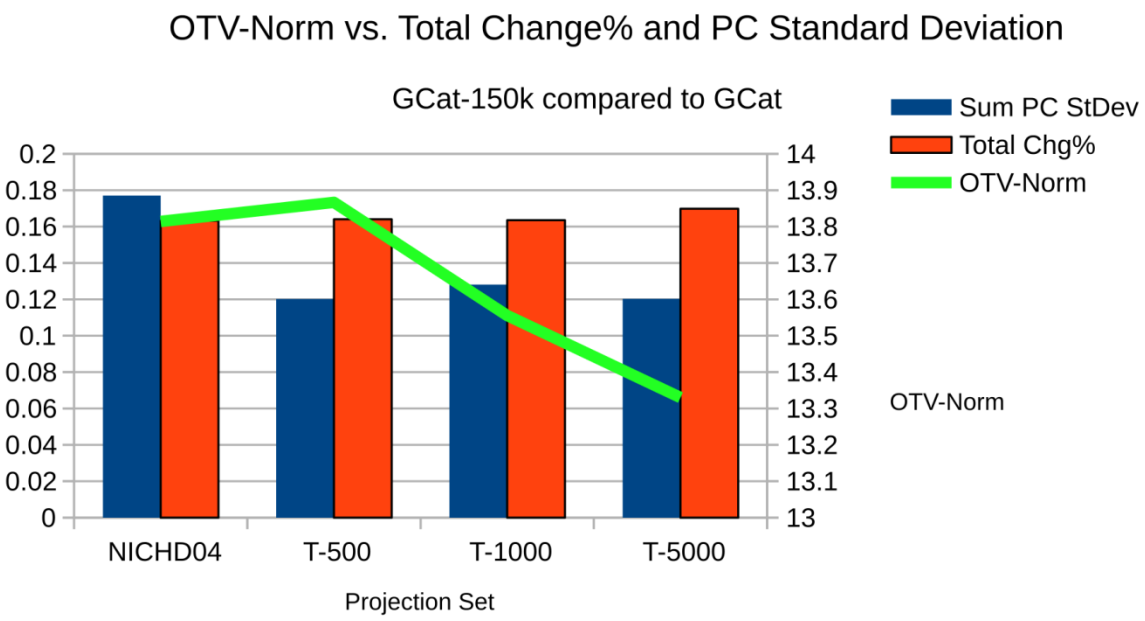
Projection/Anchor Sets

Set	Documents	Unique Terms	Term Instances
NICHD04	1,060	5,912	70,063
T-500	500	16,317	123,668
T-1000	1,000	24,319	252,372
T-5000	5,000	49,995	1,281,749

