Big Five Personality Test

CLUSTERING PROJECT

Uria Levy, Elie Cahen, Shlomi Cohen

Big Five Personality Test (or OCEAN model):

Openness (OPN)	-	**Openness to experience, conservativism\liberalism, adventurous, in/conventional thinking**
Conscientiousness (CSN)	-	**Self discipline, responsibility, self organization, carelessness**
Extroversion (EXT)	-	**How Extroverted or Introverted an individual would be?**
Agreeableness (AGR)	-	**Sociality, care\carelessness for others, aggression and selfishness**
Neuroticism (EST)	-	**Mood stability, indifference, calmness, self-confidence**

DATA

- Retrieved from Kaggle
- +1,000,000 cases
- 109 columns:
  - 50 questions columns
  - Answer times columns
  - Date
  - No. of entries from the same IP (IPC)
  - Screen data
  - Location data

```
df.tail(5)
```

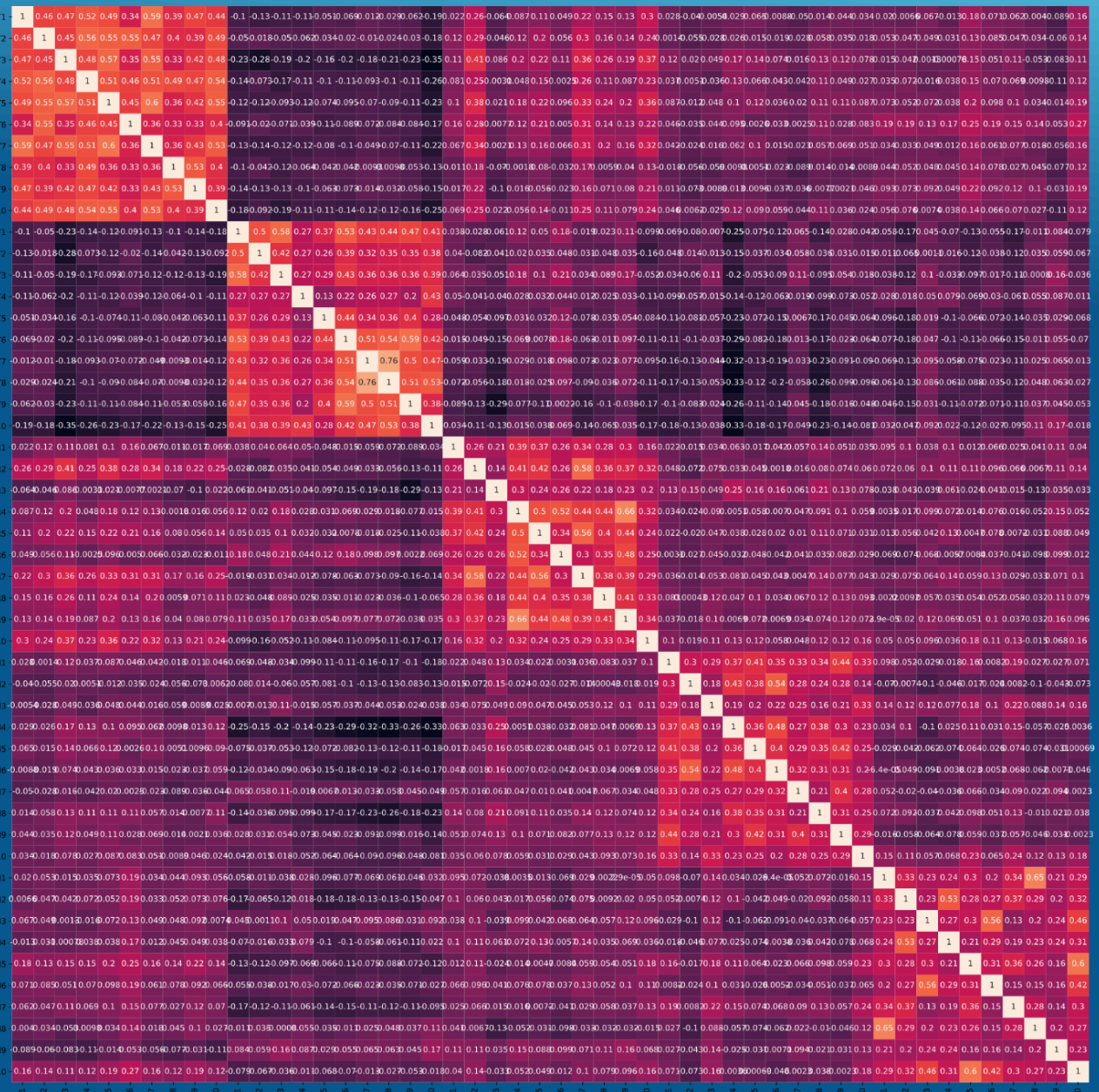
	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	EST1	EST2	EST3	EST4	EST5	EST6	EST7	EST8	EST9	EST10	AGR1	AGR2	AGR3	AGR4	AGR5	AGR6	AGR7	AG
1015336	4.0	2.0	4.0	3.0	4.0	3.0	3.0	3.0	3.0	3.0	4.0	3.0	3.0	3.0	4.0	3.0	4.0	3.0	3.0	3.0	5.0	4.0	2.0	5.0	2.0	4.0	2.0	4
1015337	4.0	3.0	4.0	3.0	3.0	3.0	4.0	4.0	3.0	3.0	4.0	3.0	5.0	1.0	5.0	5.0	4.0	4.0	4.0	5.0	2.0	4.0	1.0	4.0	3.0	5.0	3.0	3
1015338	4.0	2.0	4.0	3.0	5.0	1.0	4.0	2.0	4.0	4.0	3.0	2.0	4.0	3.0	2.0	2.0	4.0	2.0	4.0	1.0	3.0	5.0	5.0	3.0	2.0	3.0	2.0	4
1015339	2.0	4.0	3.0	4.0	2.0	2.0	1.0	4.0	2.0	4.0	4.0	3.0	4.0	2.0	4.0	4.0	2.0	2.0	4.0	4.0	2.0	3.0	2.0	4.0	3.0	4.0	2.0	4
1015340	4.0	2.0	4.0	2.0	4.0	1.0	4.0	2.0	4.0	4.0	4.0	3.0	4.0	3.0	2.0	3.0	3.0	1.0	4.0	2.0	1.0	5.0	2.0	4.0	3.0	5.0	2.0	3

```
df.shape
```

```
(1015341, 109)
```

OUR OBJECTIVE:
IDENTIFYING TYPICAL PERSONALITIES AMONG SUBJECTS

EDA:

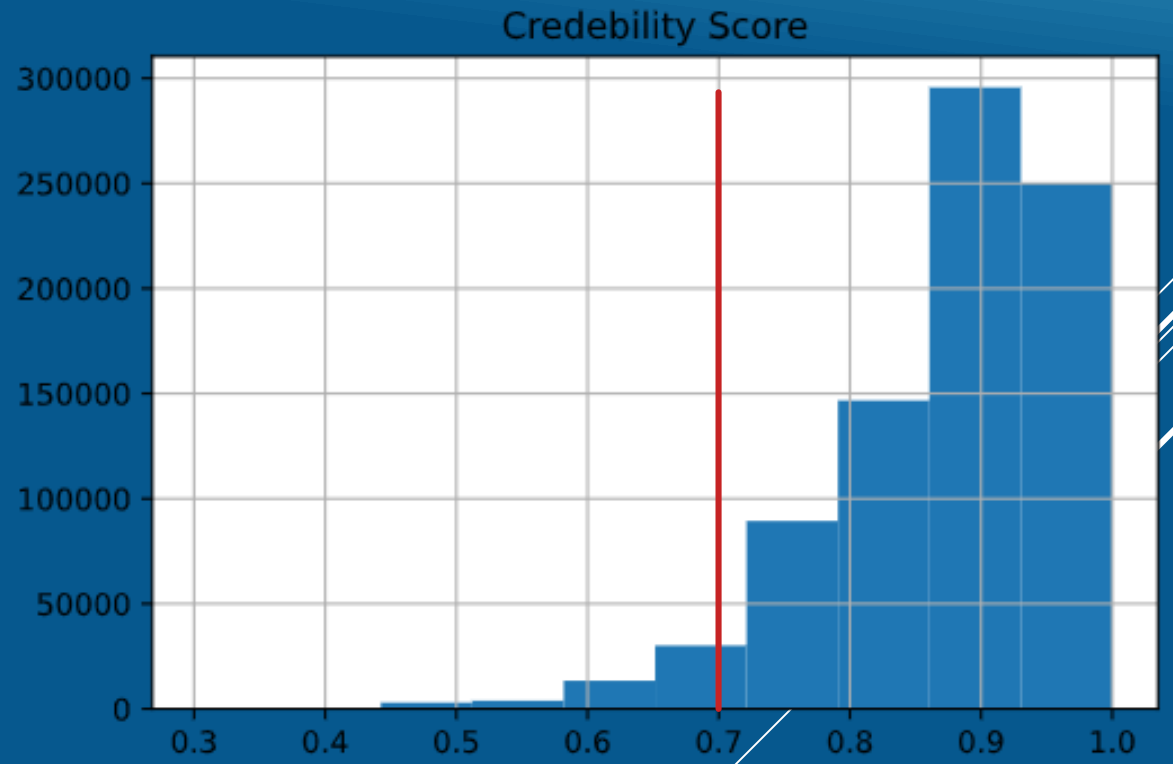

EXT1 EXT2 EXT3 EXT4 EXT5 EXT6 EXT7 EXT8 EXT9 EXT10
EST1 EST2 EST3 EST4 EST5 EST6 EST7 EST8 EST9 EST10
AGR1 AGR2 AGR3 AGR4 AGR5 AGR6 AGR7 AGR8 AGR9 AGR10
CSN1 CSN2 CSN3 CSN4 CSN5 CSN6 CSN7 CSN8 CSN9 CSN10
OPN1 OPN2 OPN3 OPN4 OPN5 OPN6 OPN7 OPN8 OPN9 OPN10

