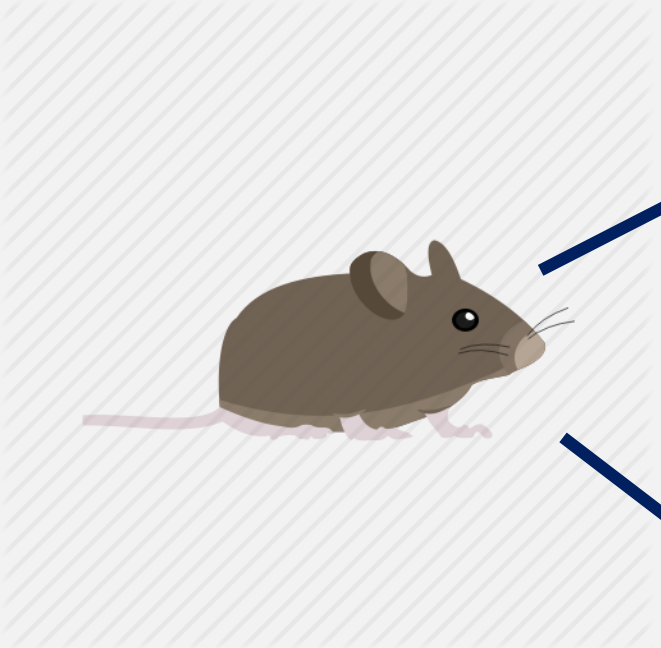
Mouse Protein Clustering

0416235 劉昱劭

Syllabus

- ✓ Dataset
- ✓ Dimensionality reduction to 2dim
  - ✓ Clustering
- ✓ Dimensionality reduction to 3dim
  - ✓ Clustering
- ✓ Conclustion

Dataset



- Control
  - Normal
- Ts65Dn
  - Down Syndrome (唐氏症基因)
  - Learning Disabilities (學習障礙)

2 Experiments

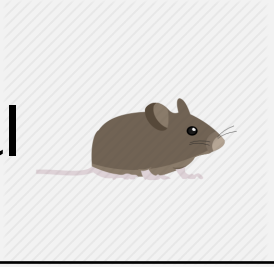
Drug (Memantine)

- To estimate whether Memantine helps Down syndrome mice.

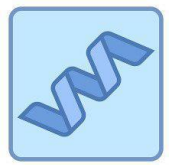
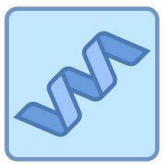
Stimulus

- Give some stimulus to make mice learn.

8 Class

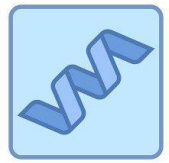
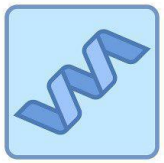
		Stimulus	
		Yes	not
Drug	Yes	normal  illed	normal  illed
	not	normal  illed	normal  illed

Dataset


MouseID	SYP_N	H3AcK18_N	EGR1_N	H3MeK4_N	CaNA_N
309_1	0.427099	0.114783	0.131790	0.128186	1.675652
294_1	0.464092	0.185664	0.183012	0.168740	1.027627
3477_1	0.400048	0.154416	0.135307	0.172007	1.399615
3422_1	0.418097	0.136876	0.175802	0.179059	0.959882
3414_1	0.400078	0.112587	0.107095	0.123739	1.829242
293_1	0.489383	0.400228	0.139862	0.284145	1.198272
18899_1	0.397663	0.155484	0.158174	0.187052	1.357802
3421_1	0.400070	0.205017	0.174371	0.204785	1.322558

Dataset


MouseID	SYP_N	H3AcK18_N	EGR1_N	H3MeK4_N	CaNA_N	Genotype	Treatment	Behavior
309_1	0.427099	0.114783	0.131790	0.128186	1.675652	Control	Memantine	C/S
294_1	0.464092	0.185664	0.183012	0.168740	1.027627	Control	Memantine	S/C
3477_1	0.400048	0.154416	0.135307	0.172007	1.399615	Control	Saline	C/S
3422_1	0.418097	0.136876	0.175802	0.179059	0.959882	Control	Saline	S/C
3414_1	0.400078	0.112587	0.107095	0.123739	1.829242	Ts65Dn	Memantine	C/S
293_1	0.489383	0.400228	0.139862	0.284145	1.198272	Ts65Dn	Memantine	S/C
18899_1	0.397663	0.155484	0.158174	0.187052	1.357802	Ts65Dn	Saline	C/S
3421_1	0.400070	0.205017	0.174371	0.204785	1.322558	Ts65Dn	Saline	S/C

