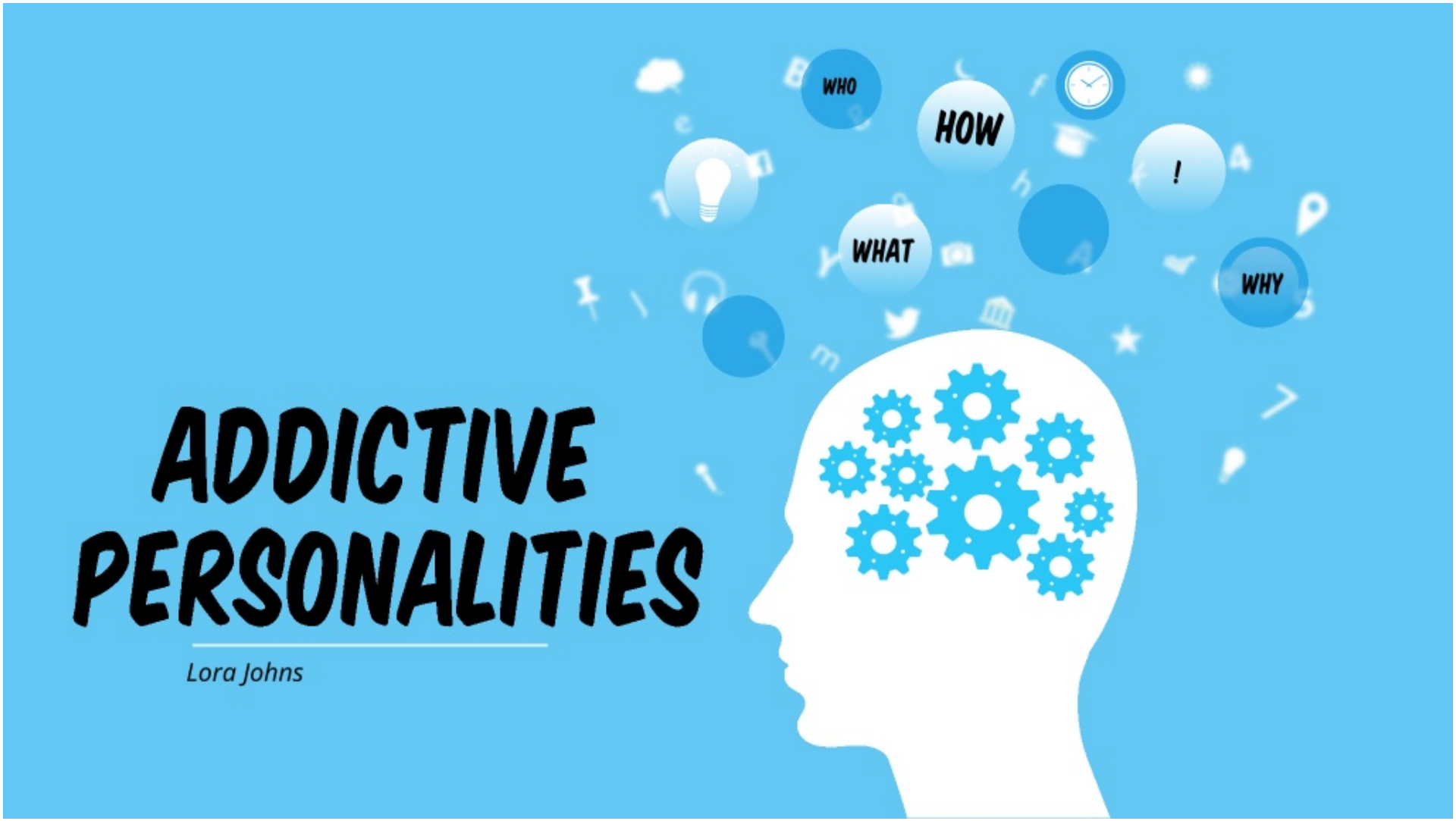
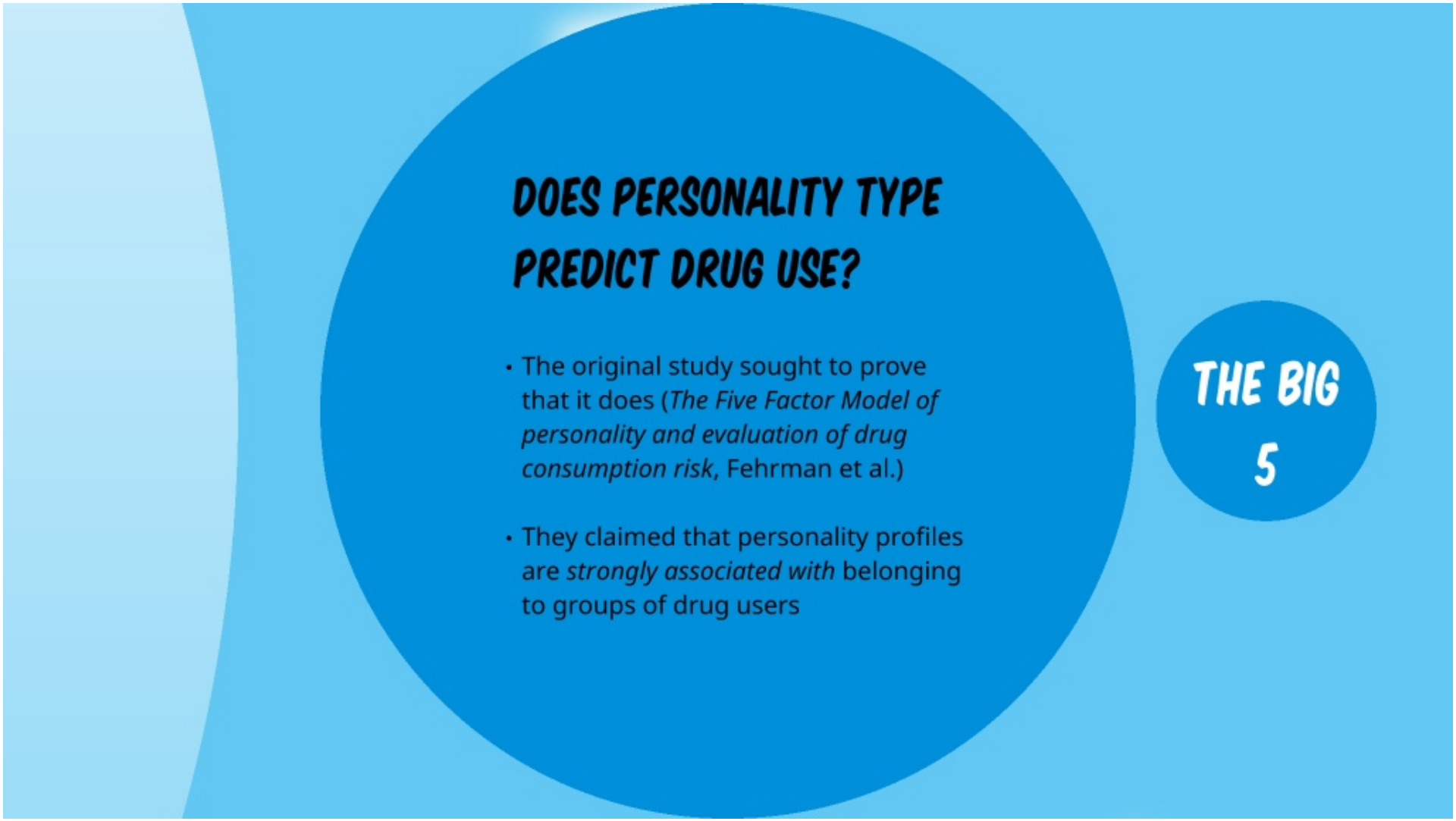