EDA:

Subjects authentication:

- Multiple subjects of same IP –

  Same user or shared PC?
- Rapid repliers – less than 3 sec\questions
- Subject credibility - credibility scale (0-1):



R=0.76

EST7: I change my mood a lot. 1 2 3 4 5
EST8: I have frequent mood swings. 1 2 3 4 5

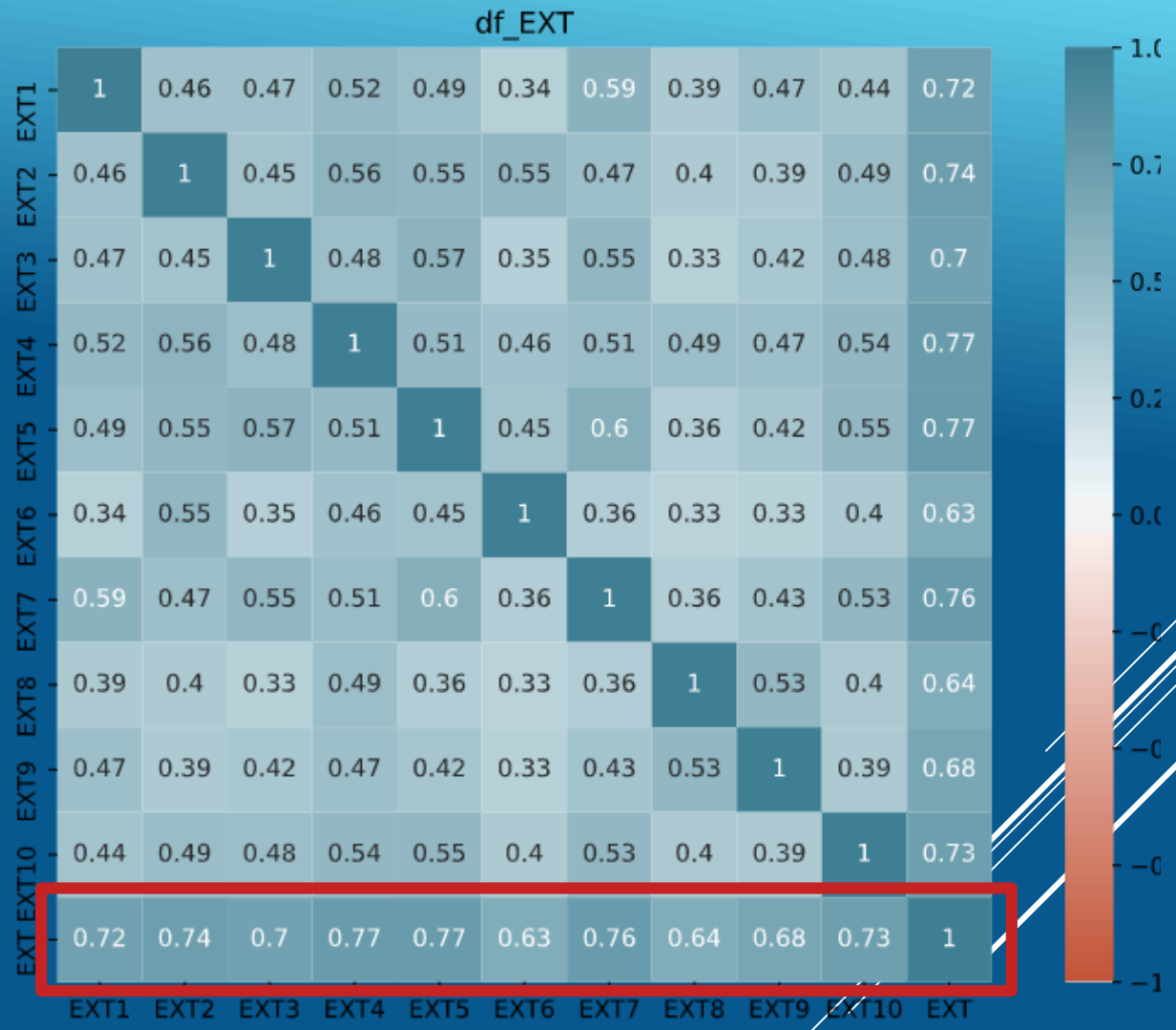
d=|5-2| = 3

EDA:

Missing data:

- Missing > 5 = drop
- Using inner scale correlations to replace missing data (linear model) -
- Iterative generation of regression model for each column:

  Bayesian Ridge estimator



EXT_7 = ?

EXT_7: I talk to a lot of different people at parties

1 2 3 4 5

FEATURE ENGINEERING:

- Drop all irrelevant data (GIS, screen, times, etc.)

Questions data:

- Aligning negative questions:

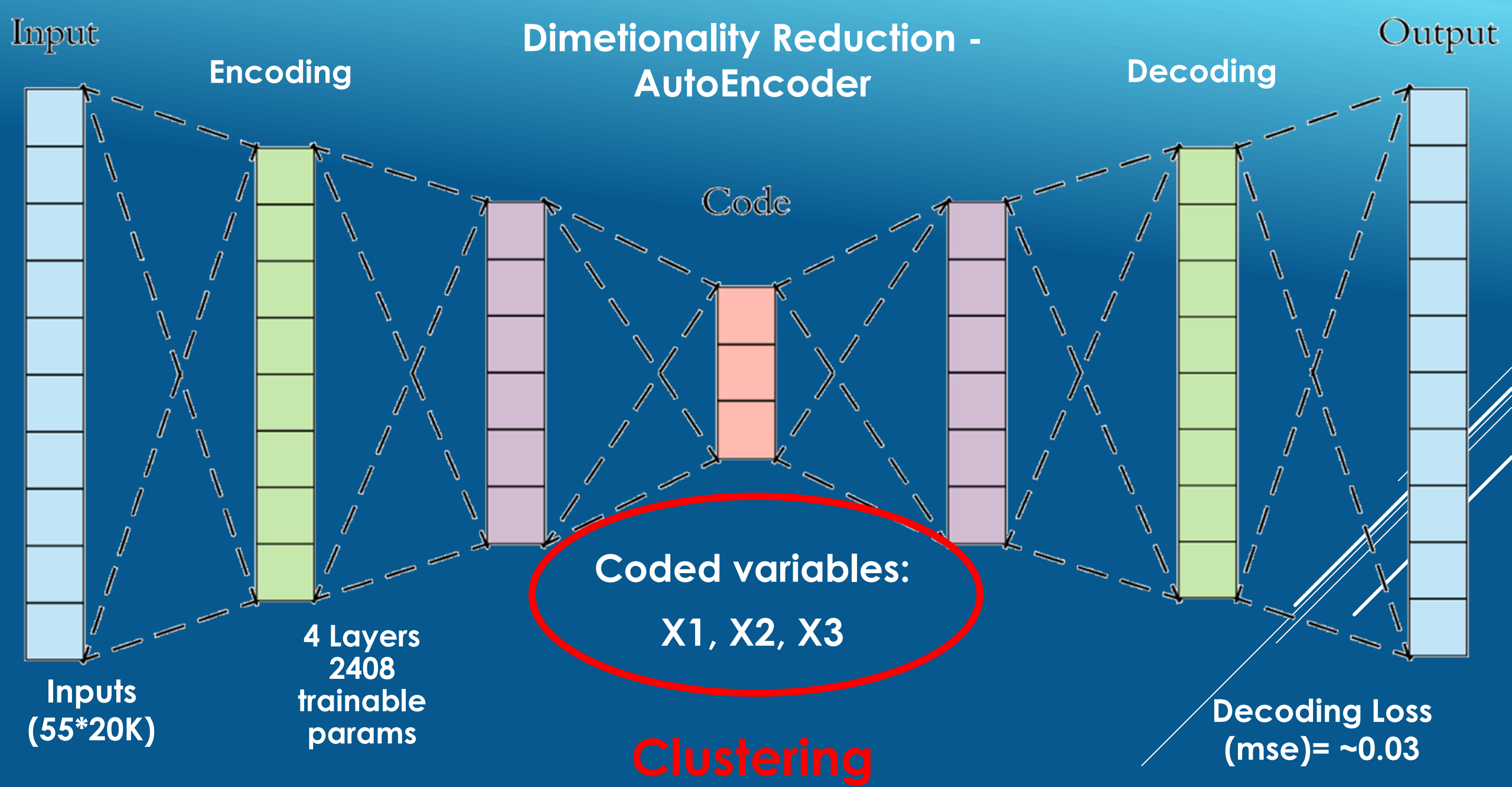
OPN_3: I have a vivid imagination (+)
OPN_6: I don't have a good imagination. (-)