Dataset

encode

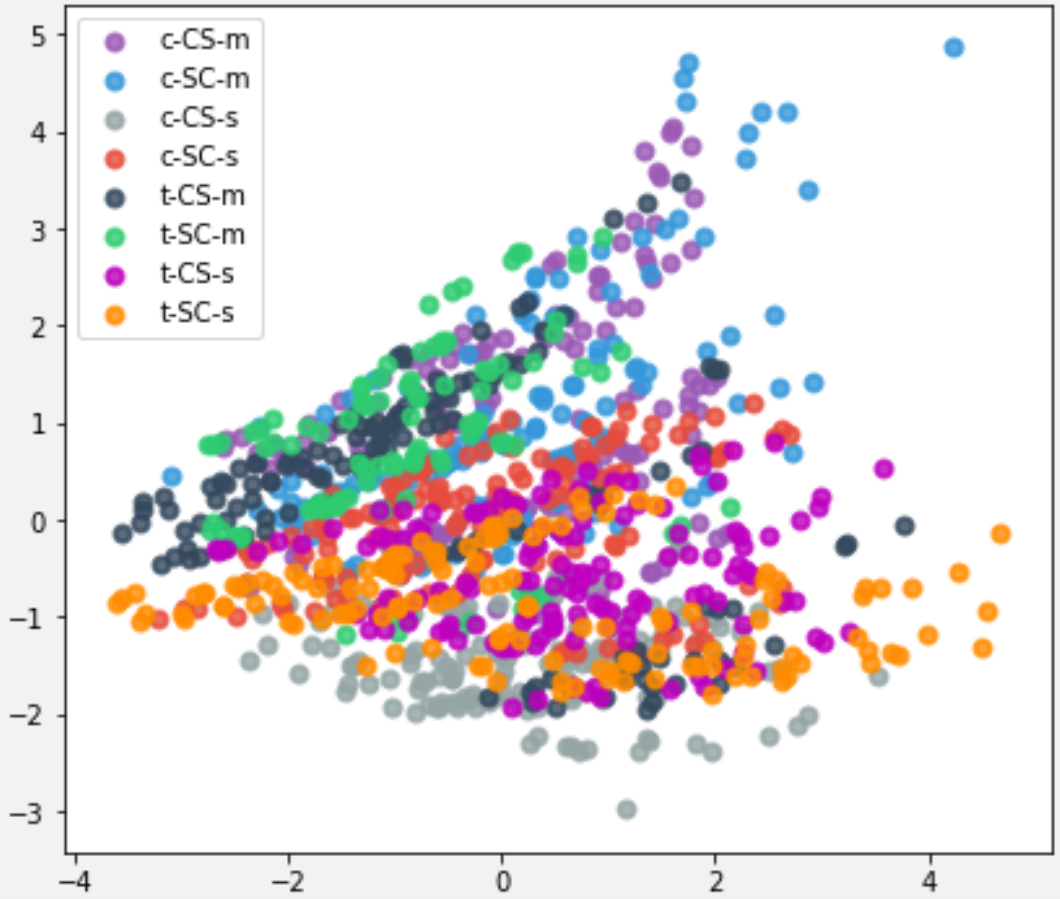
MouseID	SYP_N	H3AcK18_N	EGR1_N	H3MeK4_N	CaNA_N	Genotype	Treatment	Behavior	class	target
309_1	0.427099	0.114783	0.131790	0.128186	1.675652	Control	Memantine	C/S	c-CS-m	0
294_1	0.464092	0.185664	0.183012	0.168740	1.027627	Control	Memantine	S/C	c-SC-m	2
3477_1	0.400048	0.154416	0.135307	0.172007	1.399615	Control	Saline	C/S	c-CS-s	1
3422_1	0.418097	0.136876	0.175802	0.179059	0.959882	Control	Saline	S/C	c-SC-s	3
3414_1	0.400078	0.112587	0.107095	0.123739	1.829242	Ts65Dn	Memantine	C/S	t-CS-m	4
293_1	0.489383	0.400228	0.139862	0.284145	1.198272	Ts65Dn	Memantine	S/C	t-SC-m	6
18899_1	0.397663	0.155484	0.158174	0.187052	1.357802	Ts65Dn	Saline	C/S	t-CS-s	5
3421_1	0.400070	0.205017	0.174371	0.204785	1.322558	Ts65Dn	Saline	S/C	t-SC-s	7