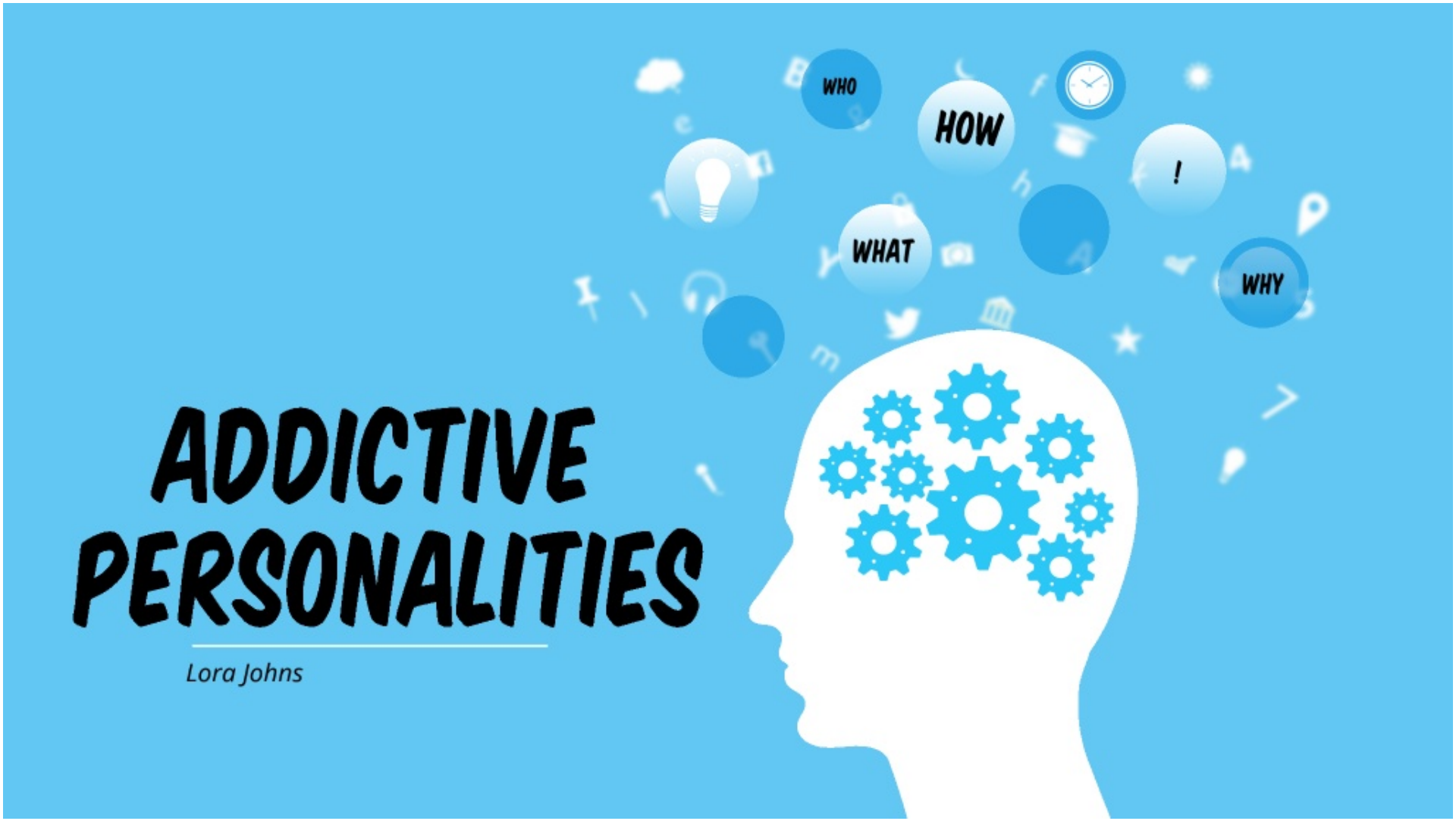
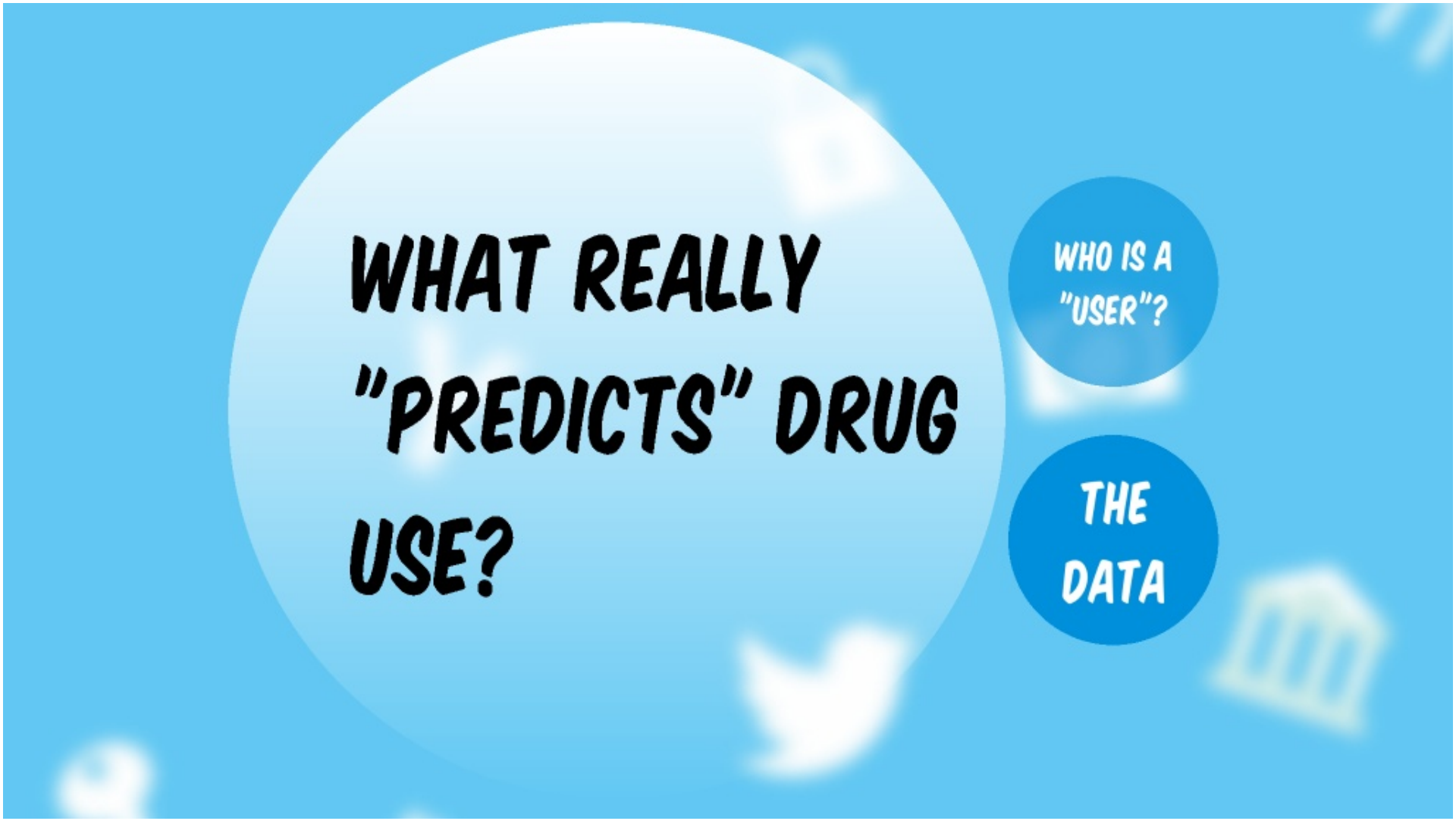