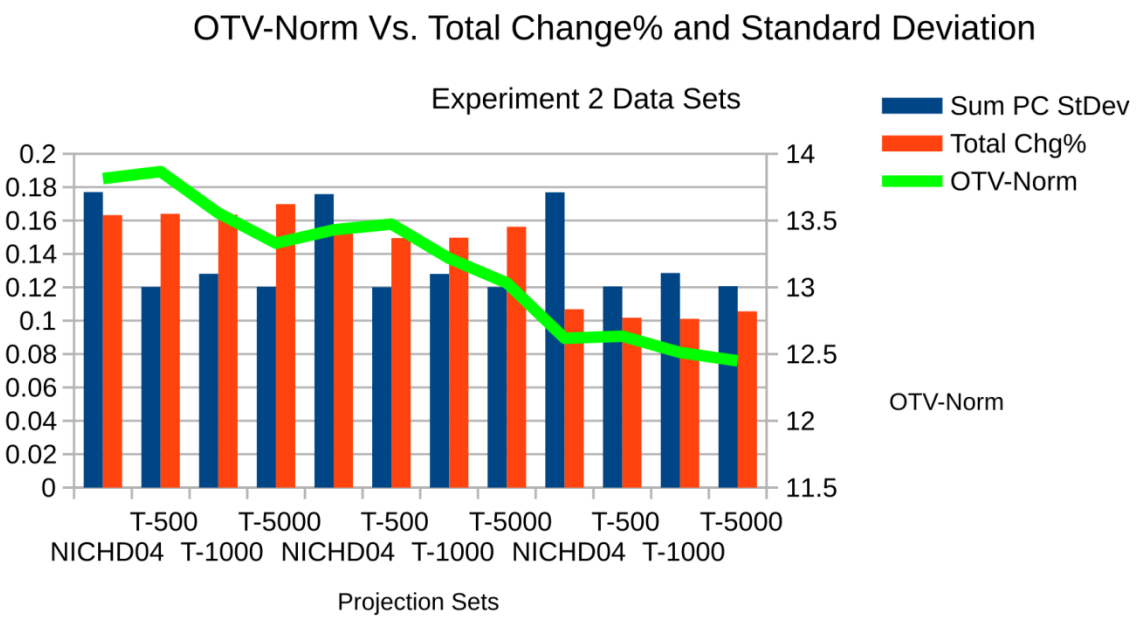
Control Experiment



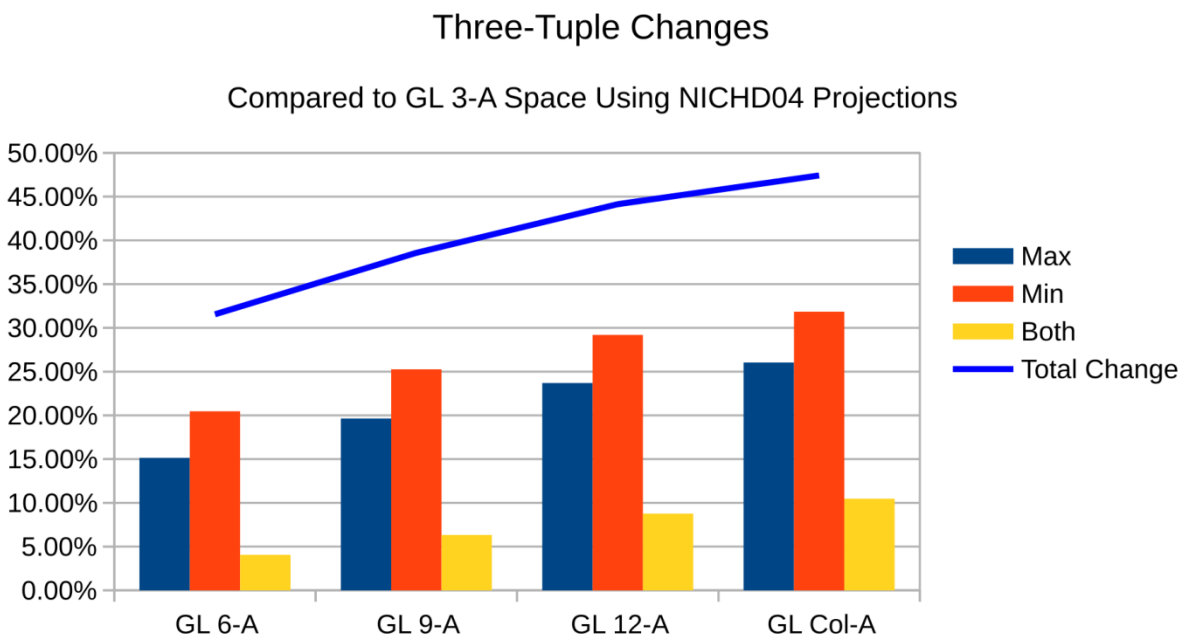
General Experiment

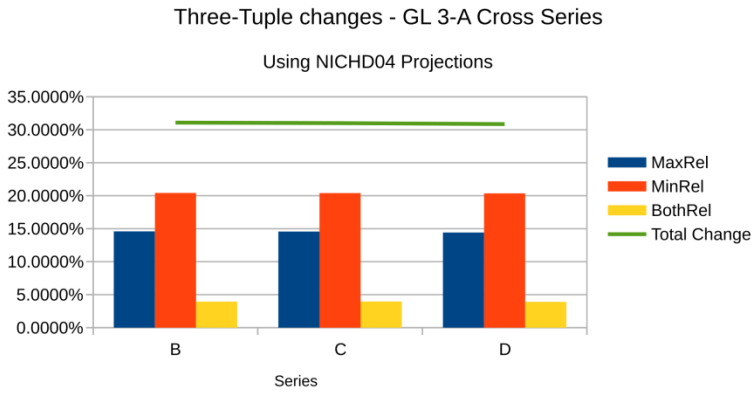


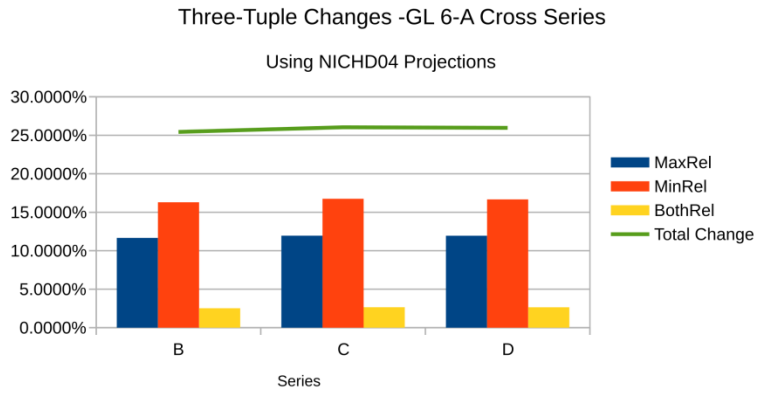
General Experiment

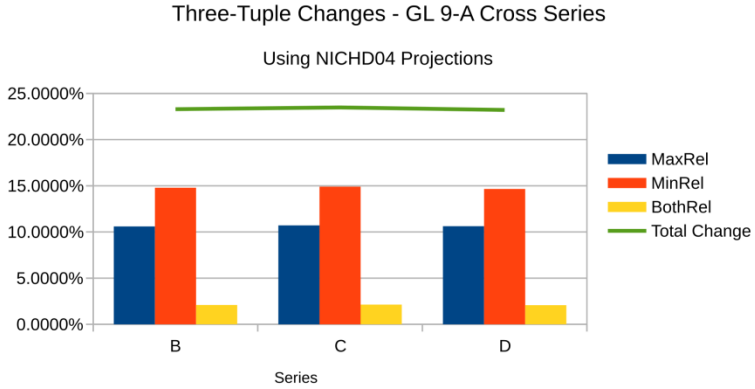


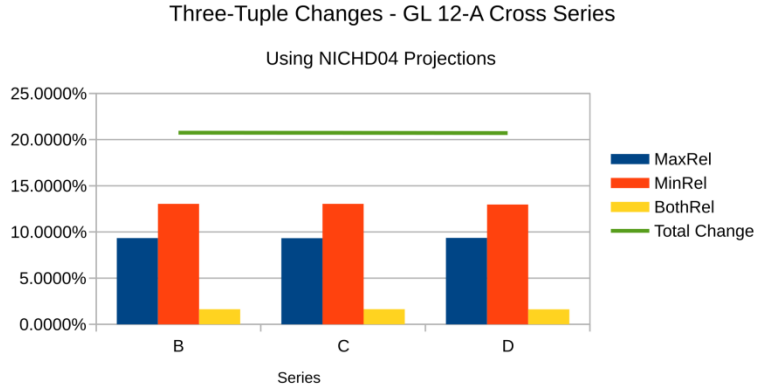
Grade Level Series Experiment

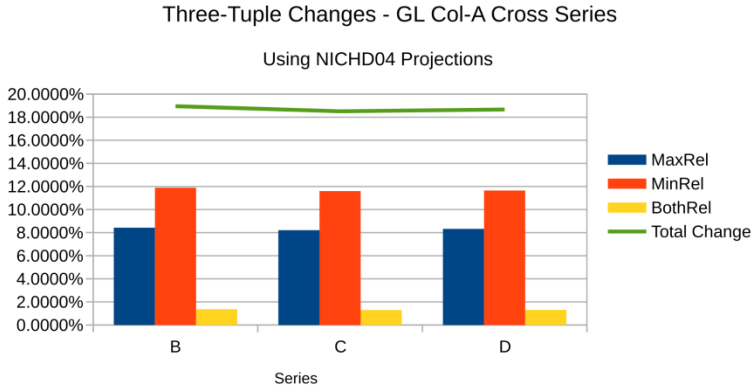



Three-Tuple changes - GL 3-A Cross Series
Using NICHD04 Projections
35.0000%
30.0000%
25.0000%
20.0000%
15.0000%
10.0000%
5.0000%
0.0000%
B
C
D
Series
MaxRel
MinRel
BothRel
Total Change