Dimensionality Reduction

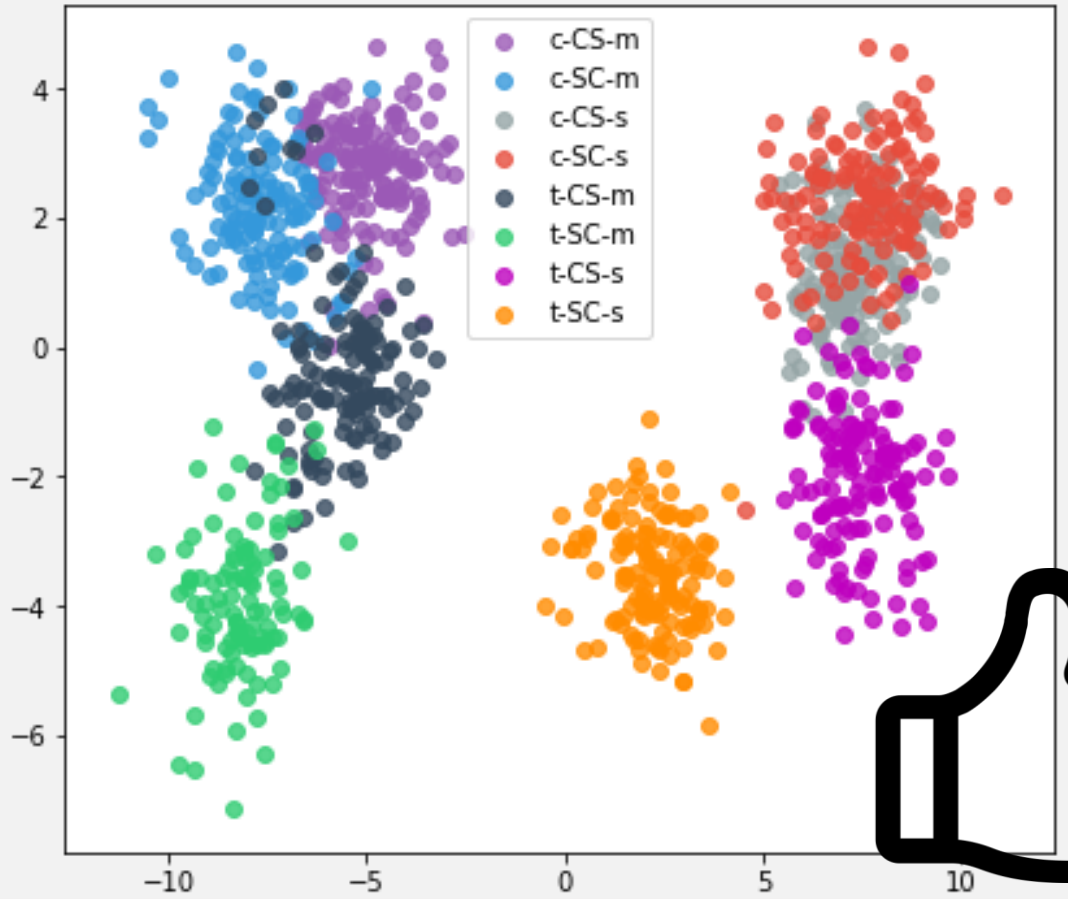
- ✓ For visualization
- ✓ Speed up clustering