WHO
HOW
!
WHAT
WHY
ADDICTIVE PERSONALITIES
Lora Johns


DRUG USE &
THE BIG FIVE
WHAT'S
YOUR TYPE?

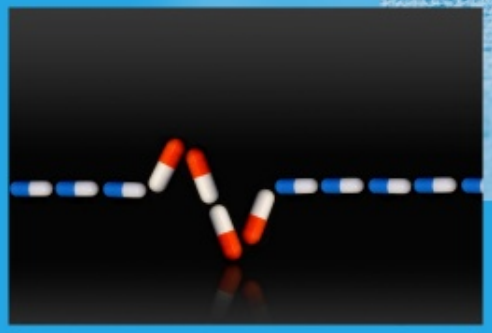

DOES PERSONALITY TYPE PREDICT DRUG USE?
• The original study sought to prove that it does (*The Five Factor Model of personality and evaluation of drug consumption risk*, Fehrman et al.)
• They claimed that personality profiles are *strongly associated with* belonging to groups of drug users
THE BIG 5

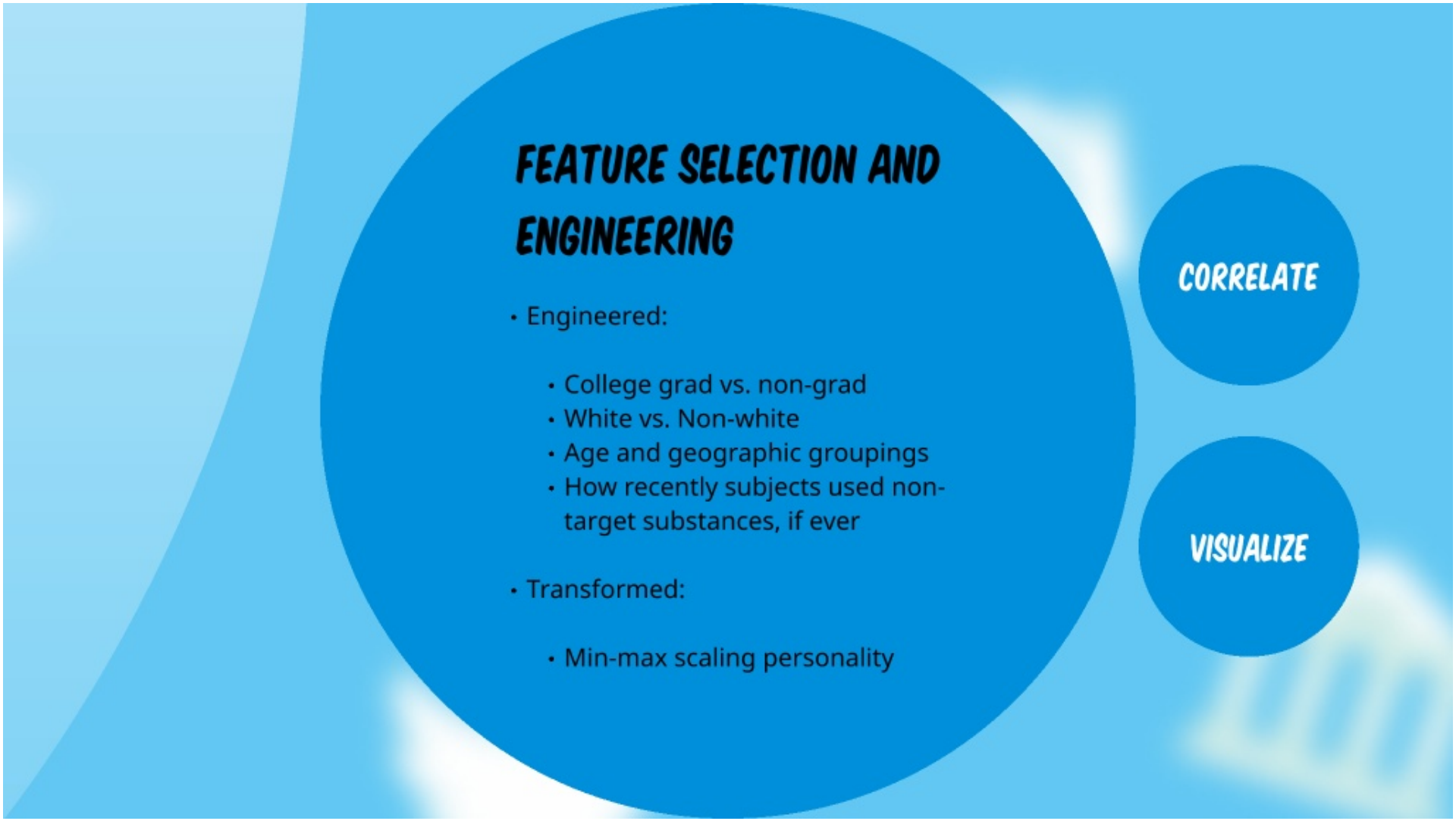

• NEUROTICISM
• EXTRAVERSION
• OPENNESS TO EXPERIENCE
• AGREEABLENESS