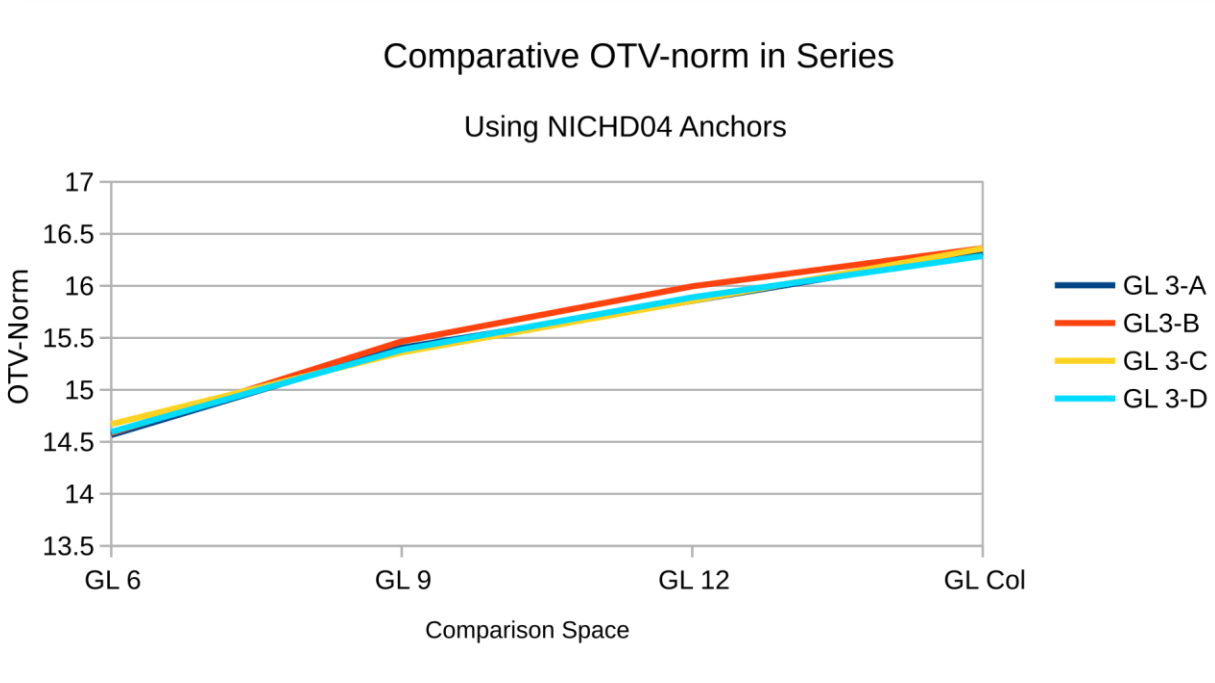

Three-Tuple Changes -GL 6-A Cross Series
Using NICHD04 Projections
30.0000%
25.0000%
20.0000%
15.0000%
10.0000%
5.0000%
0.0000%
B
C
D
Series
MaxRel
MinRel
BothRel
Total Change


Three-Tuple Changes - GL 9-A Cross Series
Using NICHD04 Projections
25.0000%
20.0000%
15.0000%
10.0000%
5.0000%
0.0000%
B
C
D
Series
MaxRel
MinRel
BothRel
Total Change

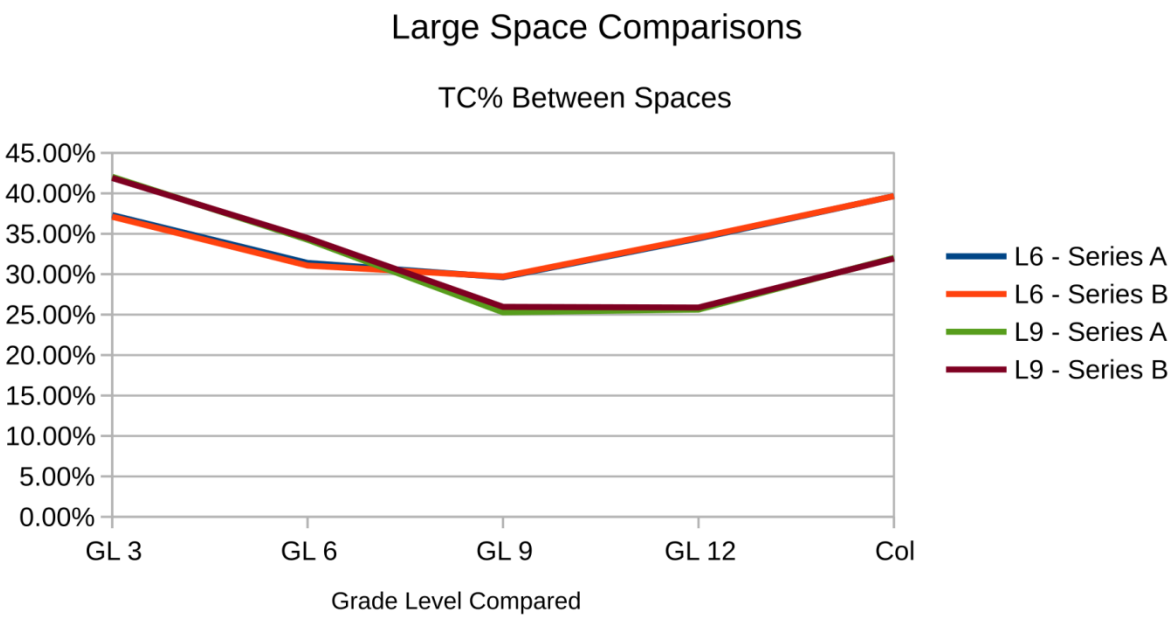

Three-Tuple Changes - GL 12-A Cross Series
Using NICHD04 Projections
25.0000%
20.0000%
15.0000%
10.0000%
5.0000%
0.0000%
B
C
D
Series
MaxRel
MinRel
BothRel
Total Change


Three-Tuple Changes - GL Col-A Cross Series
Using NICHD04 Projections
20.0000%
18.0000%
16.0000%
14.0000%
12.0000%
10.0000%
8.0000%
6.0000%
4.0000%
2.0000%
0.0000%
B
C
D
Series
MaxRel
MinRel
BothRel
Total Change