Dimensionality Reduction to 2-dim

PCA



LDA

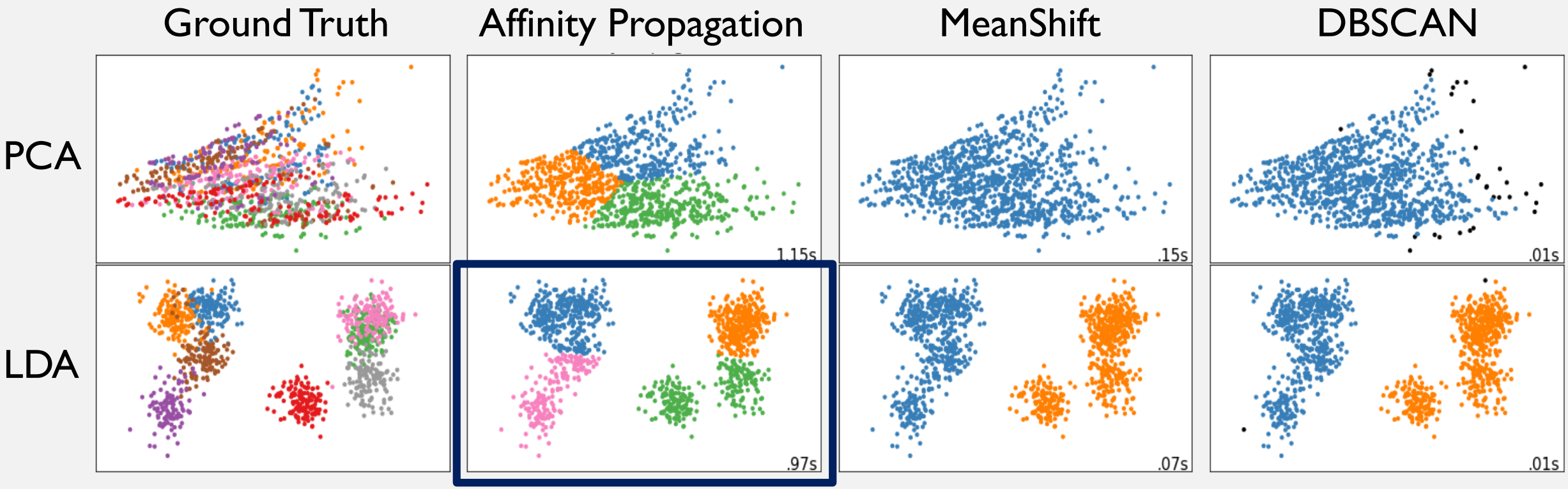


2-dim clustering without k

- V-measure
  - [0, 1]
  - Larger is better

		AffinityPropagation	MeanShift	DBSCAN
77dim	Origin	0.539819	2.14059e-16	2.14059e-16
2 dim	PCA	0.146638	2.14059e-16	0.0123773
2 dim	LDA	0.706204	0.500664	0.499559