• CONSCIENTIOUSNESS

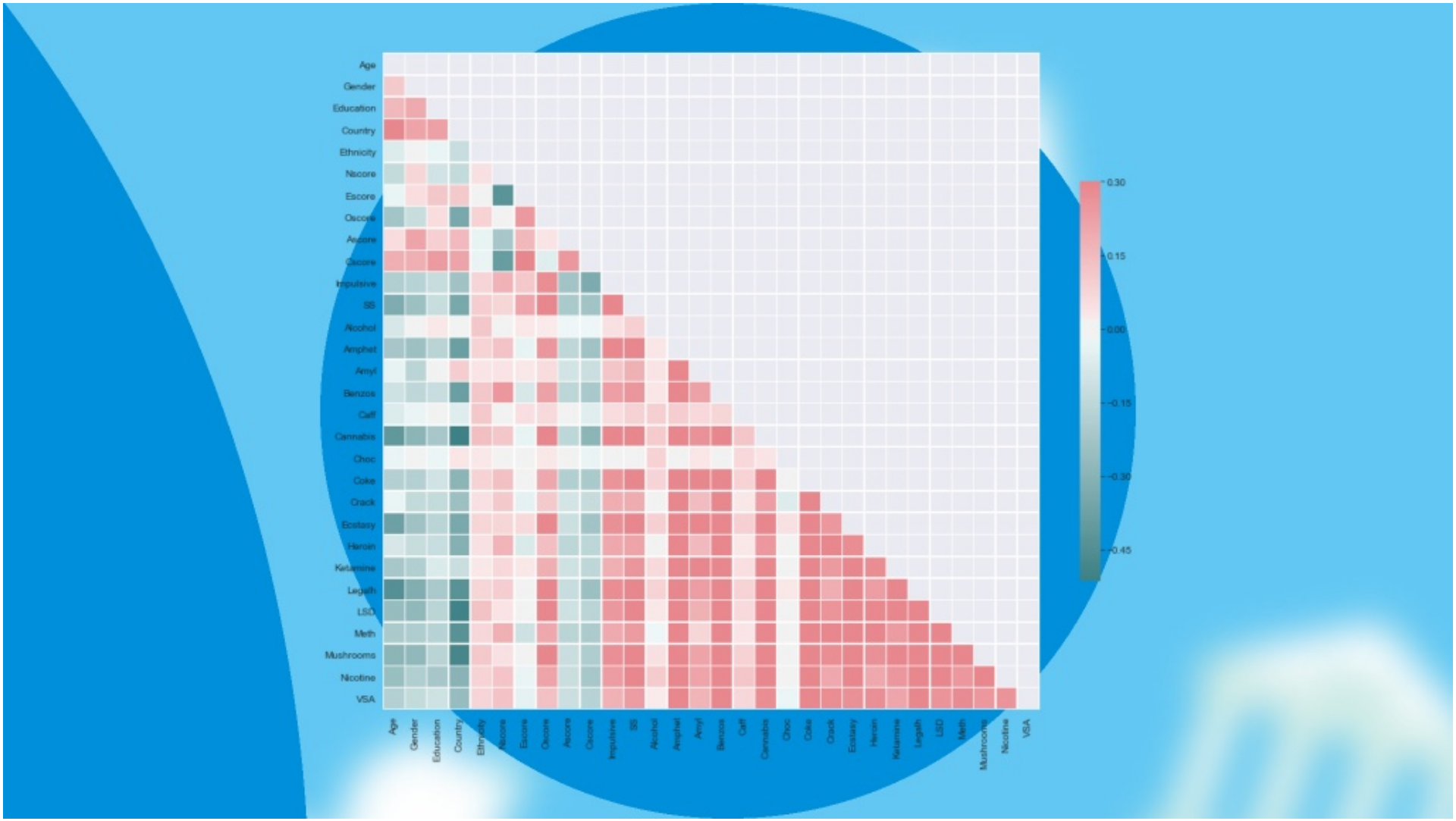

WHO
HOW
!
WHAT
WHY
ADDICTIVE PERSONALITIES
Lora Johns

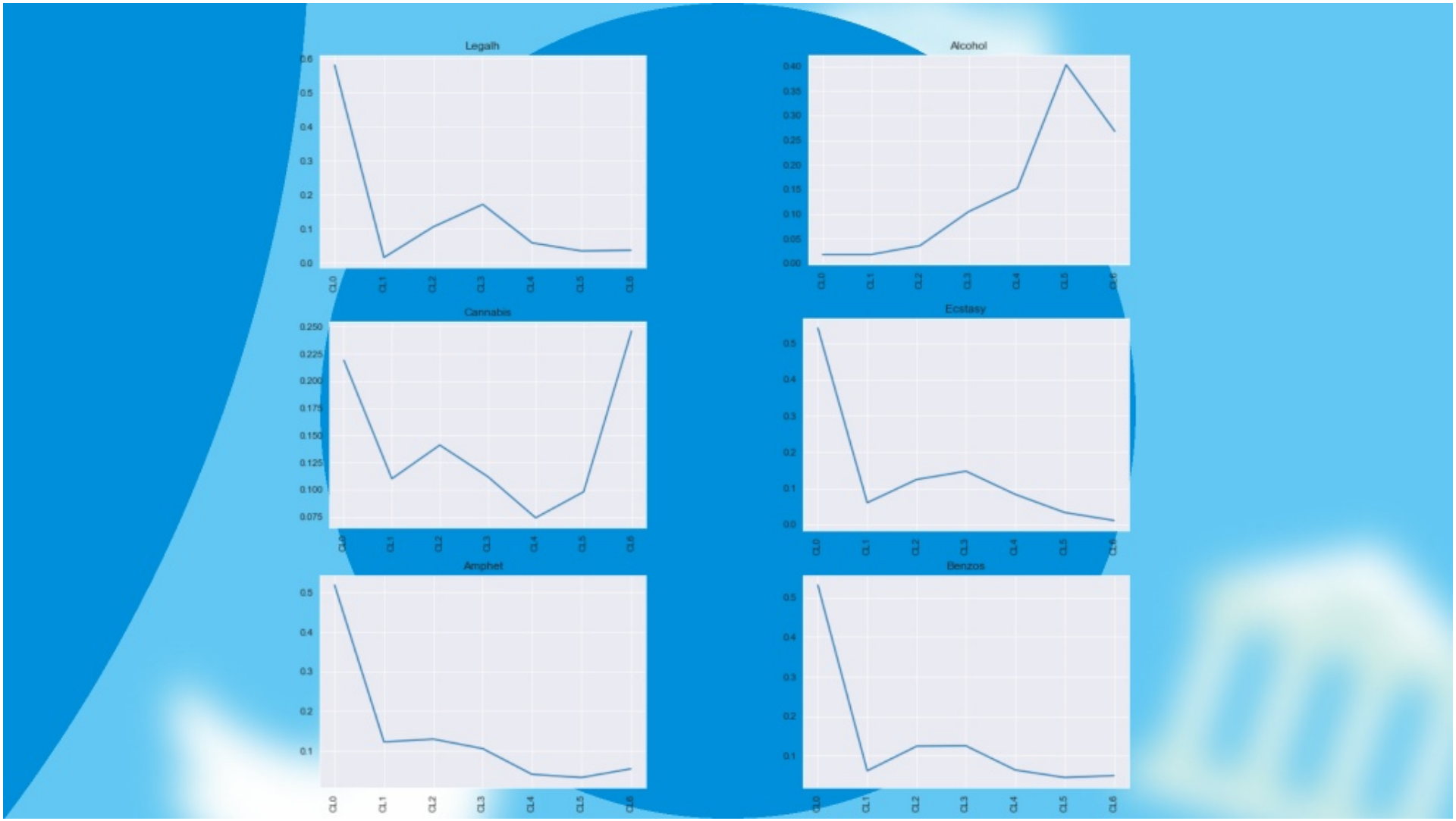

WHAT REALLY "PREDICTS" DRUG USE?
WHO IS A "USER"?
THE DATA

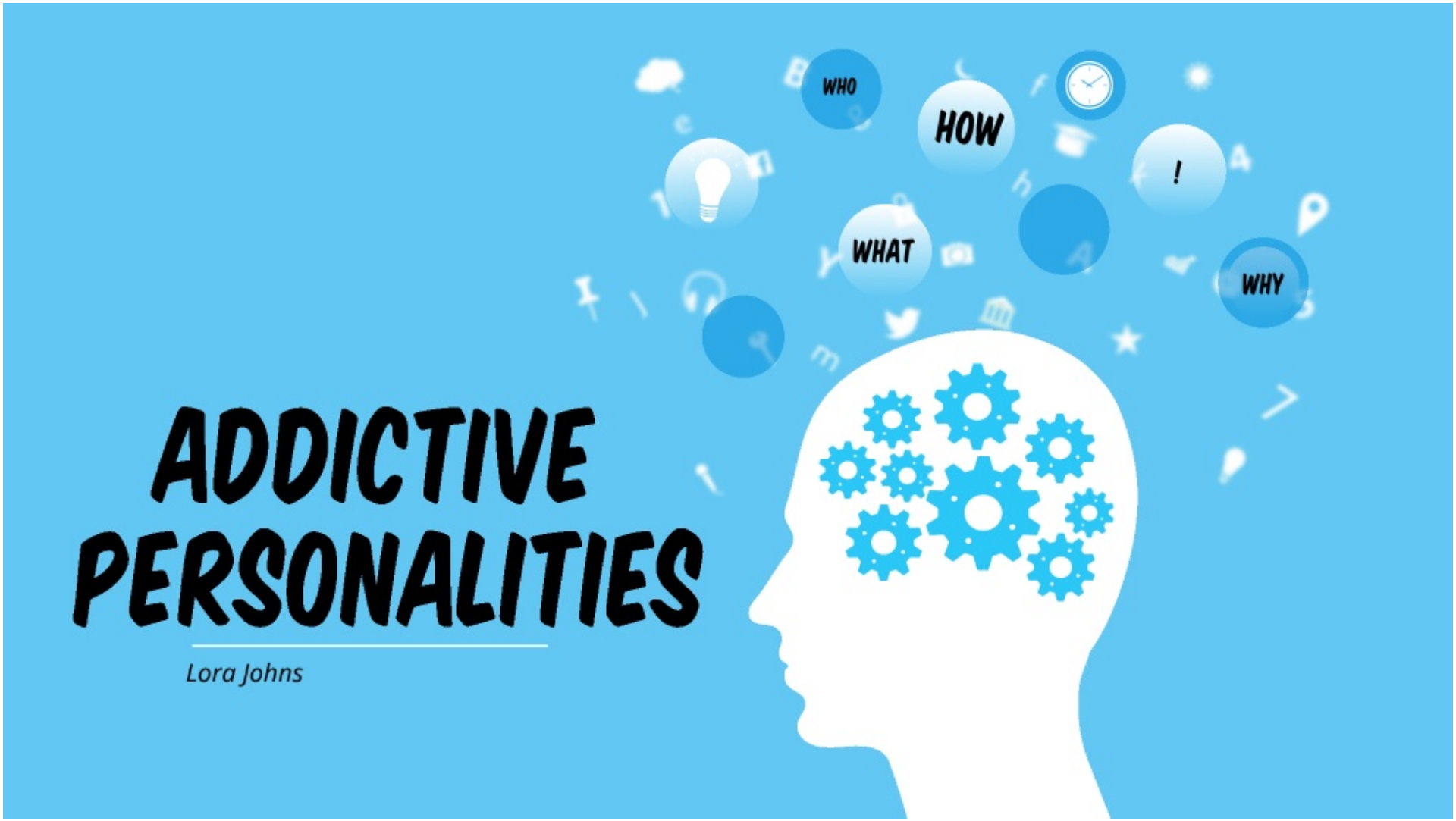
- For this analysis, an "active user":
  - Used Heroin, Methadone, Crack, or Cocaine
  - Within the last year
- This definition underlies the target variable. Why?
  - Highest risk population (real-world)
  - Create a reasonable target variable without introducing confusion (pragmatic)