Dimensionality Reduction:

- For each trait:
  - Sums of all trait-related scores to one scale:

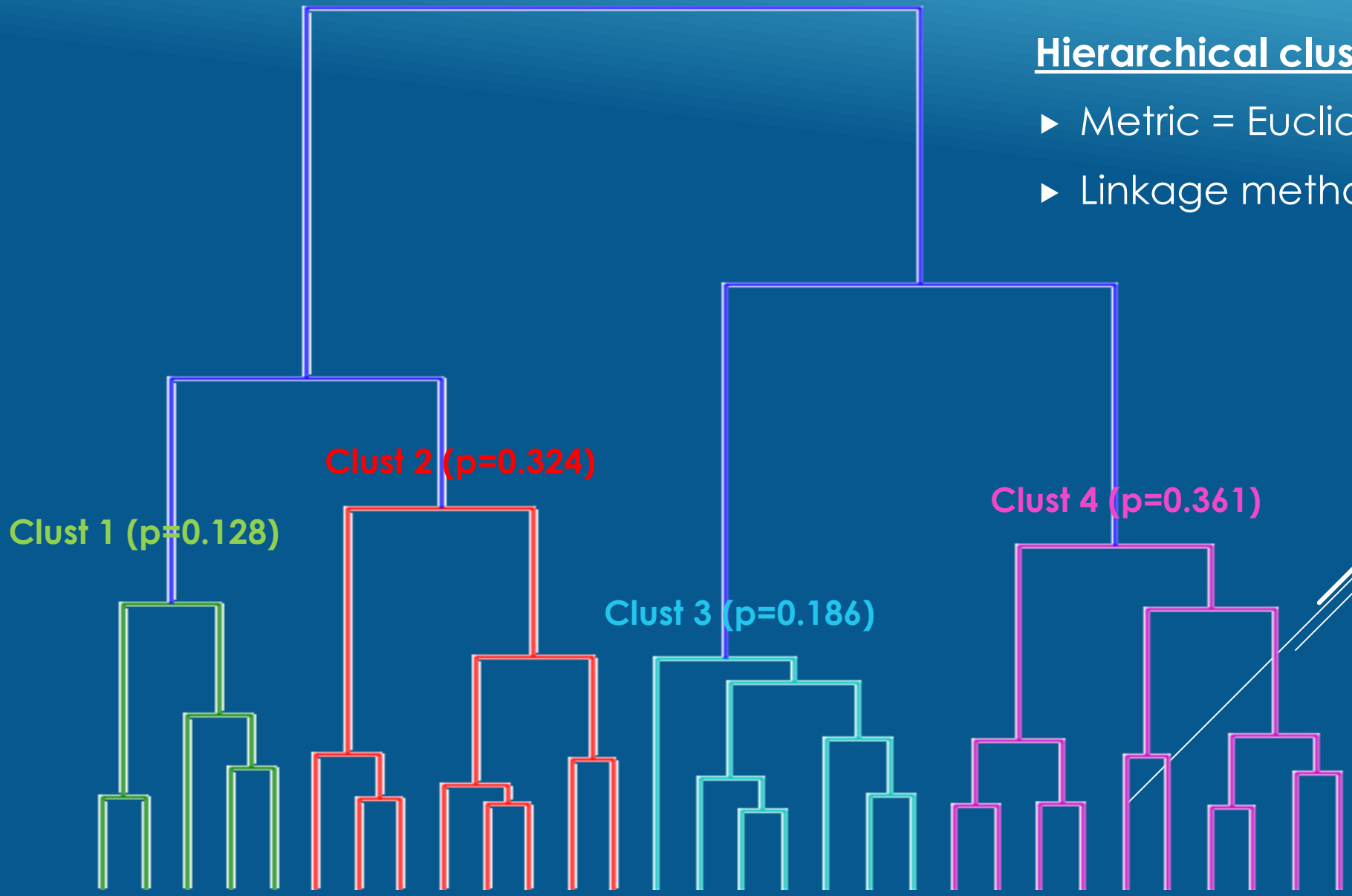
Increase of column variance
(range 1-5) ➔ (range 10-50)

Dimetionality Reduction - AutoEncoder



Clustering

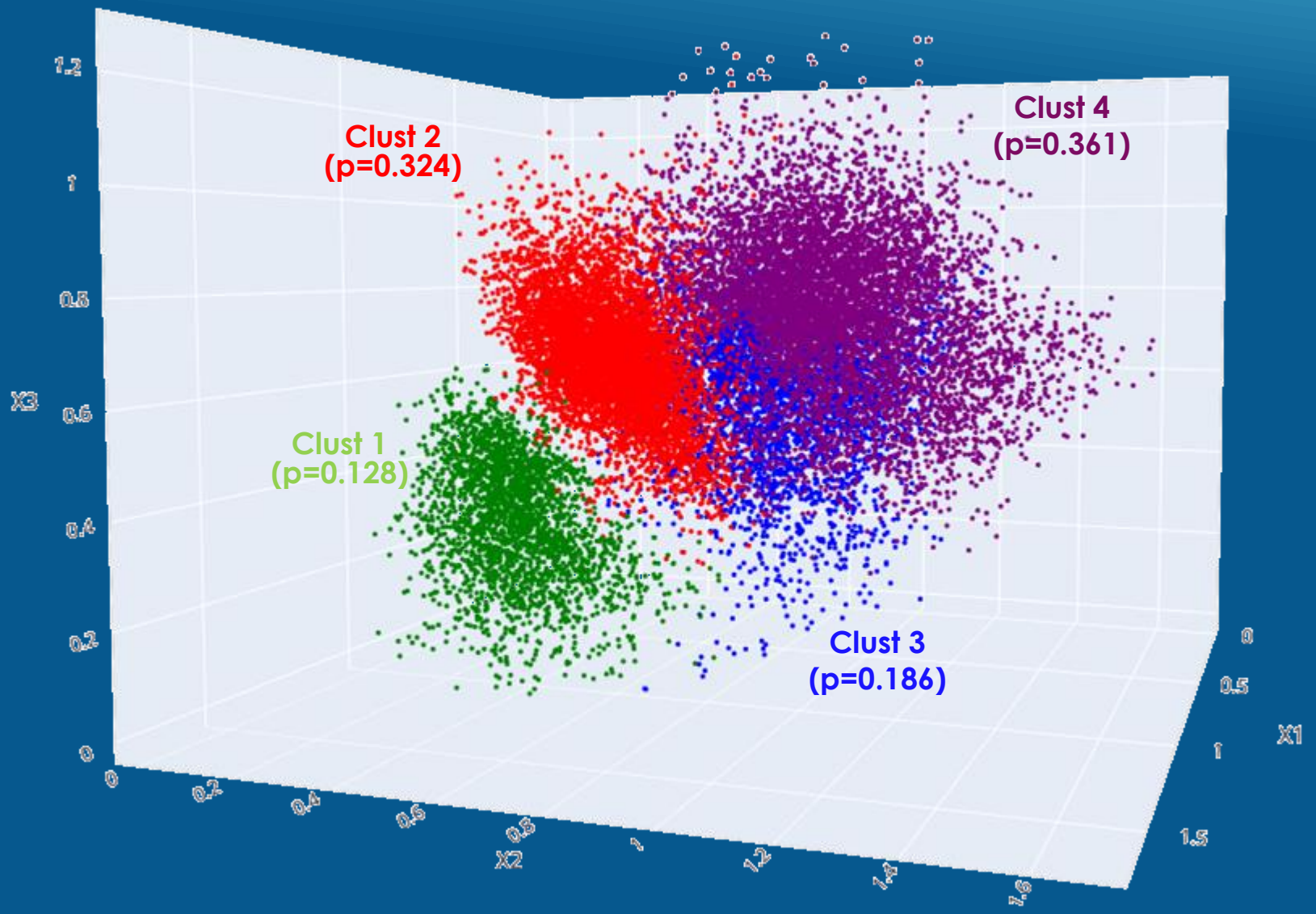
CLUSTERING (AE X's dimensions):