2-dim clustering without k



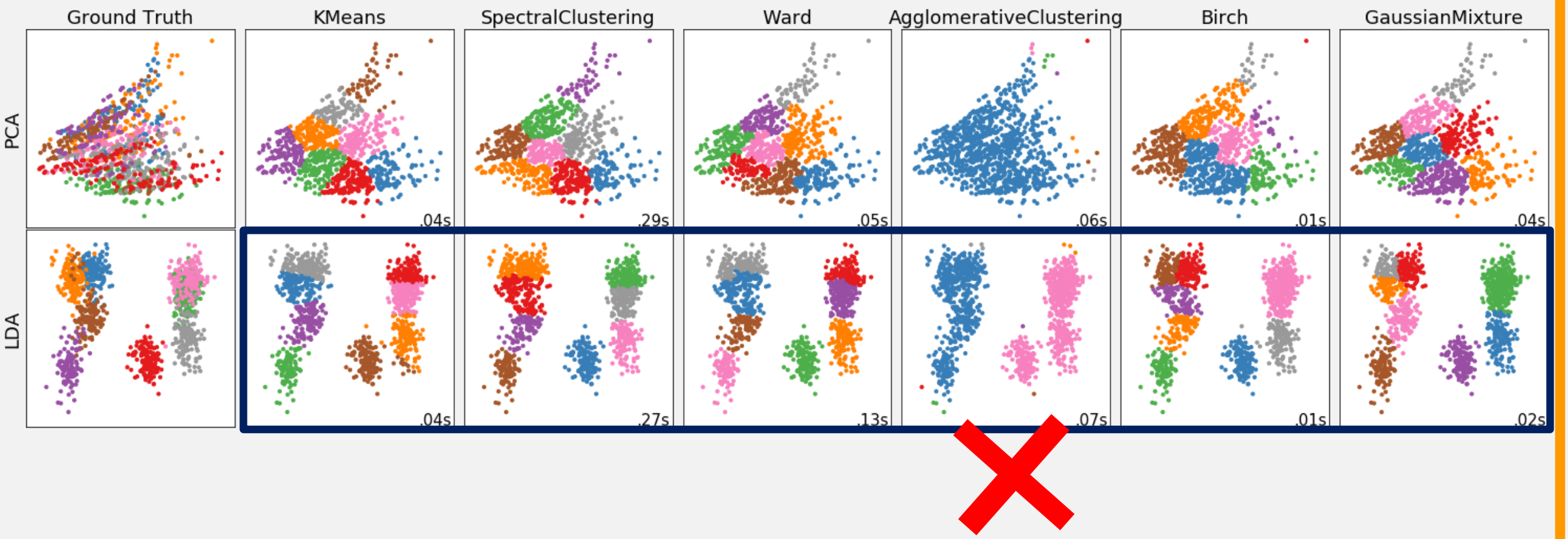
- Good
- But only get 4 classes

2-dim clustering with k=8

- V-measure
  - [0, 1]
  - Larger is better

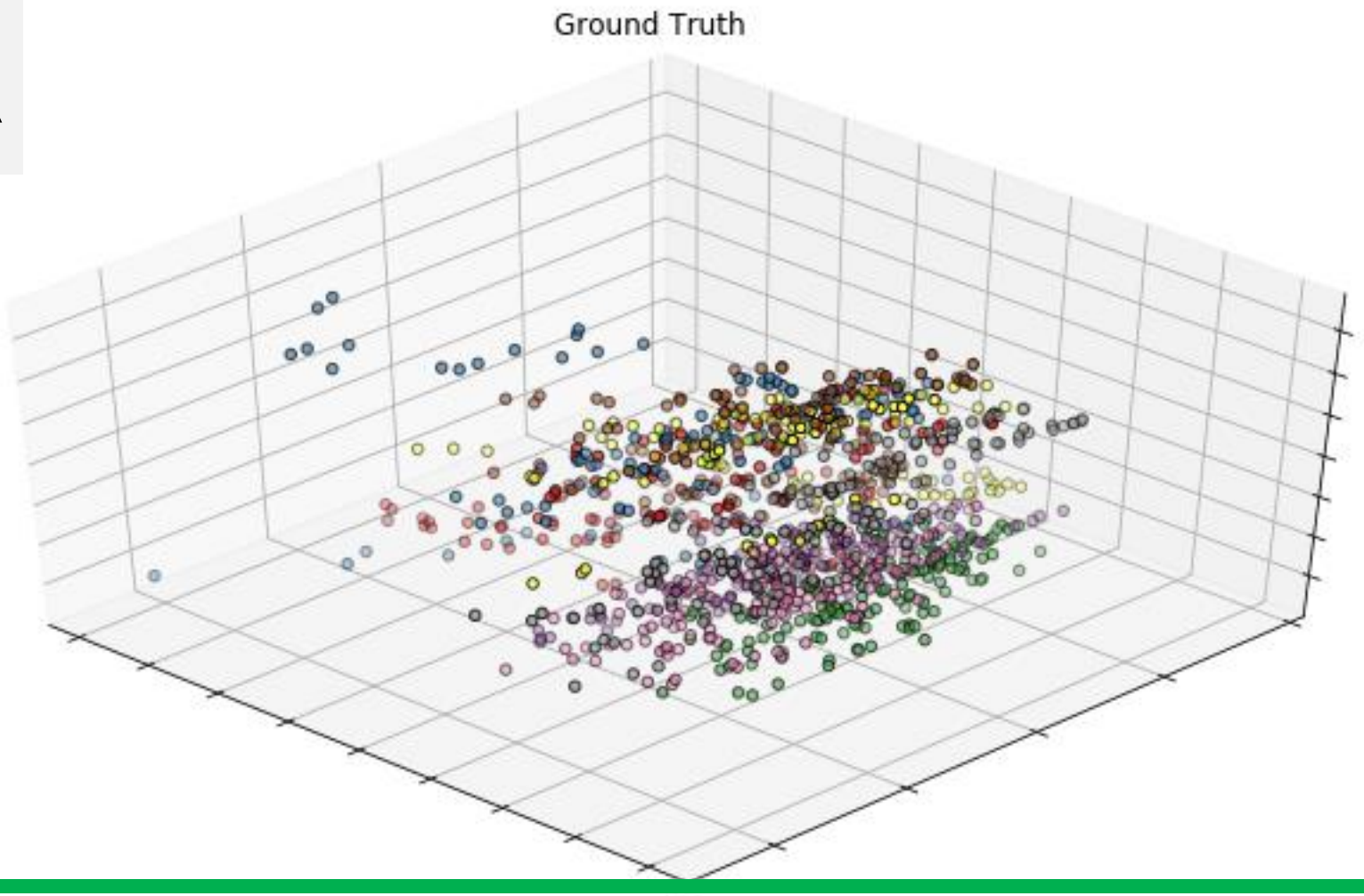
		KMeans	Spectral	Ward	Agglo_Avglink	Birch	GaussianMixture
77d	Origin	0.2643	0.451411	0.322522	0.0300846	0.264992	0.262896
2 d	PCA	0.205813	0.225799	0.217847	0.0256257	0.169129	0.199601
2 d	LDA	0.744963	0.739274	0.725042	0.49933	0.781119	0.765893