FEATURE SELECTION AND ENGINEERING
• Engineered:
• College grad vs. non-grad
• White vs. Non-white
• Age and geographic groupings
• How recently subjects used non-target substances, if ever
• Transformed:
• Min-max scaling personality
CORRELATE
VISUALIZE


Age
Gender
Education
Country
Ethnicity
Nscore
Escore
Oscore
Ascore
Cscore
Impulsive
SS
Alcohol
Amphet
Amyl
Benzos
Caff
Cannabis
Choc
Coke
Crack
Ecstasy
Heroin
Ketamine
Legalh
LSD
Meth
Mushrooms
Nicotine
VSA
0.30
0.15
0.00
−0.15
−0.30
−0.45

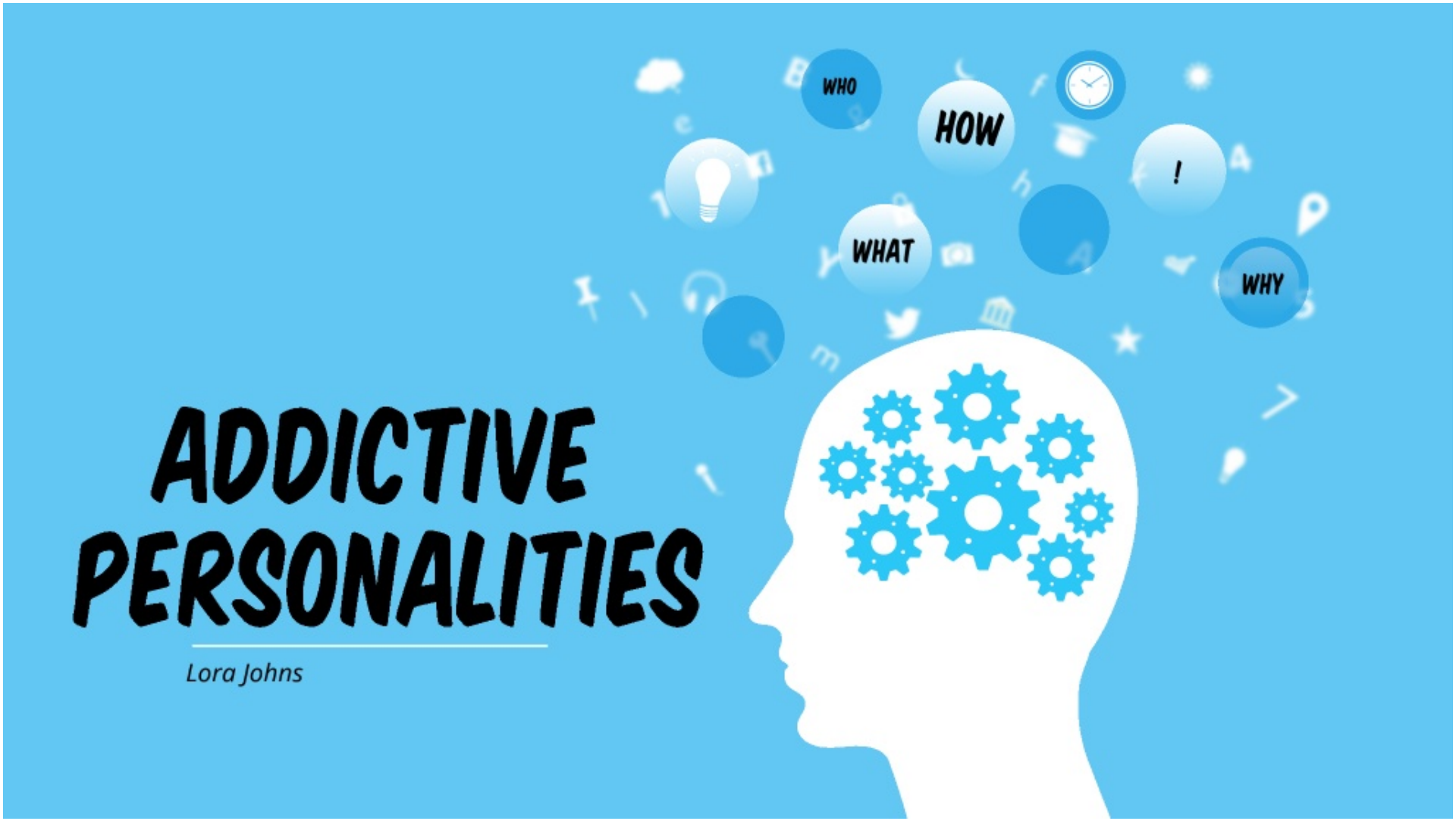

Legalh
Alcohol
Cannabis
Ecstasy
Amphet
Benzos

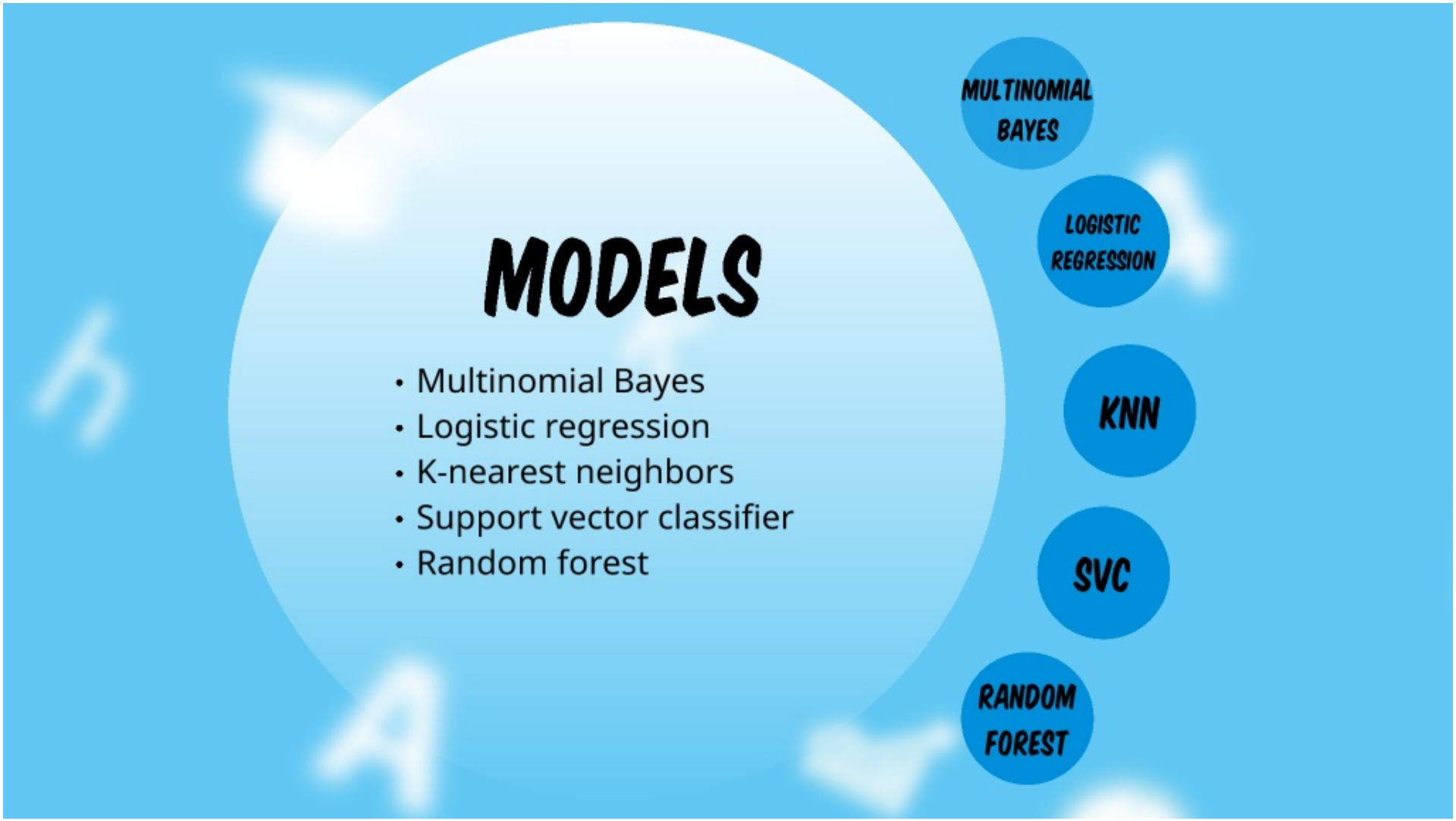

WHO
HOW
!
WHAT
WHY
ADDICTIVE PERSONALITIES