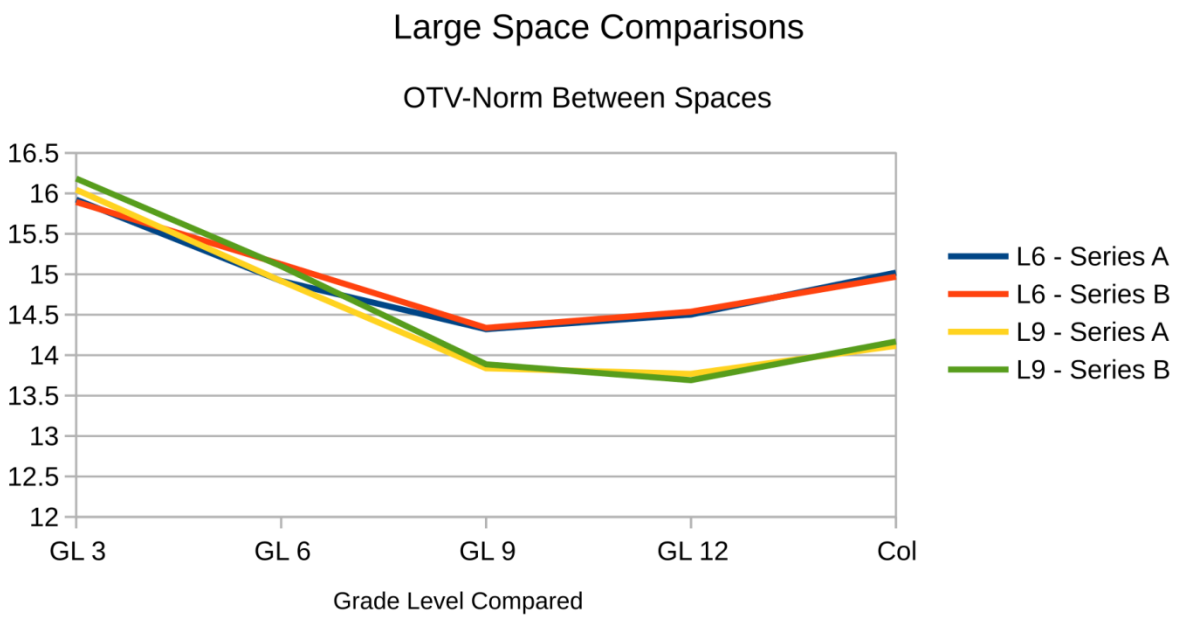
Grade Level Series OTV-Norm



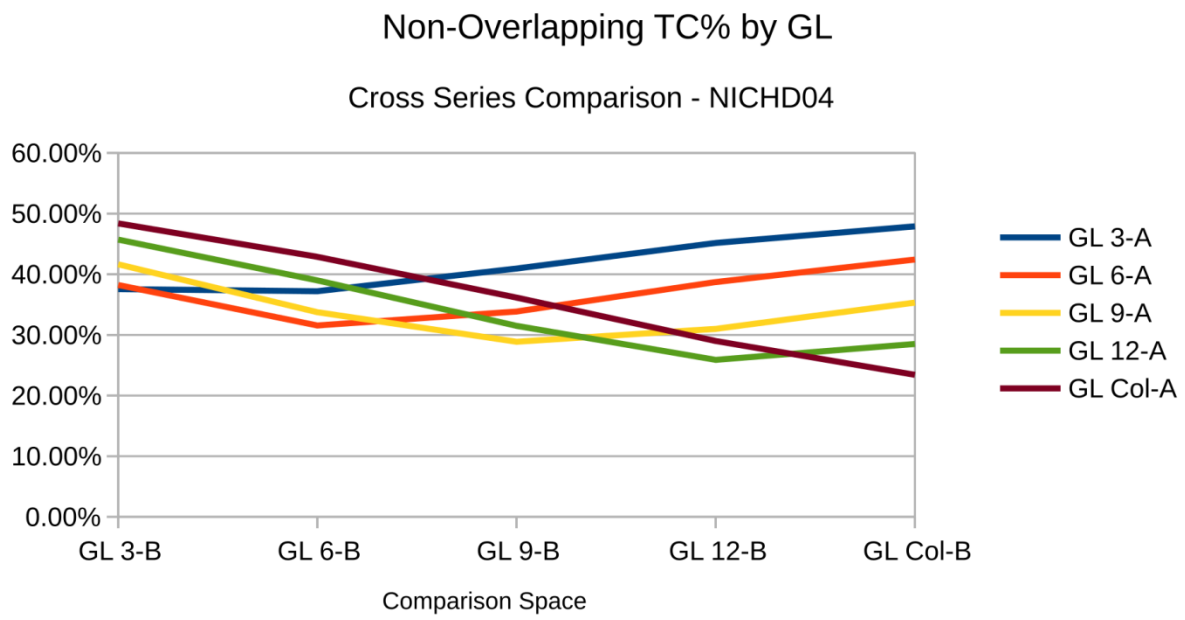
Large Volume Experiment



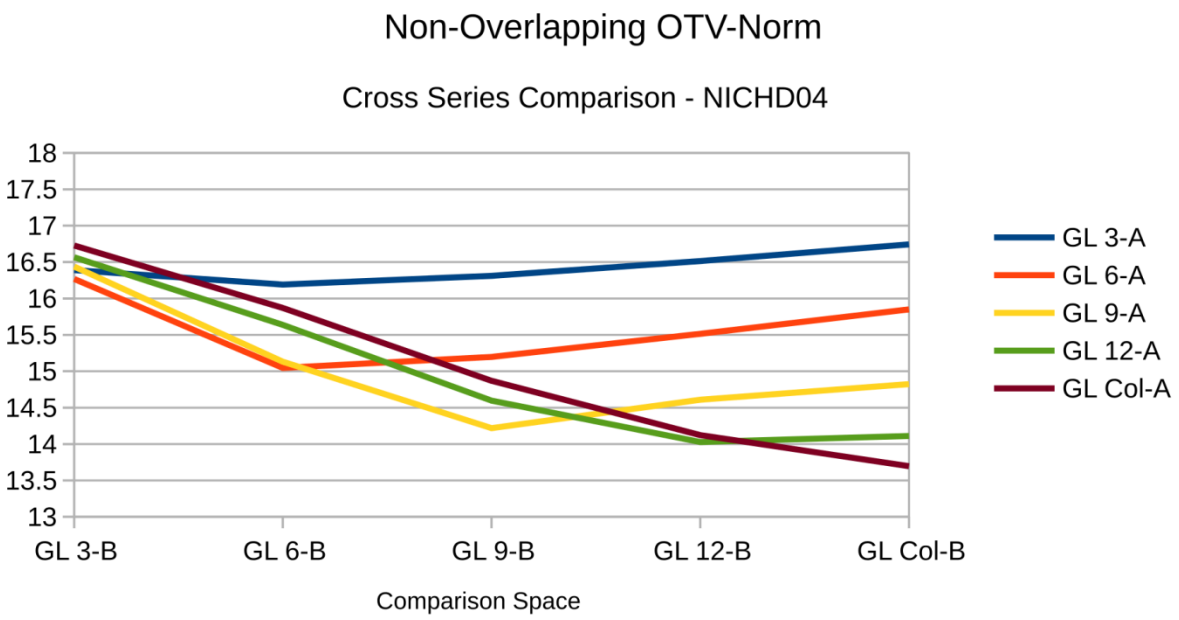
Large Volume Experiment



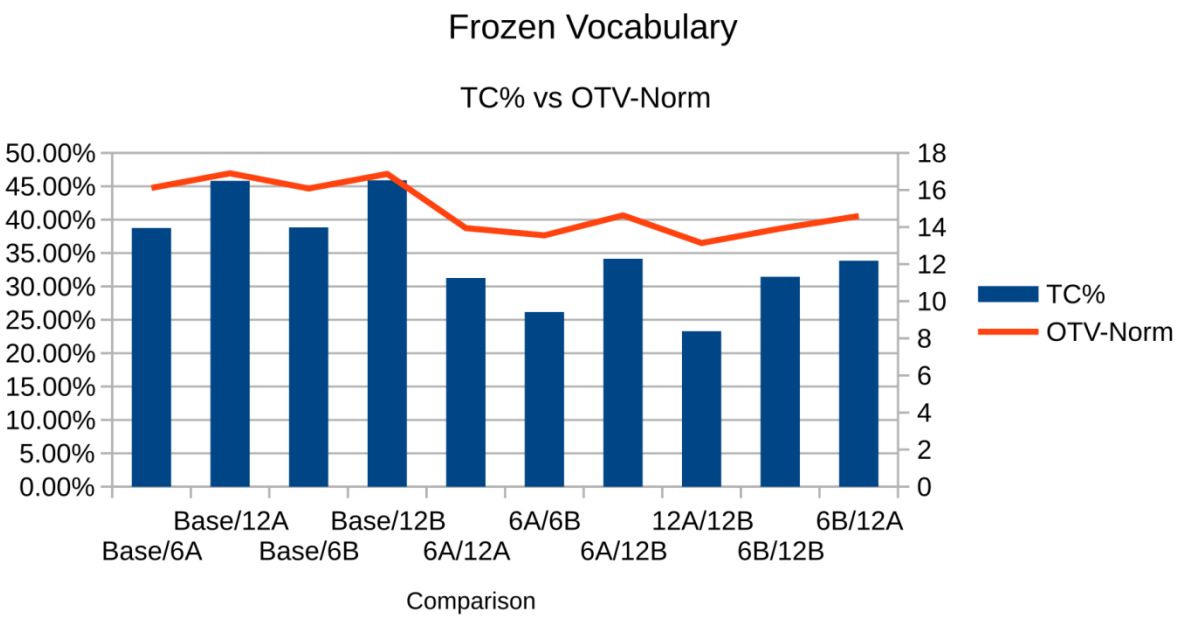
Non-overlapping Series Experiment



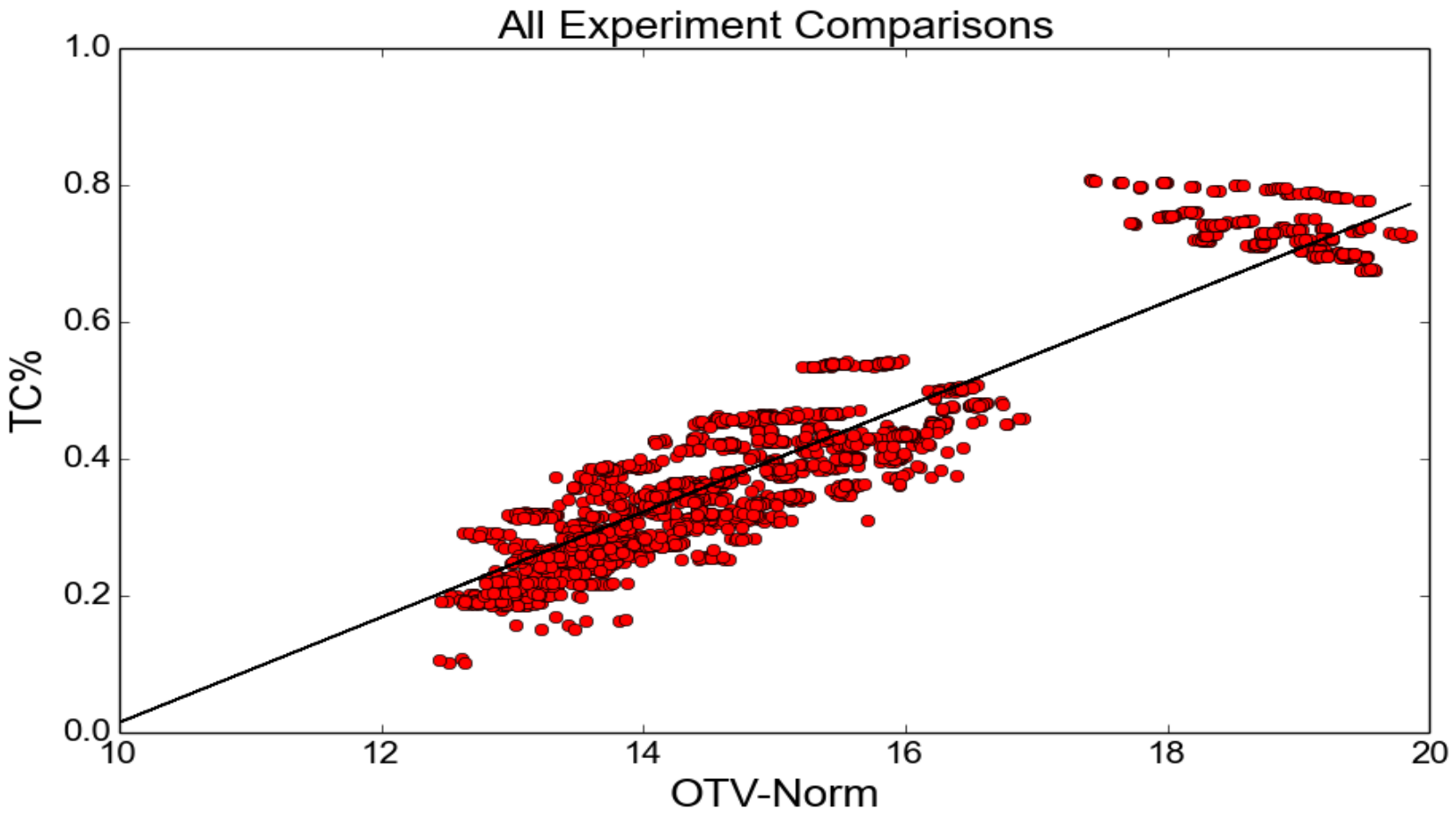
Non-Overlapping Series OTV-Norm