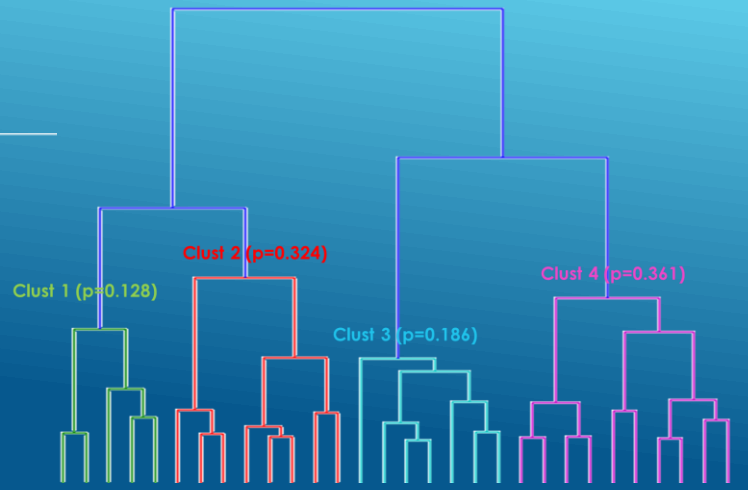


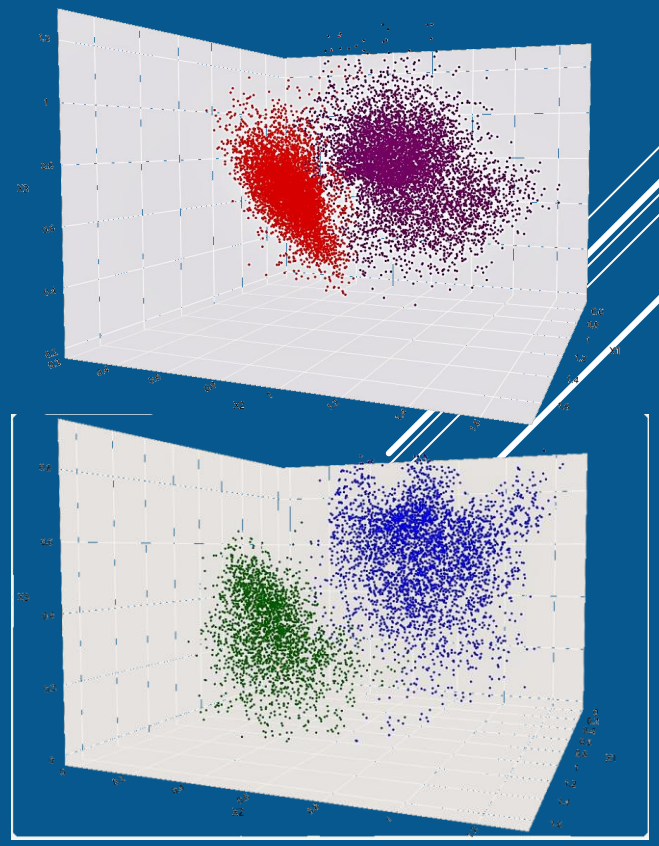
Hierarchical clustering

- Metric = Euclidean ,
- Linkage method = Ward

CLUSTERING (AE X's dimensions):