Lora Johns


ANALYSIS
METRICS

PRELIMINARY METRICS AND TESTS

- Chi squared contingency table for predictors indicated statistical significance (test statistic of 36081.0 and p value 1.0)
- Initial multinomial Bayes classification (without feature scaling) produced accuracy of ~ 0.56


WHO
HOW
!
WHAT
WHY
ADDICTIVE PERSONALITIES
Lora Johns

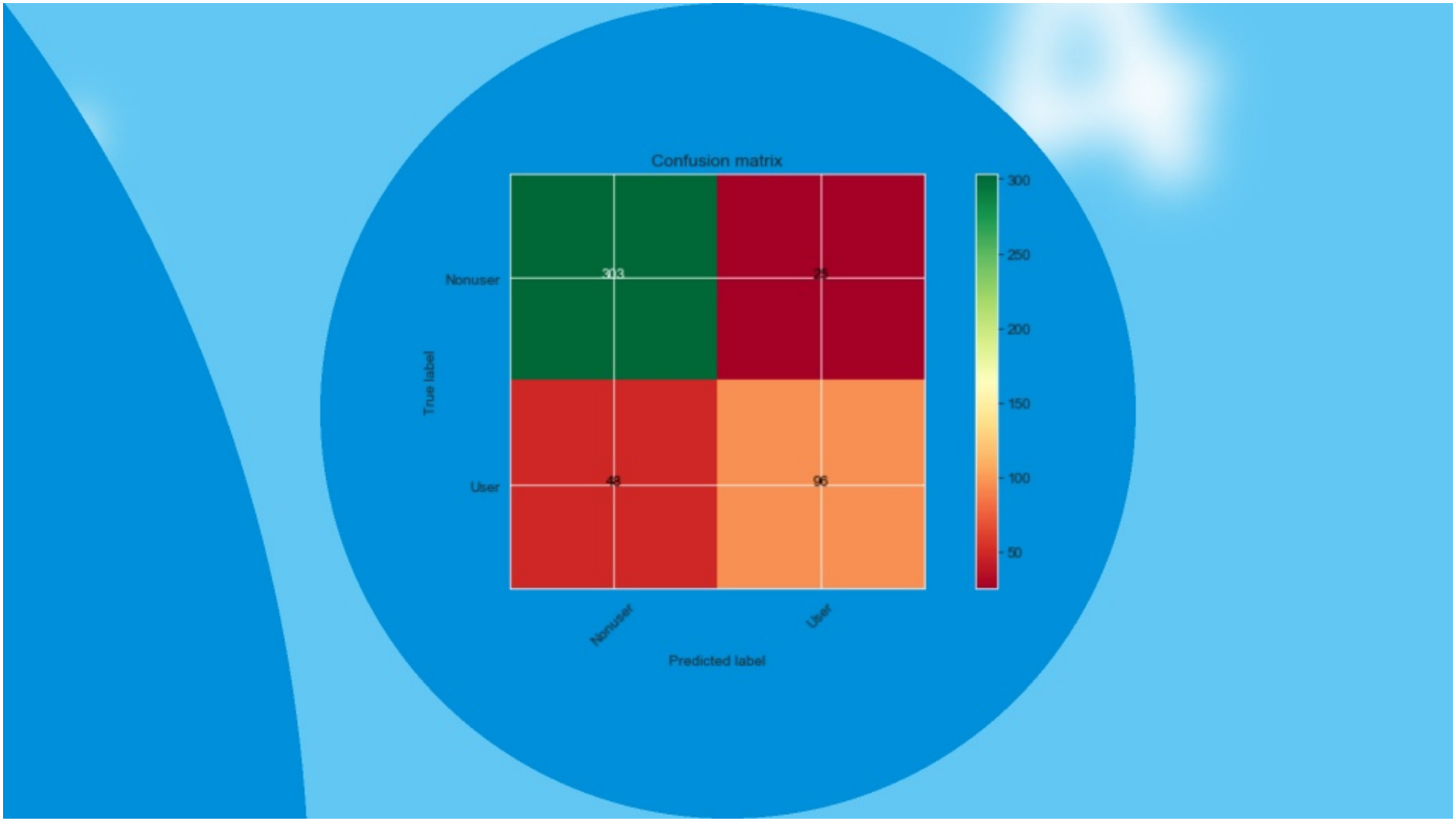

MODELS
• Multinomial Bayes
• Logistic regression
• K-nearest neighbors
• Support vector classifier
• Random forest
MULTINOMIAL BAYES
LOGISTIC REGRESSION
KNN
SVC
RANDOM FOREST

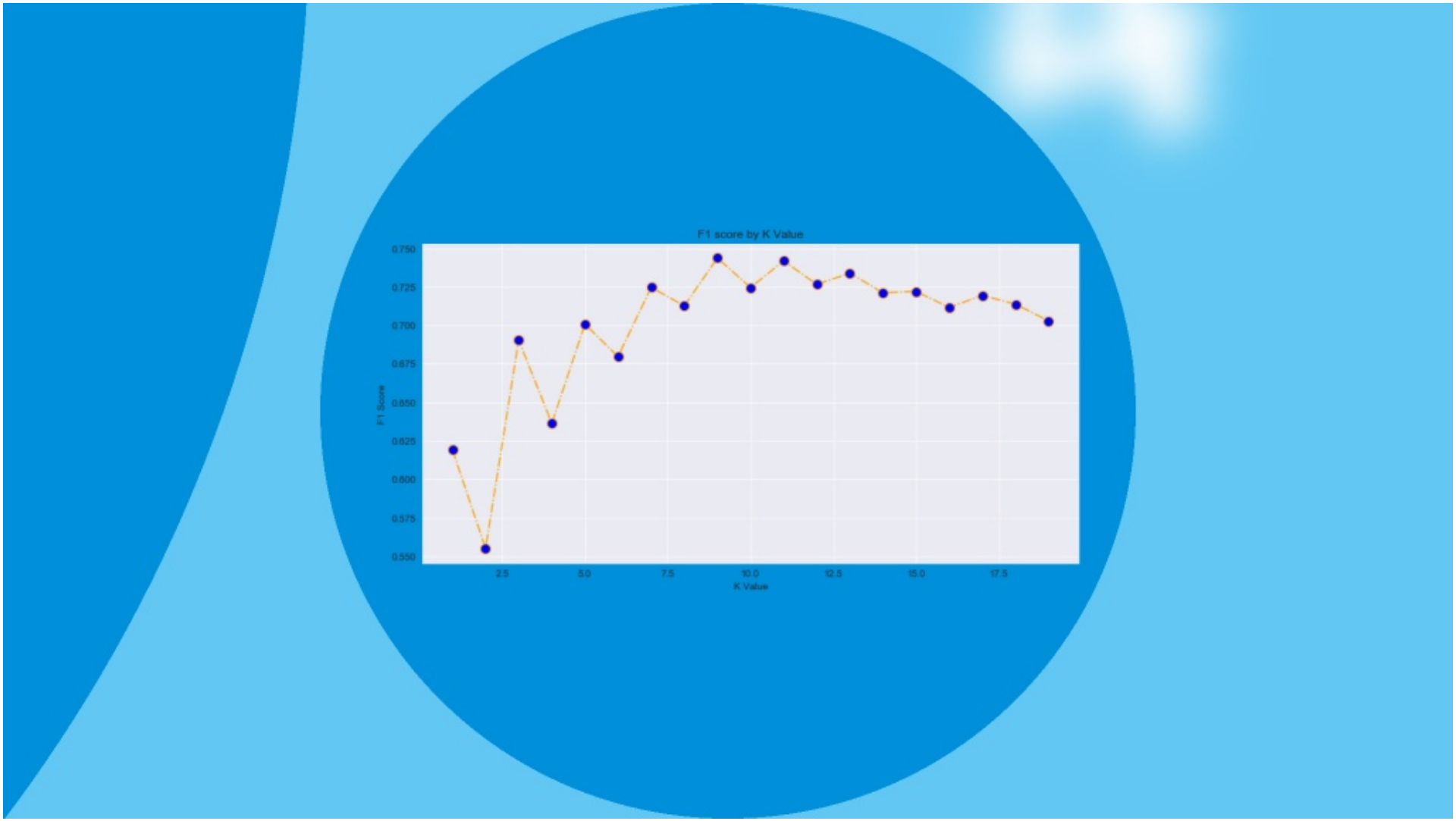
- Accuracy: 0.801
- Cross-validation: stratified k-fold