Frozen Vocabulary Experiment

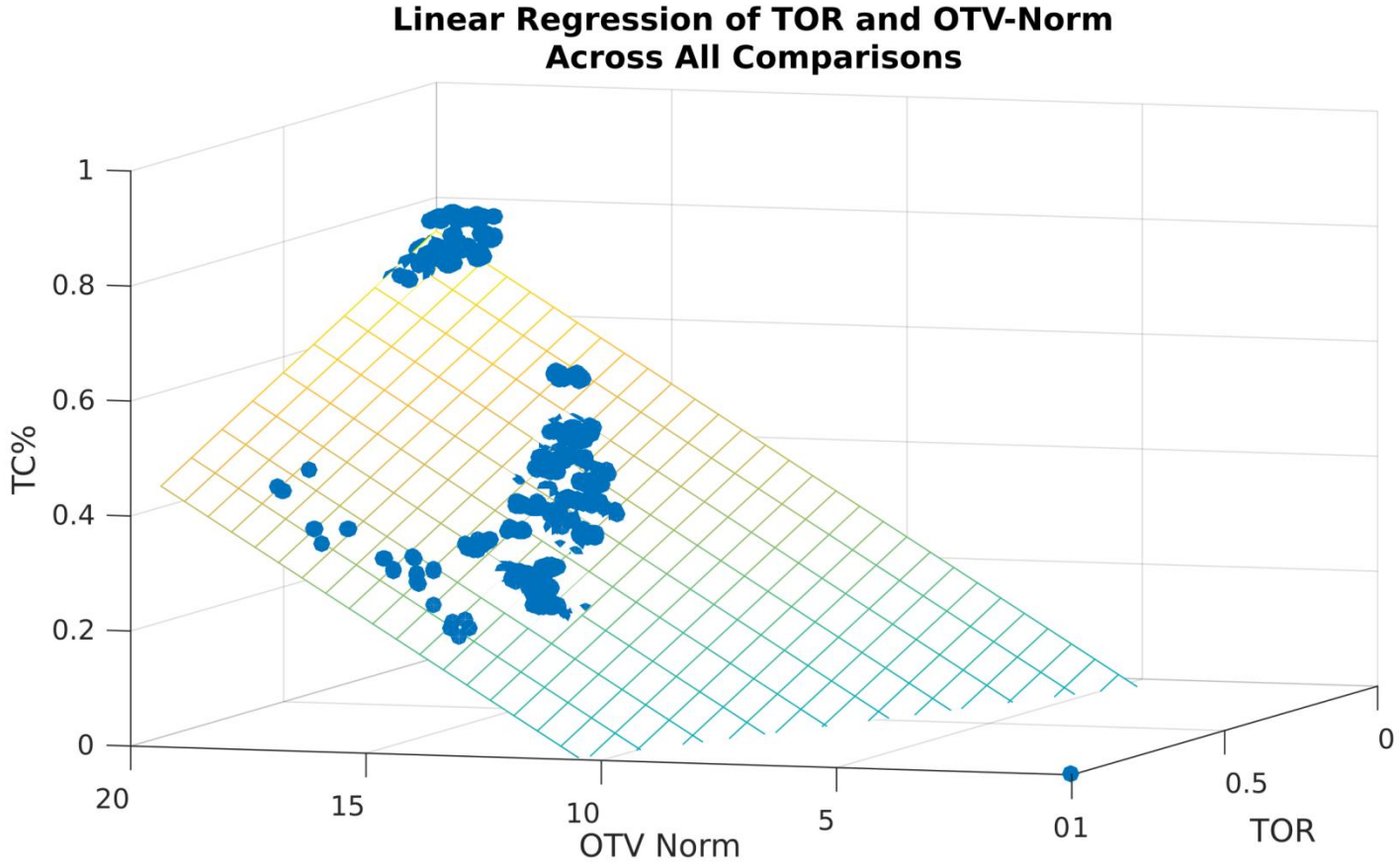


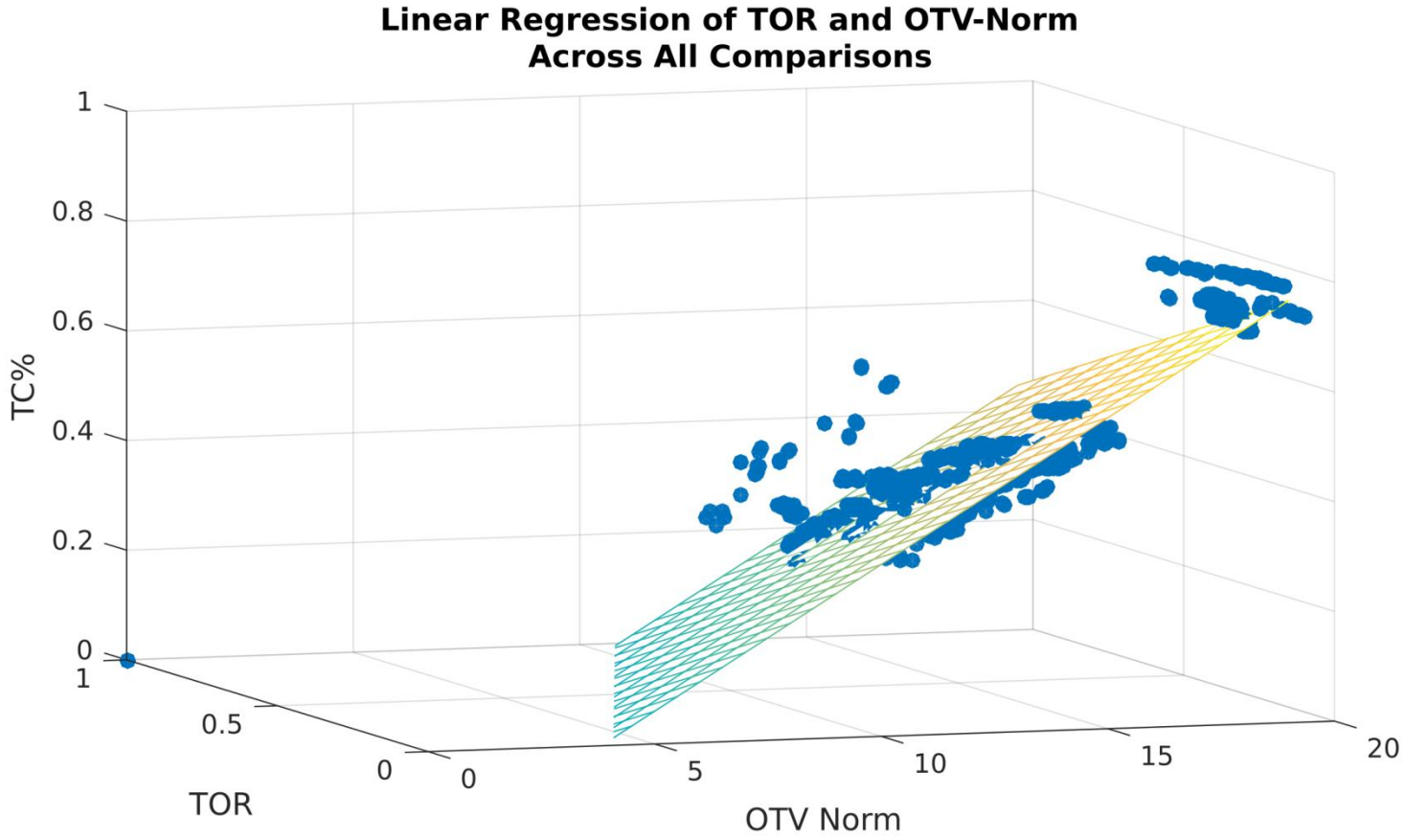
OTV-Norm

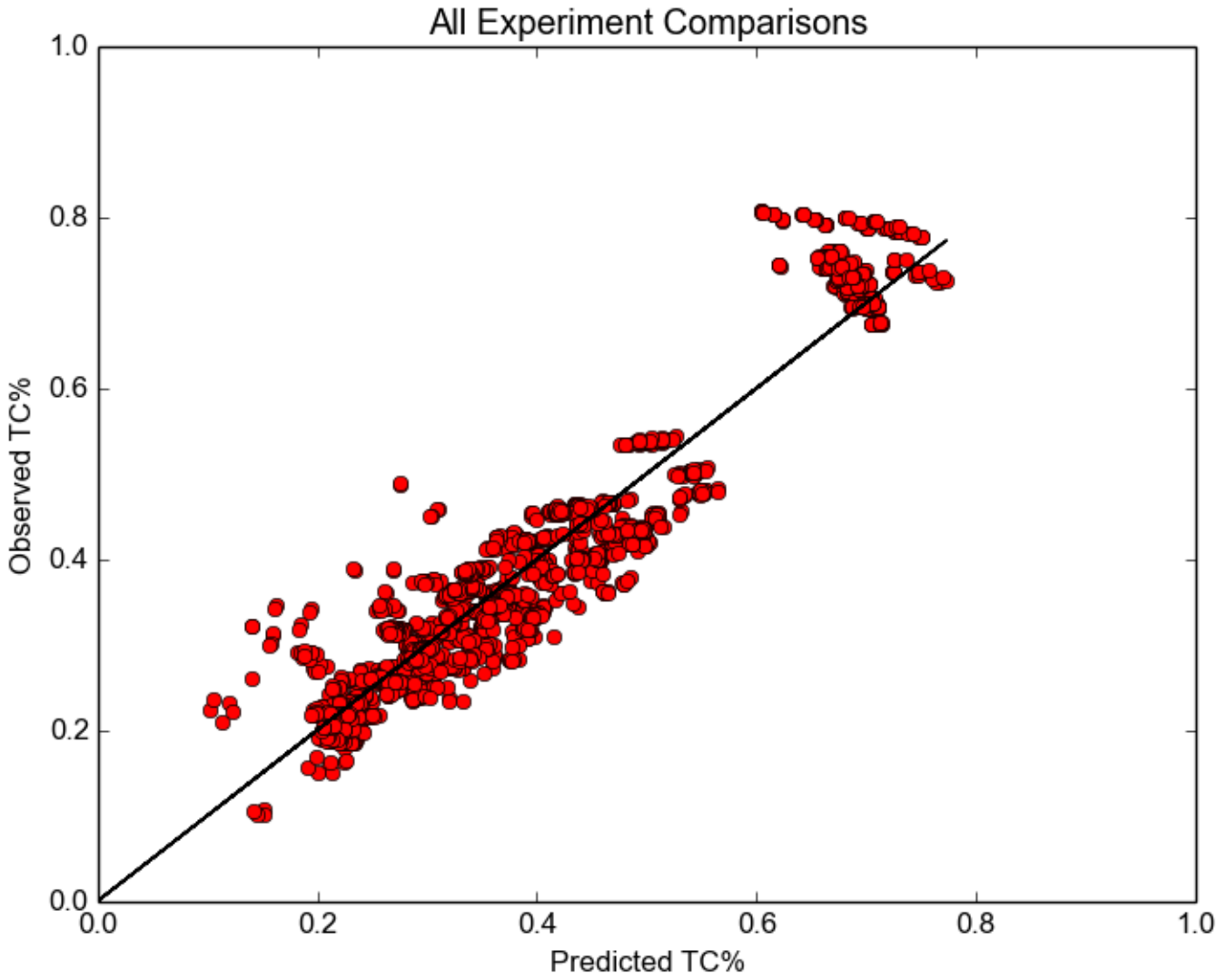

All Experiment Comparisons
TC%
OTV-Norm
1.0
0.8
0.6
0.4
0.2
0.0
10
12
14
16
18
20

Semantic Measurement Model

$$TC\% \approx -0.207882 + 0.0507194(OTVNorm) + -0.339339(TOR)$$


Linear Regression of TOR and OTV-Norm
Across All Comparisons
TC%
1
0.8
0.6
0.4
0.2
0
20
15
10
5
01
OTV Norm
0
0.5
TOR


Linear Regression of TOR and OTV-Norm
Across All Comparisons
TC%
1
0.8
0.6
0.4
0.2
0
TOR
1
0.5
0
OTV Norm
0
5
10
15
20


All Experiment Comparisons
Observed TC%
Predicted TC%
0.0
0.2
0.4
0.6
0.8
1.0

Summary of Contributions

- Semantic differences are observable
  - Measurable
  - Quality based
- Similarity not dependent on overlapping content
- OTV-Norm & Semantic Measurement Model
  - Whole-space measurement

Further Research

- Refine the model
  - Anchor set selection/influence
  - Account for non-overlapping terms
  - Investigate non-linear model
- Other questions raised