2-dim clustering with k=8



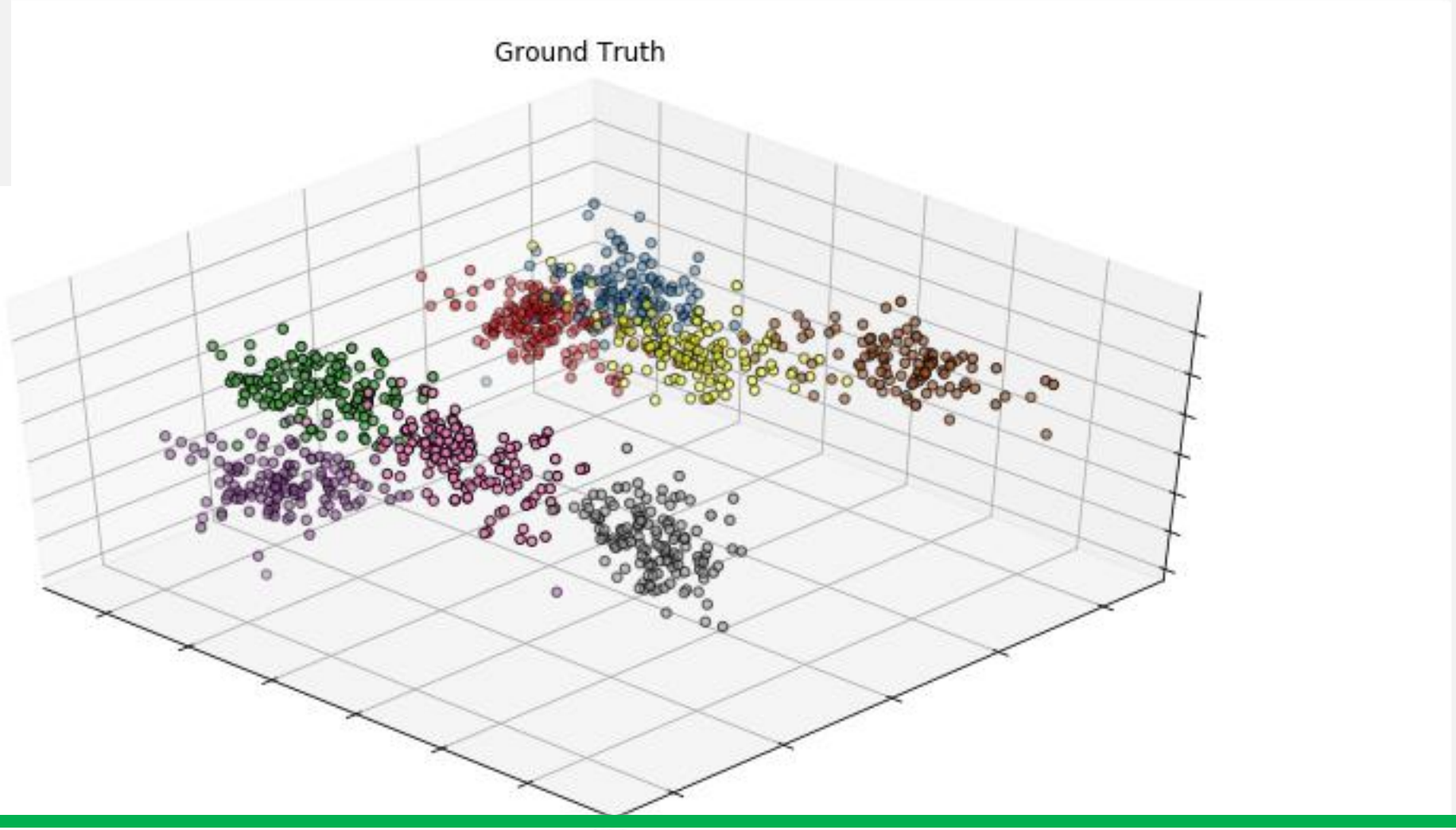
Dimensionality Reduction to 3-dim

PCA



Dimensionality Reduction to 3-dim

LDA



3-dim clustering without k

- V-measure

		AffinityPropagation	MeanShift	DBSCAN
77dim	Origin	0.539819	2.14059e-16	2.14059e-16
2 dim	PCA	0.146638	2.14059e-16	0.0123773
2 dim	LDA	0.706204	0.500664	0.499559
	3 dim PCA	0.162405	0.152277	0.284773
	3 dim LDA	0.891241	0.760094	0.027584