LOGISTIC REGRESSION

- F1: 0.739
- Accuracy decreased on testing set
- Ergo, this model probably overfit the training data
- More false labels than other models


K-NEAREST NEIGHBORS
ACCURACY: 0.8453
F1: 0.7245
TRAINING RMSE: 0.403
TEST RMSE: 0.393
• Performance improved on test
• Multiple models run to find model with best F1 score
MATRIX
MULTIPLE MODELS


Confusion matrix
Nonuser
User
True label
313
25
48
96
Nonuser
User
Predicted label
300
250
200
150
100
50


F1 score by K Value
F1 Score
0.750
0.725
0.700
0.675
0.650
0.625
0.600
0.575
0.550
2.5
5.0
7.5
10.0
12.5
15.0
17.5
K Value

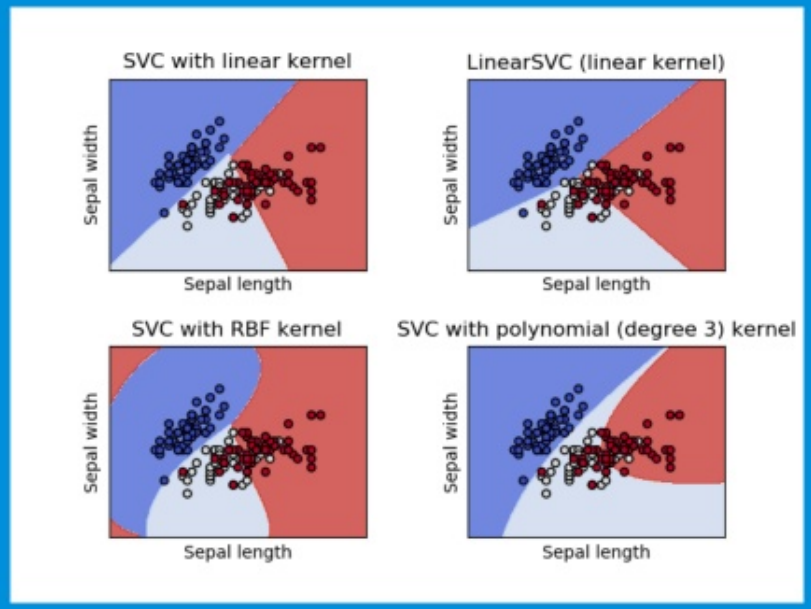
SUPPORT VECTOR MACHINE

ACCURACY: 0.831

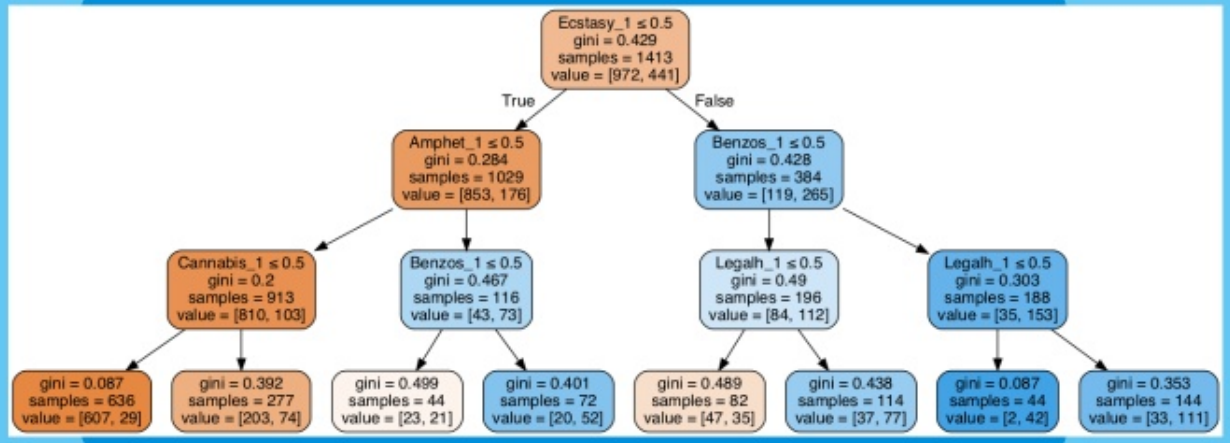
CONFUSION MATRIX: [285, 43]
[37, 107]

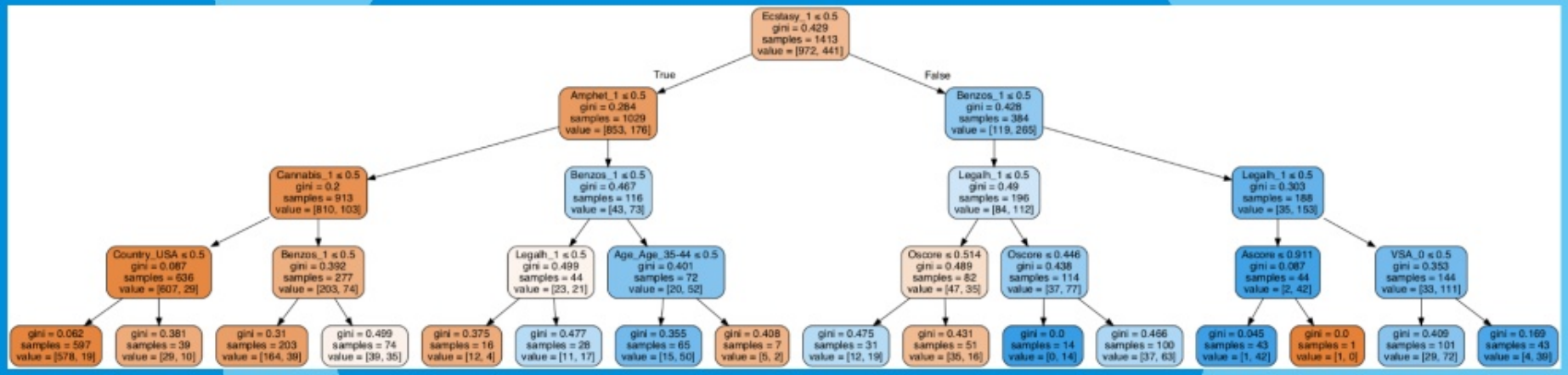
- Tuned hyperparameters with grid search
- Few false negatives (which, if we care about intervention, is good); many false positives (which is problematic in this context)