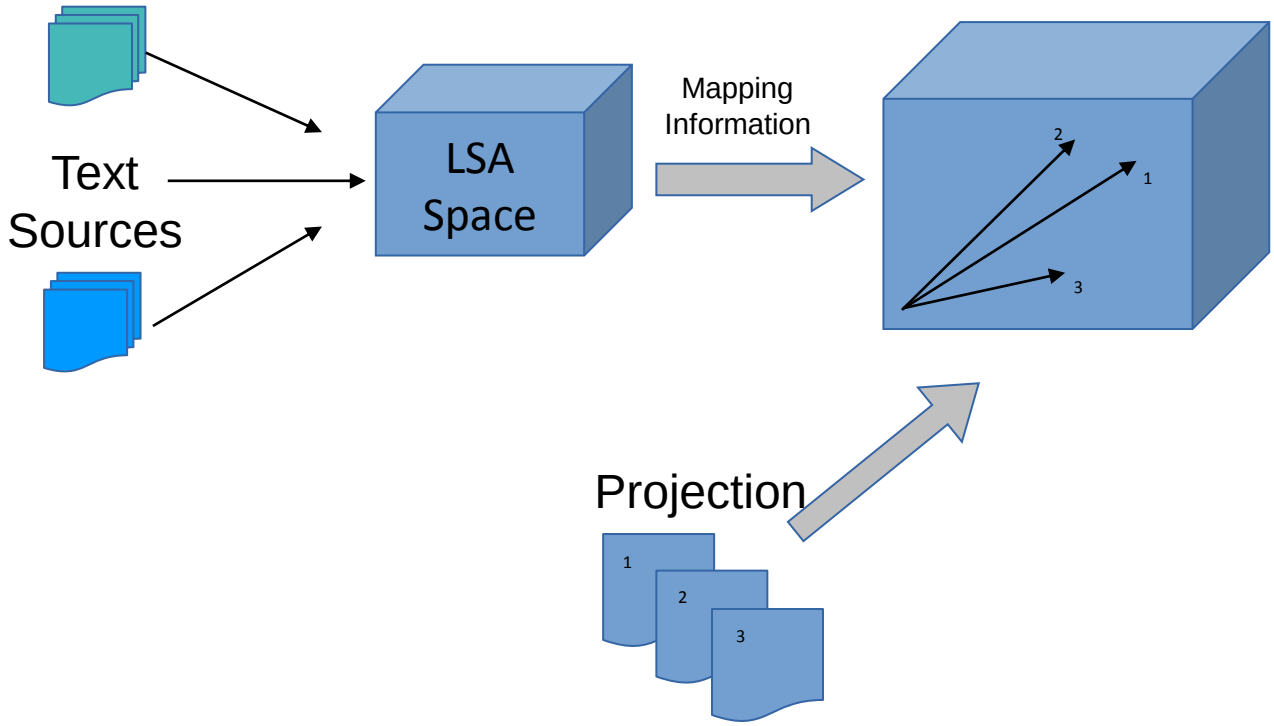
Leverage for Answering Other Questions

- Is it possible to identify key documents that affect the meaning of a space?
- Do additional items added to a space have any impact?
- Is there a point at which adding any items to a space makes no difference?
- Is it possible to identify necessary knowledge that would align two spaces?

Q&A

Backup Slides

Projection of New Content

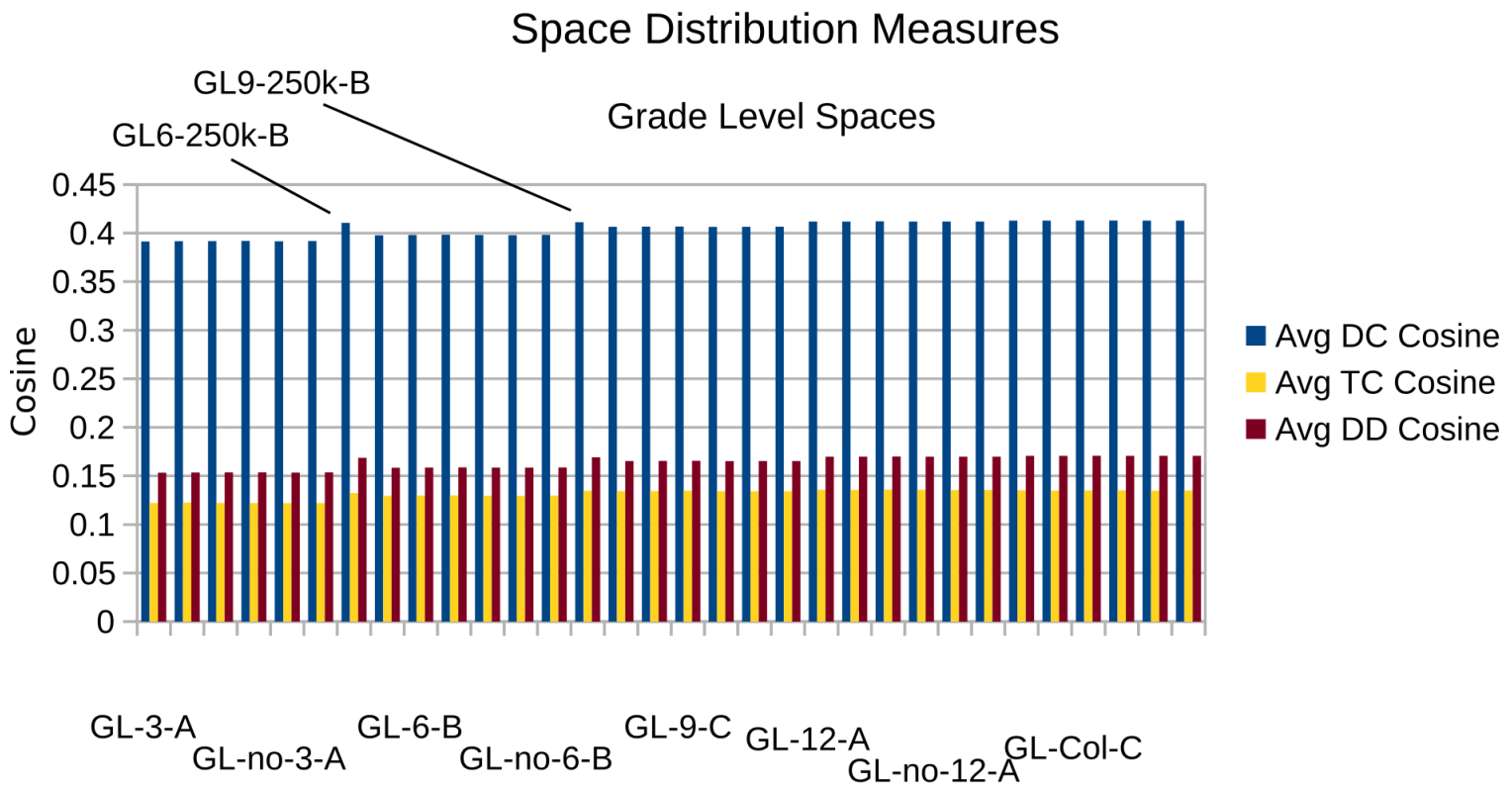


Data

- 42 Spaces
- 592 Comparisons
- 4 Projection Sets
- 4 Anchor Sets
- 26 Measures

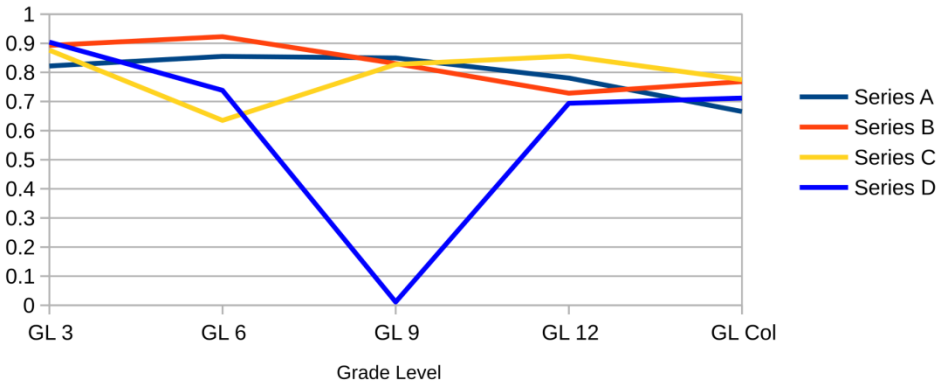
61,568 Data Items Collected

Distribution Analysis



Document Centroid Cosine Comparison

RTRC-GCat Compared to Grade Level Series Using NICHD04 Anchors



Comparative Term Centroid Cosines in Series

Using NICHD04 Anchors