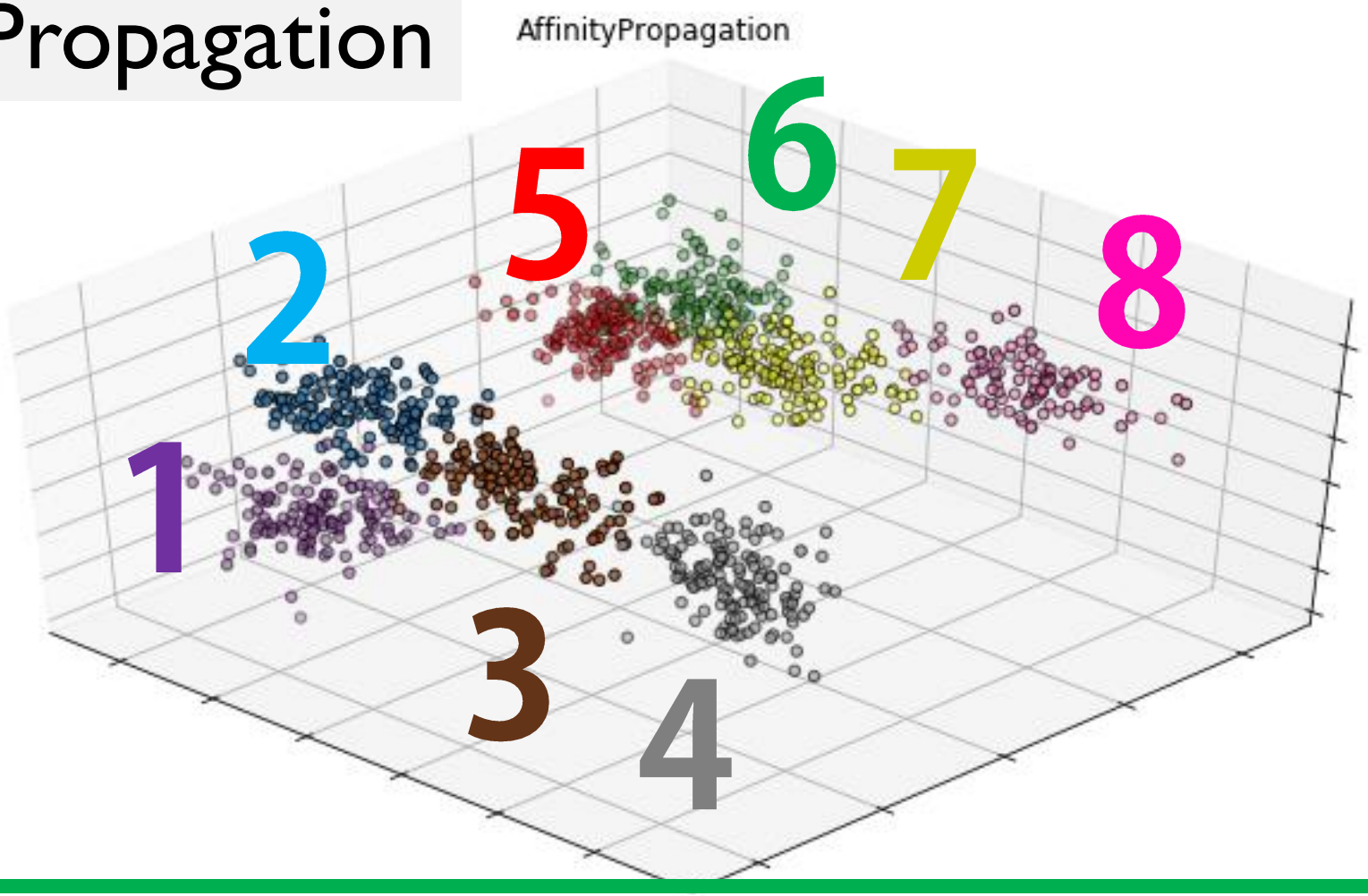
3-dim clustering without k

- V-measure

		AffinityPropagation	MeanShift	DBSCAN
77dim	Origin	0.539819	2.14059e-16	2.14059e-16
2 dim	PCA	0.146638	2.14059e-16	0.0123773
2 dim	LDA	0.706204	0.500664	0.499559
	3 dim PCA	0.162405	0.152277	0.284773
	3 dim LDA	0.891241	0.760094	0.027584

3-dim clustering without k

Affinity Propagation



3-dim clustering with k=8

- V-measure

		KMeans	Spectral	Ward	Agglo_Avglink	Birch
77dim	Origin	0.2643	0.451411	0.322522	0.0300846	0.264992
2 dim	PCA	0.205813	0.225799	0.217847	0.0256257	0.169129
2 dim	LDA	0.744963	0.739274	0.725042	0.49933	0.781119
	3dim PCA	0.255006	0.327668	0.560827	0.021417	0.197289
	3dim LDA	0.893534	0.884505	0.831242	0.62531	0.854395