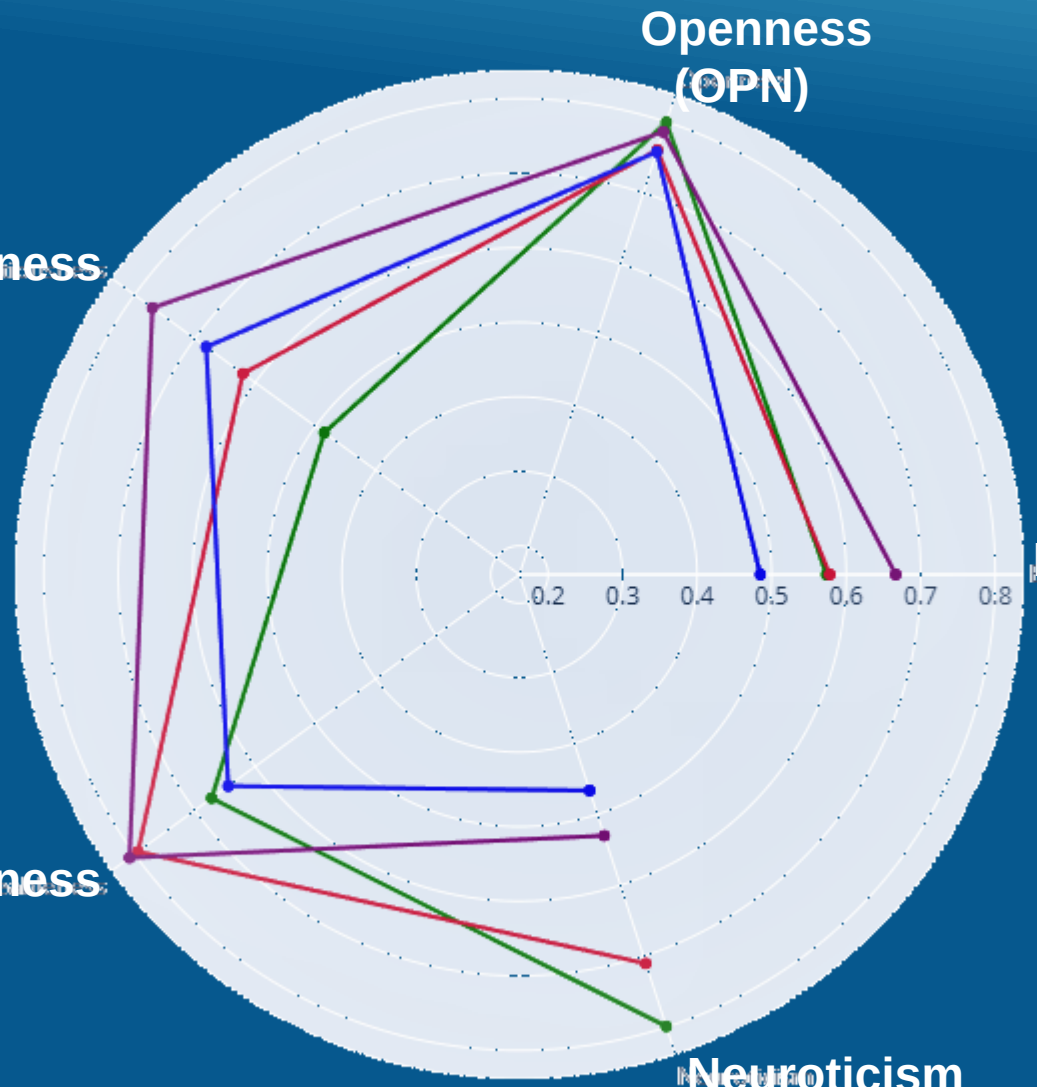


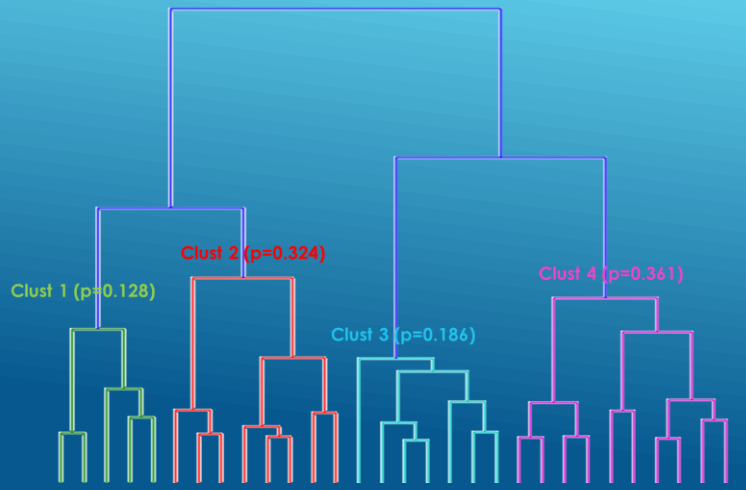




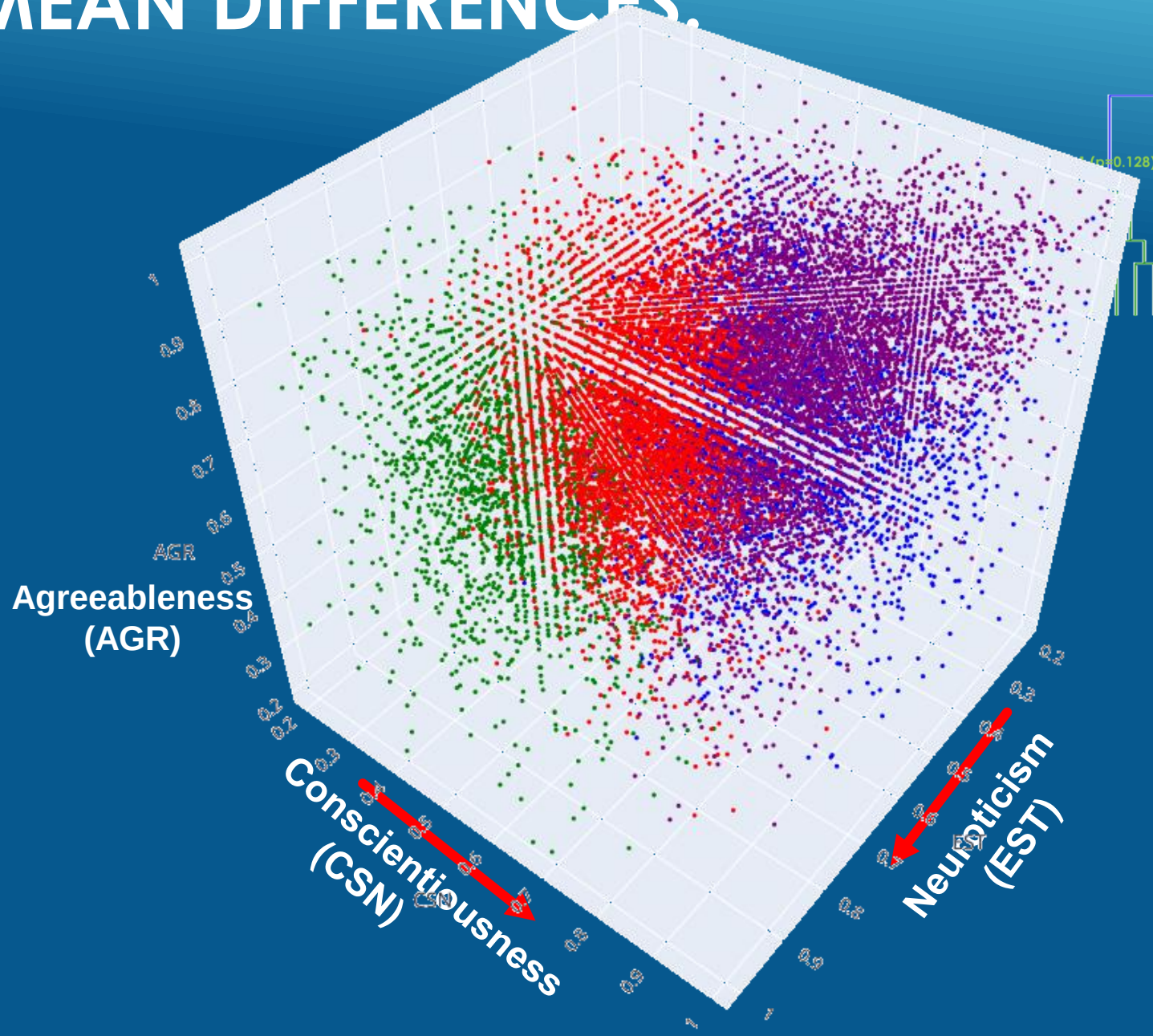

SO WHERE ARE THE DIFFERENCES?

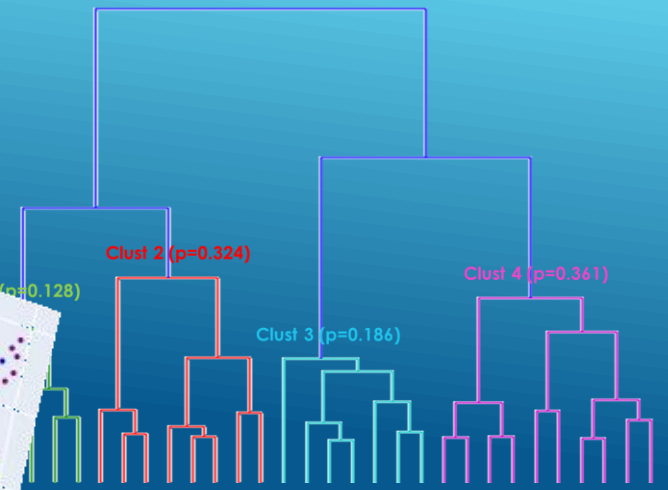
TRAIT'S MEAN DIFFERENCES:



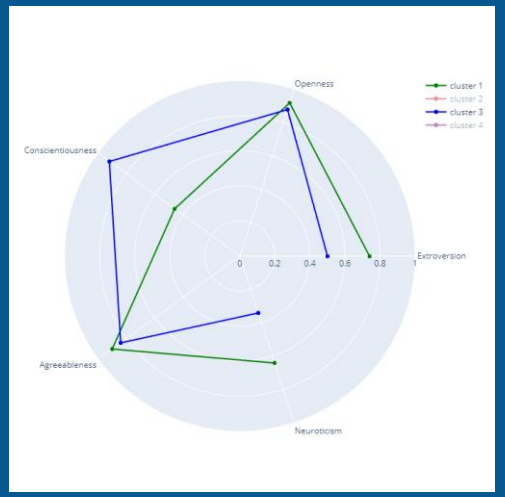


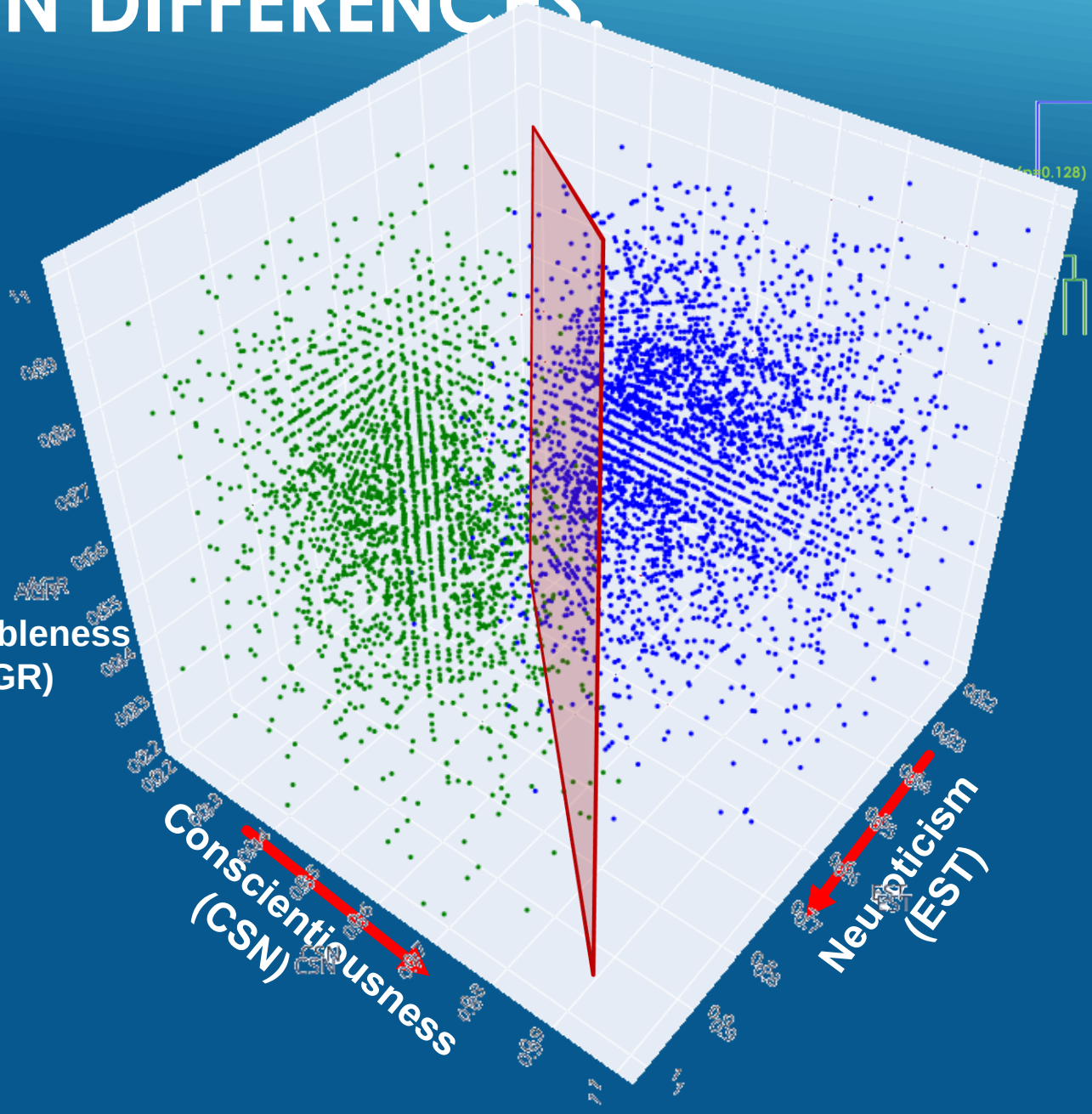
TRAIT'S MEAN DIFFERENCES:

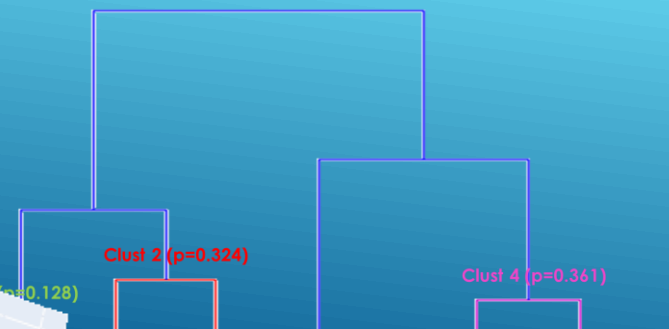


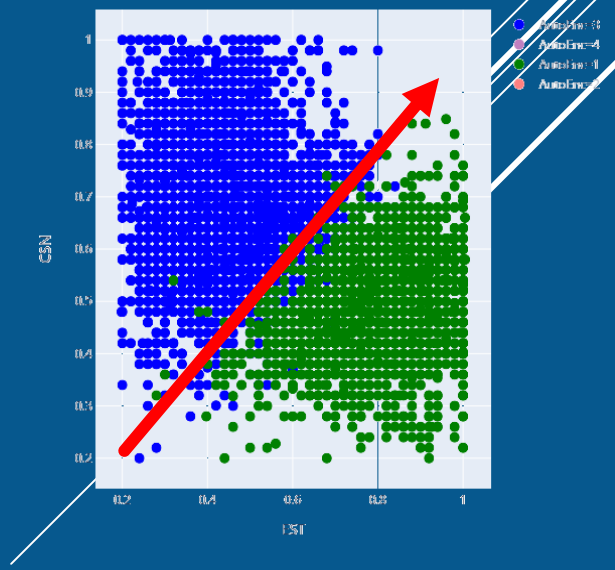


TRAIT'S MEAN DIFFERENCES:

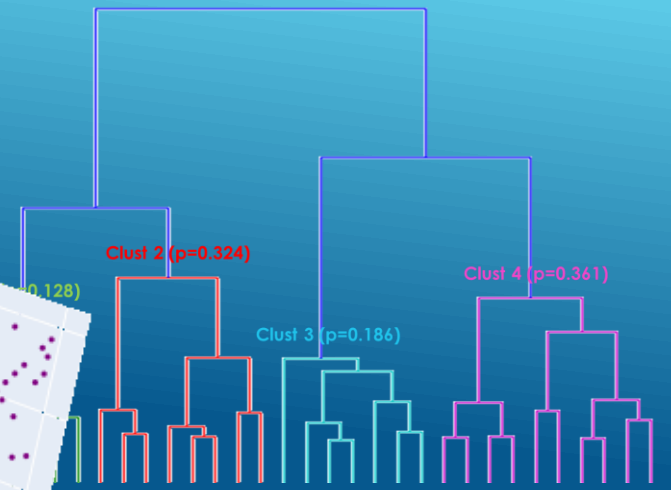


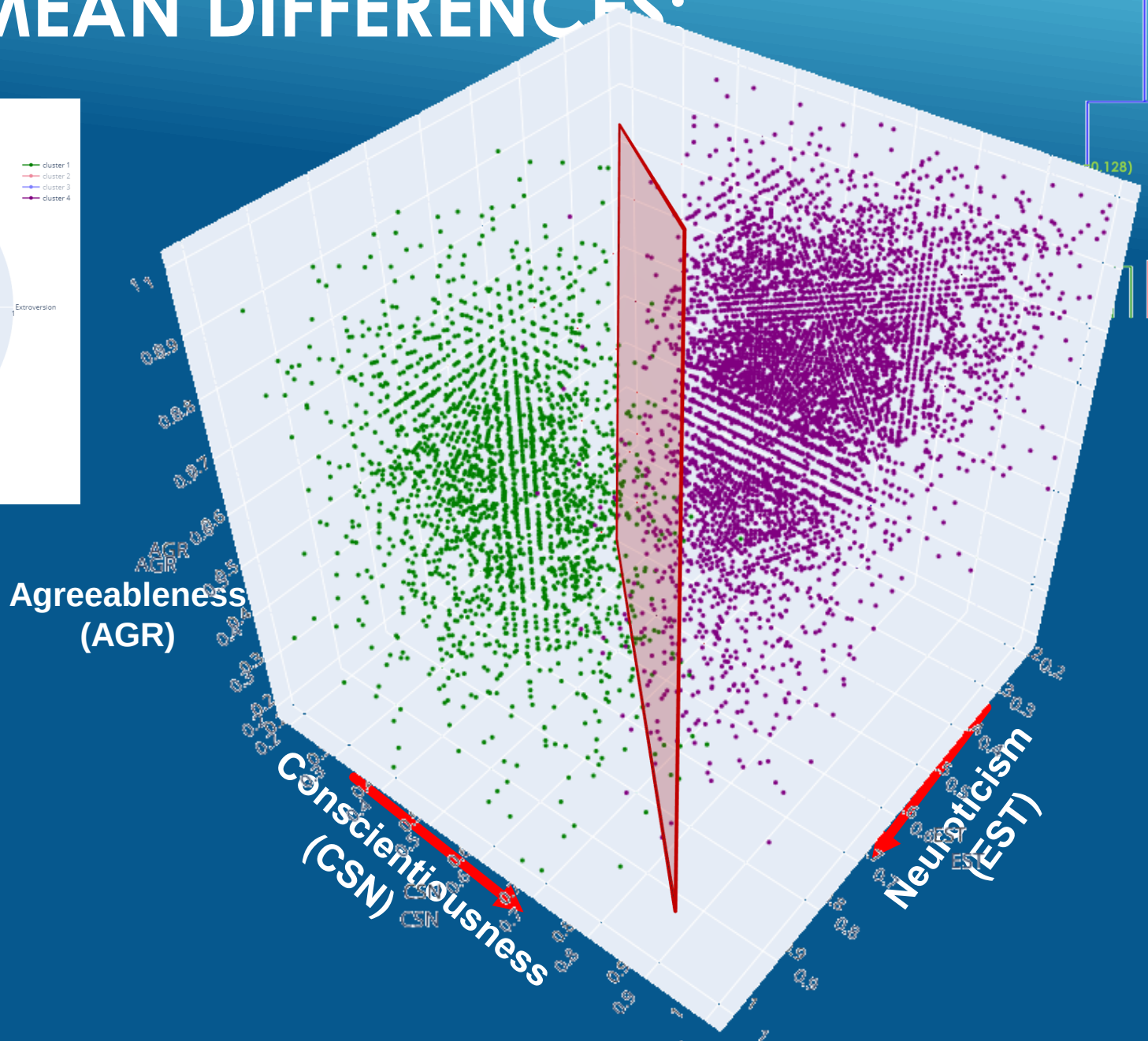


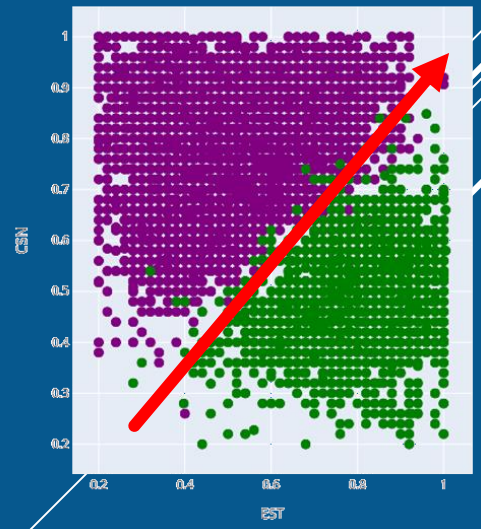




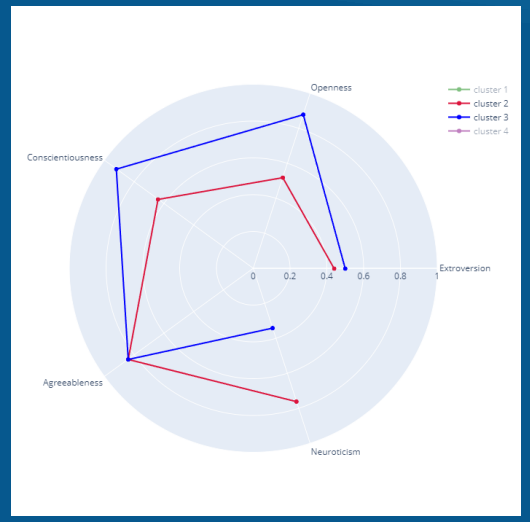
TRAIT'S MEAN DIFFERENCES:

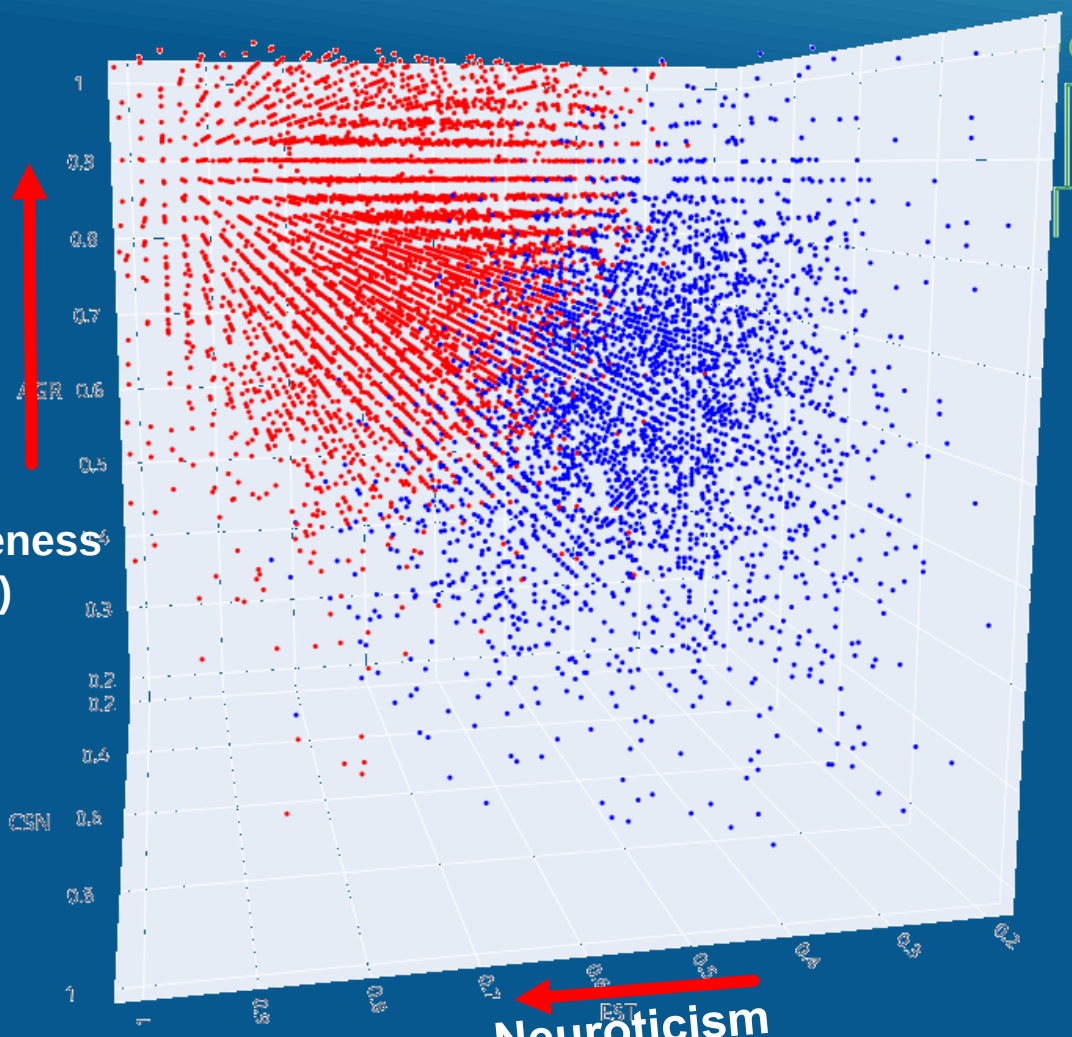






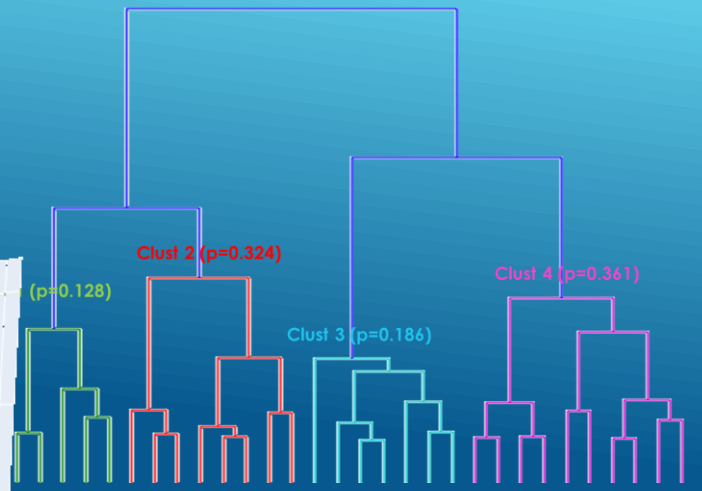
TRAIT'S MEAN DIFFERENCES:

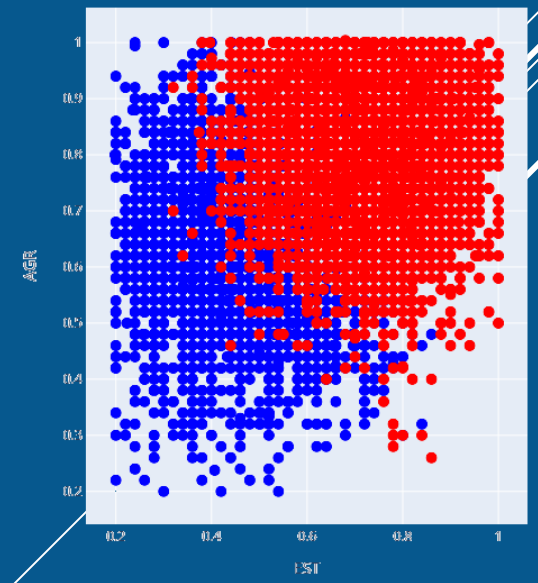




Agreeableness
(AGR)

Neuroticism
(EST)





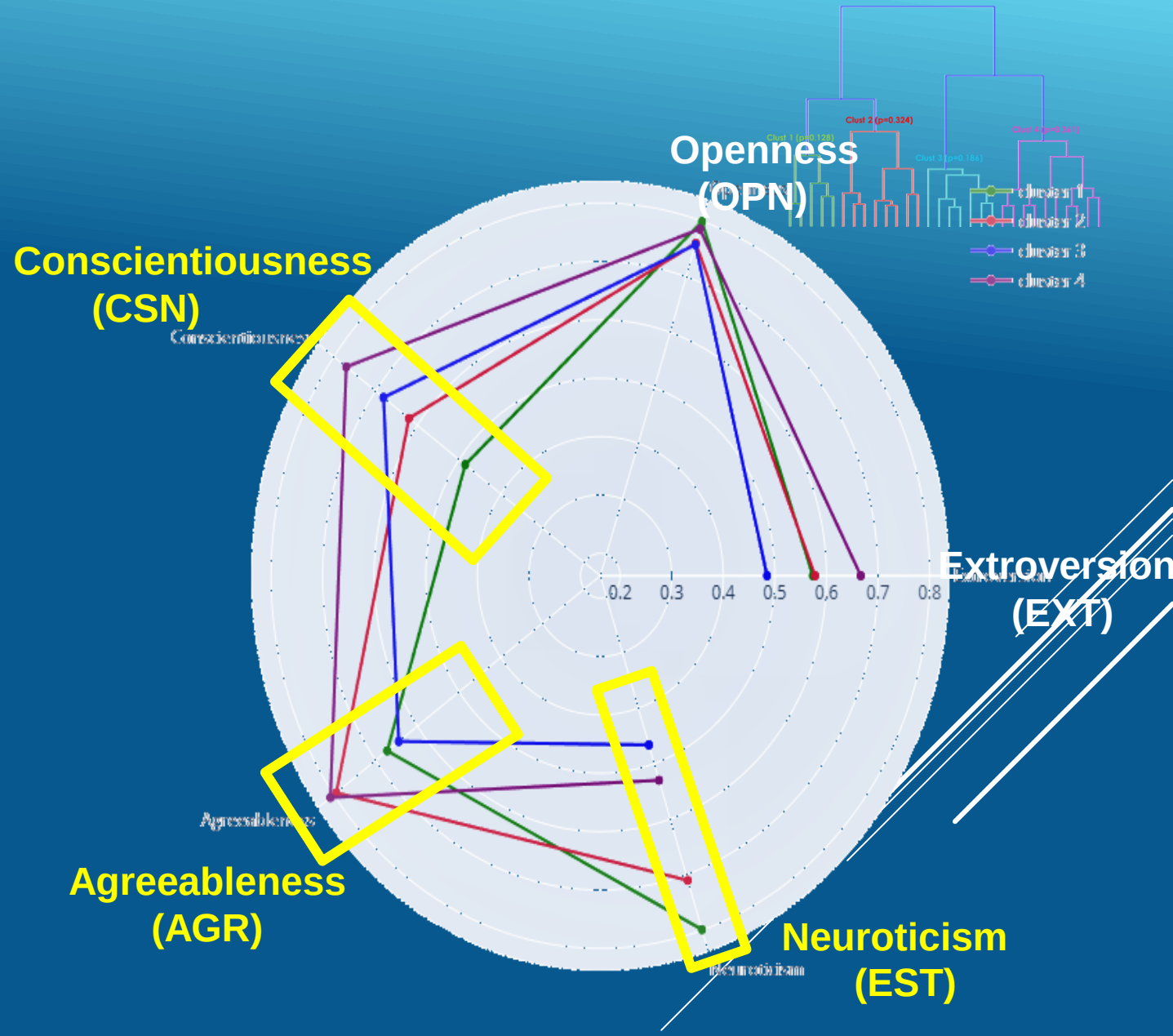
INSIGHTS:

Most robust predictors for OCEAN clustering are:

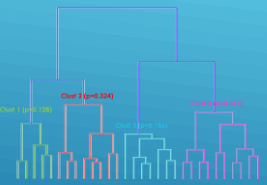
A. **Neuroticism** (i.e. mood stability, calmness, self-confidence, etc.)

B. **Conscientiousness** (i.e. self discipline, responsibility, self organization, etc.)

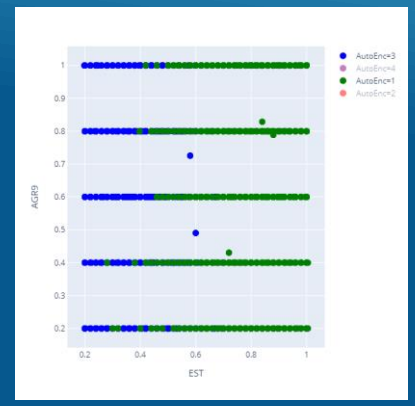
C. **Agreeableness** (Sociality, care for others, selfishness, etc.)



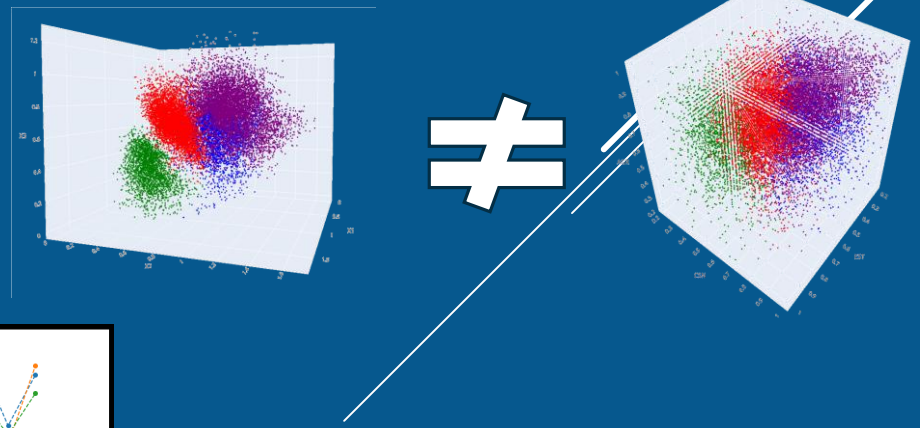
INSIGHTS:



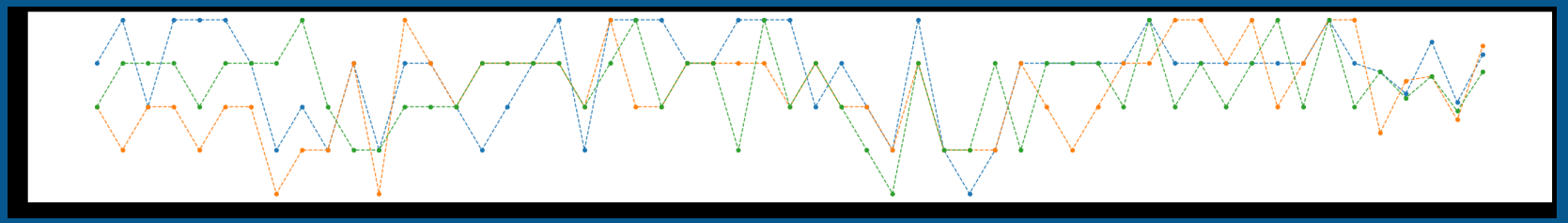
- Clustering is only a half job done
- Subjective reporting is not a continuous measure
  - Designed metrics (cosine?)
  - Very low variance



- Distinctive features aren’t Encoded features



- Possible analyses directions:
  CNN (image-like), NLP (text-like)

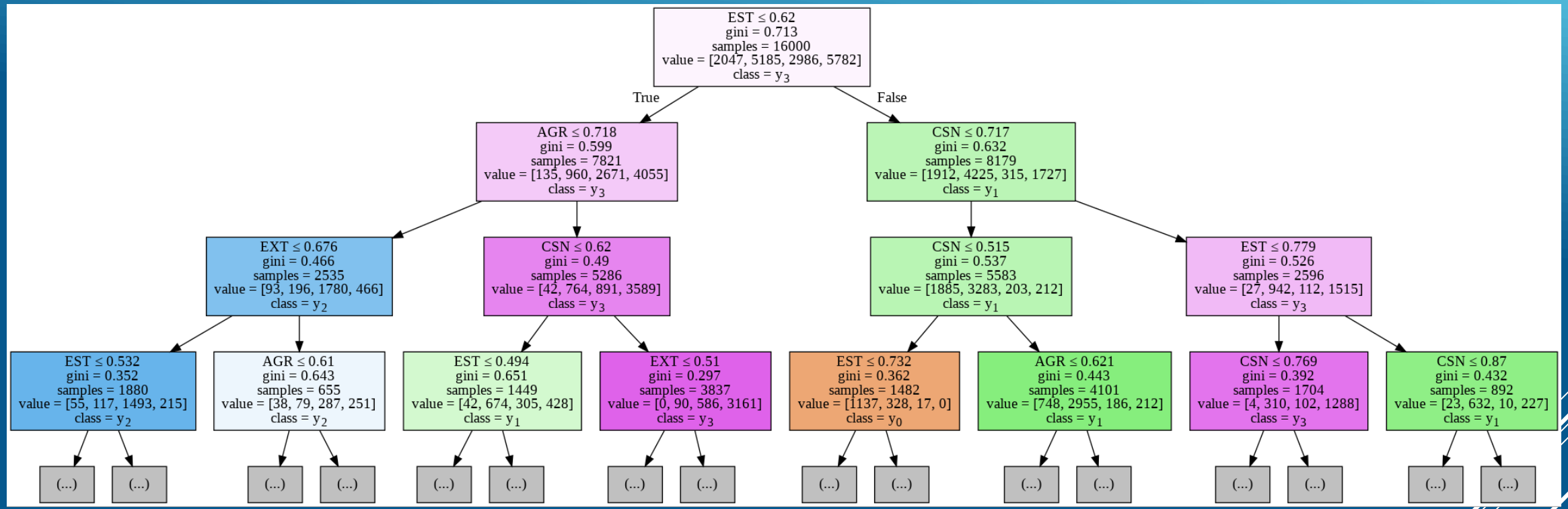


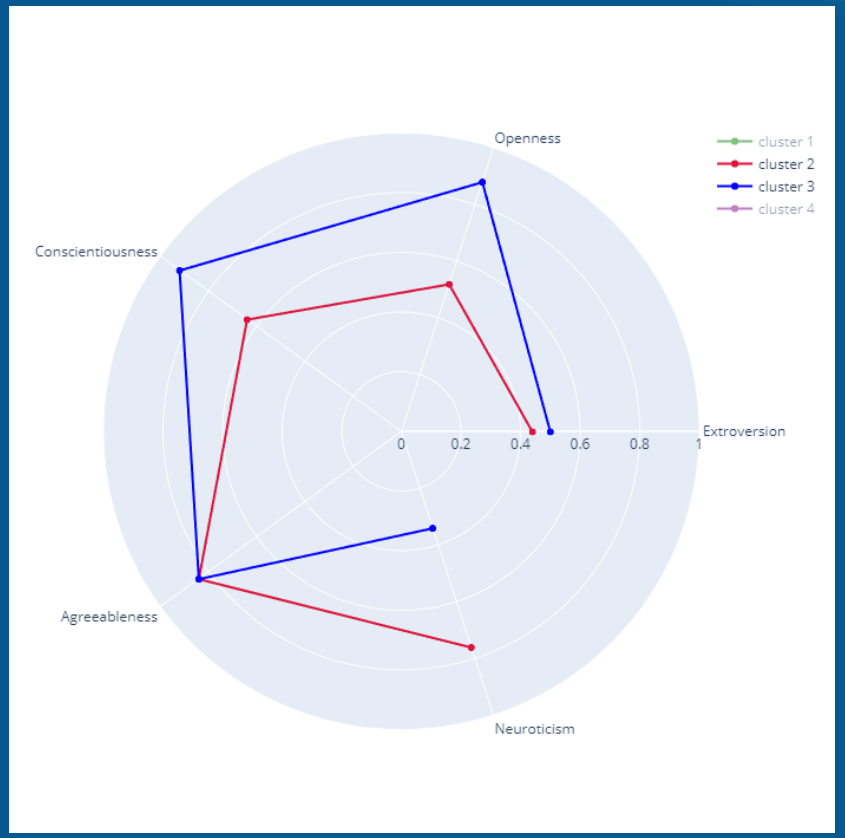
Special Thanks:

Amit Rappel

NAYA-team

THANK YOU FOR LISTENING


EST ≤ 0.62
gini = 0.713
samples = 16000
value = [2047, 5185, 2986, 5782]
class = y3
True
False
AGR ≤ 0.718
gini = 0.599
samples = 7821
value = [135, 960, 2671, 4055]
class = y3
CSN ≤ 0.717
gini = 0.632
samples = 8179
value = [1912, 4225, 315, 1727]
class = y1
EXT ≤ 0.676
gini = 0.466
samples = 2535
value = [93, 196, 1780, 466]
class = y2
CSN ≤ 0.62
gini = 0.49
samples = 5286
value = [42, 764, 891, 3589]
class = y3
CSN ≤ 0.515
gini = 0.537
samples = 5583
value = [1885, 3283, 203, 212]
class = y1
EST ≤ 0.779
gini = 0.526
samples = 2596
value = [27, 942, 112, 1515]
class = y3
EST ≤ 0.532
gini = 0.352
samples = 1880
value = [55, 117, 1493, 215]
class = y2
AGR ≤ 0.61
gini = 0.643
samples = 655
value = [38, 79, 287, 251]
class = y2
EST ≤ 0.494
gini = 0.651
samples = 1449
value = [42, 674, 305, 428]
class = y1
EXT ≤ 0.51
gini = 0.297
samples = 3837
value = [0, 90, 586, 3161]
class = y3
EST ≤ 0.732
gini = 0.362
samples = 1482
value = [1137, 328, 17, 0]
class = y0
AGR ≤ 0.621
gini = 0.443
samples = 4101
value = [748, 2955, 186, 212]
class = y1
CSN ≤ 0.769
gini = 0.392
samples = 1704
value = [4, 310, 102, 1288]
class = y3
CSN ≤ 0.87
gini = 0.432
samples = 892
value = [23, 632, 10, 227]
class = y1
(...)


Openness
Conscientiousness
Extroversion
Agreeableness
Neuroticism
0
0.2
0.4
0.6
0.8
1
cluster 1
cluster 2
cluster 3
cluster 4


Openness
Conscientiousness
E
Agreeableness
Neuroticism
0
0.2
0.4
0.6
0.8
1
cluster 1
cluster 1
cluster 1
cluster 1
cluster 1
cluster 2
cluster 2
cluster 2
cluster 2
cluster 2
cluster 3
cluster 3
cluster 3
cluster 3
cluster 3
cluster 4
cluster 4
cluster 4
cluster 4
cluster 4