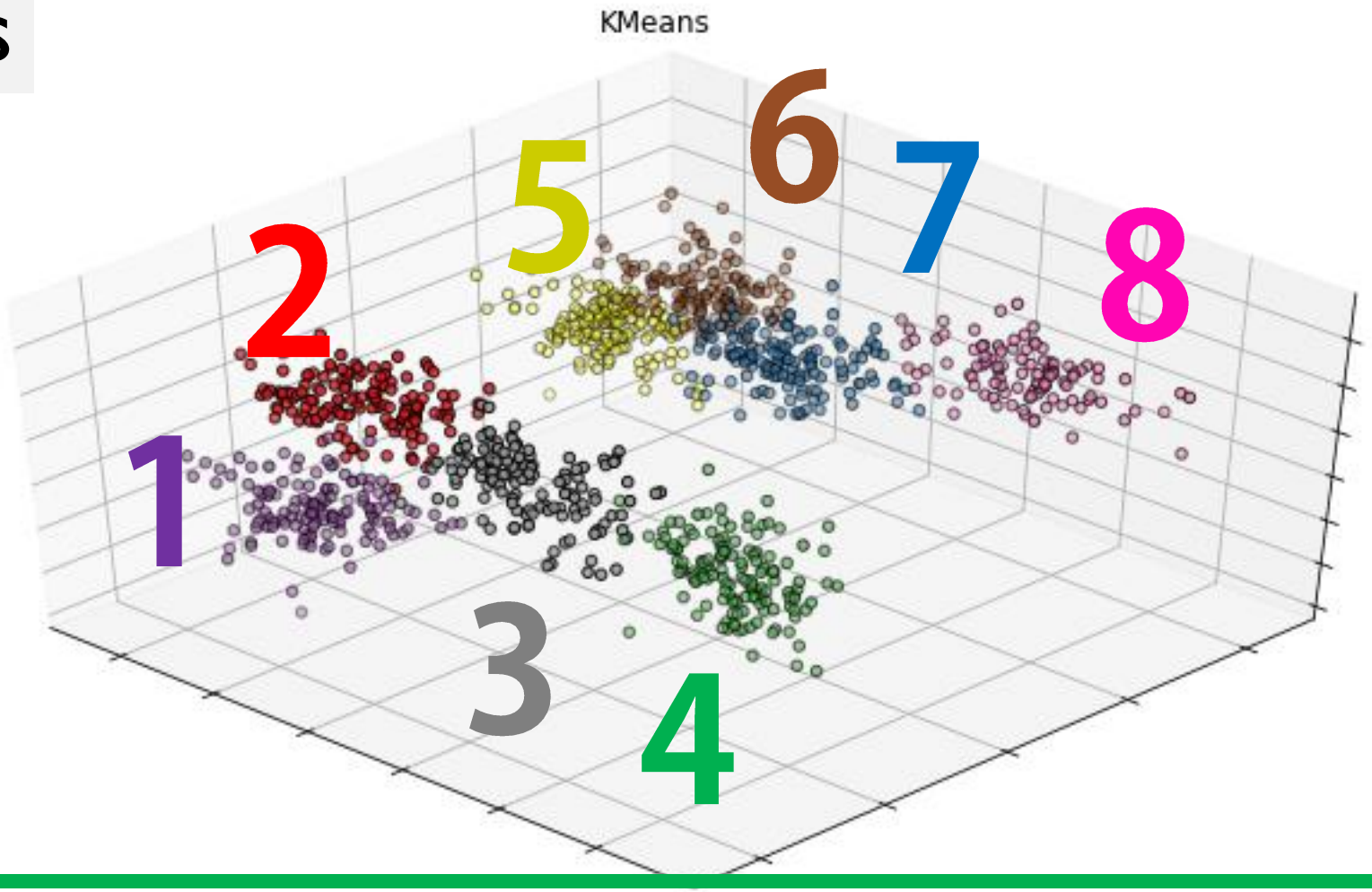
3-dim clustering with k=8

- V-measure

		KMeans	Spectral	Ward	Agglo_Avglink	Birch
77dim	Origin	0.2643	0.451411	0.322522	0.0300846	0.264992
2 dim	PCA	0.205813	0.225799	0.217847	0.0256257	0.169129
2 dim	LDA	0.744963	0.739274	0.725042	0.49933	0.781119
	3dim PCA	0.255006	0.327668	0.560827	0.021417	0.197289
	3dim LDA	0.893534	0.884505	0.831242	0.62531	0.854395

3-dim clustering with k=8

K-means



Conclusion

✓ Reduce dimension to different dimensions may get better clustering result.

✓ For touching data, clustering methods without predefined k may get few classes.

Reference Paper

✓ V-measure (For clustering accuracy)

http://www.aclweb.org/anthology/D07-1043

✓ Dataset from kaggle

https://www.kaggle.com/ruslankl/mice-protein-expression

✓ Mice Protein (提供此資料集的生物領域論文)

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0119491

✓ Mice Protein Clustering (以此資料集做分群)

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0129126

Q & A