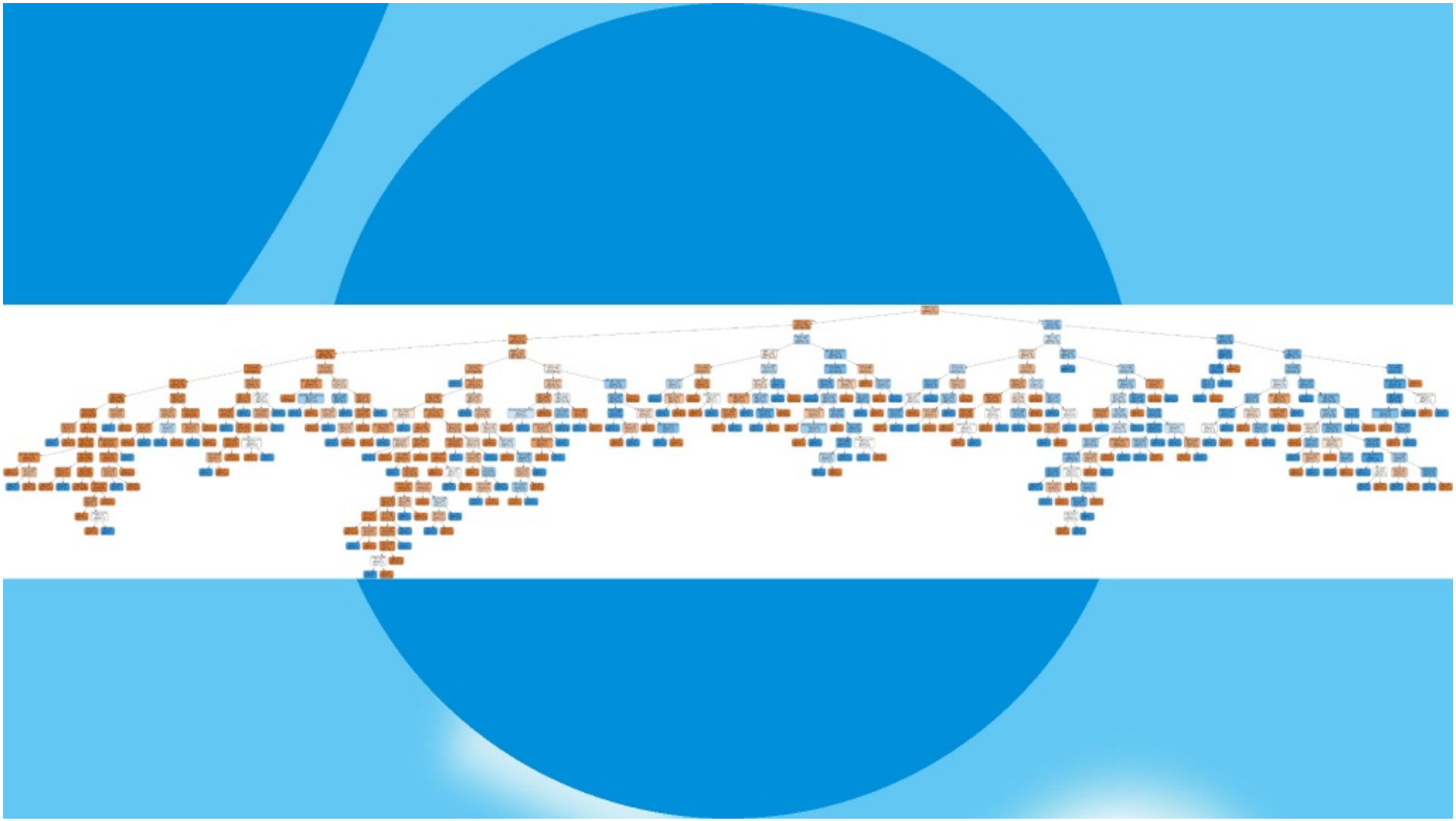
SVM

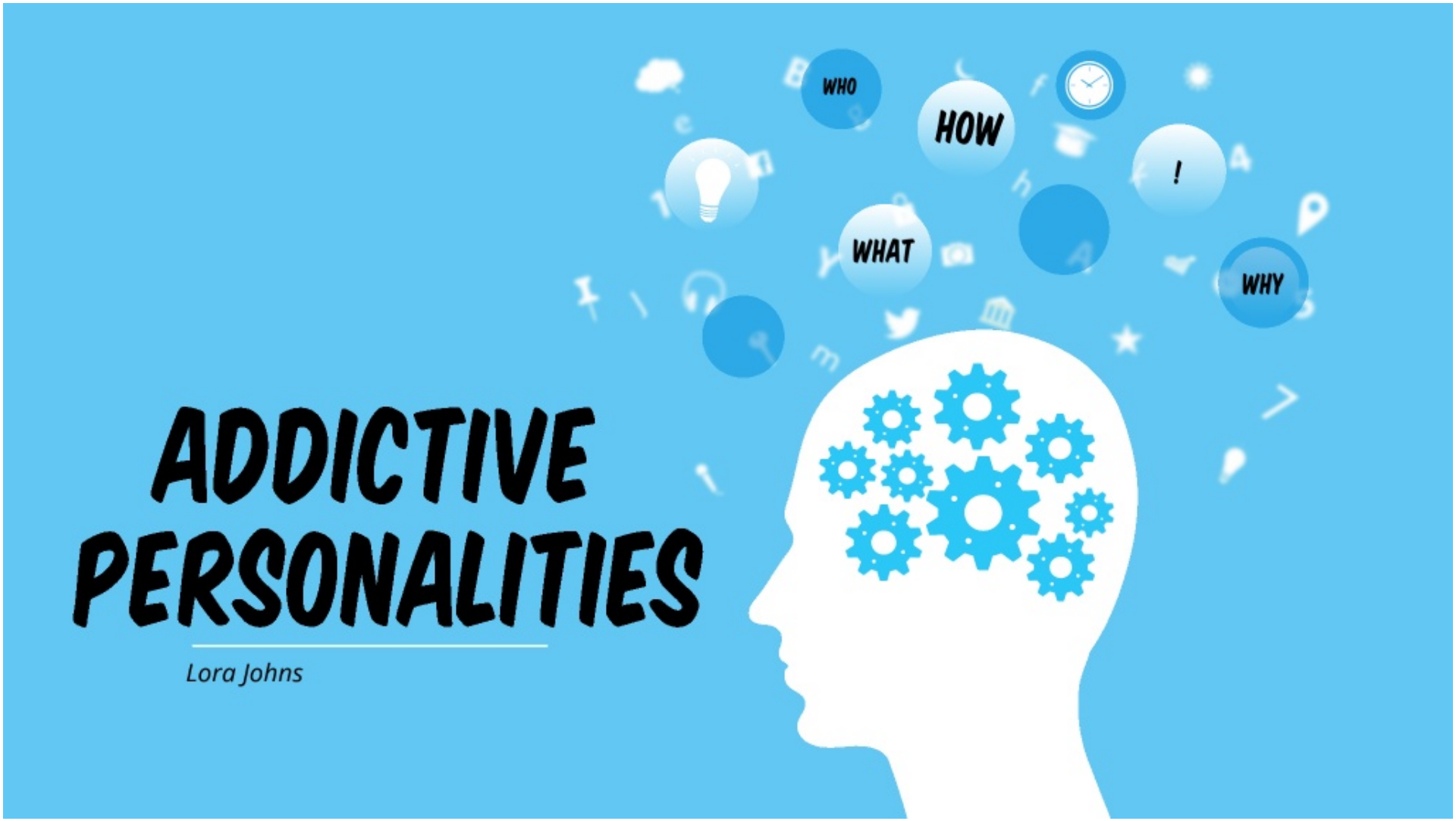

SVC with linear kernel
LinearSVC (linear kernel)
SVC with RBF kernel
SVC with polynomial (degree 3) kernel
Sepal width
Sepal length

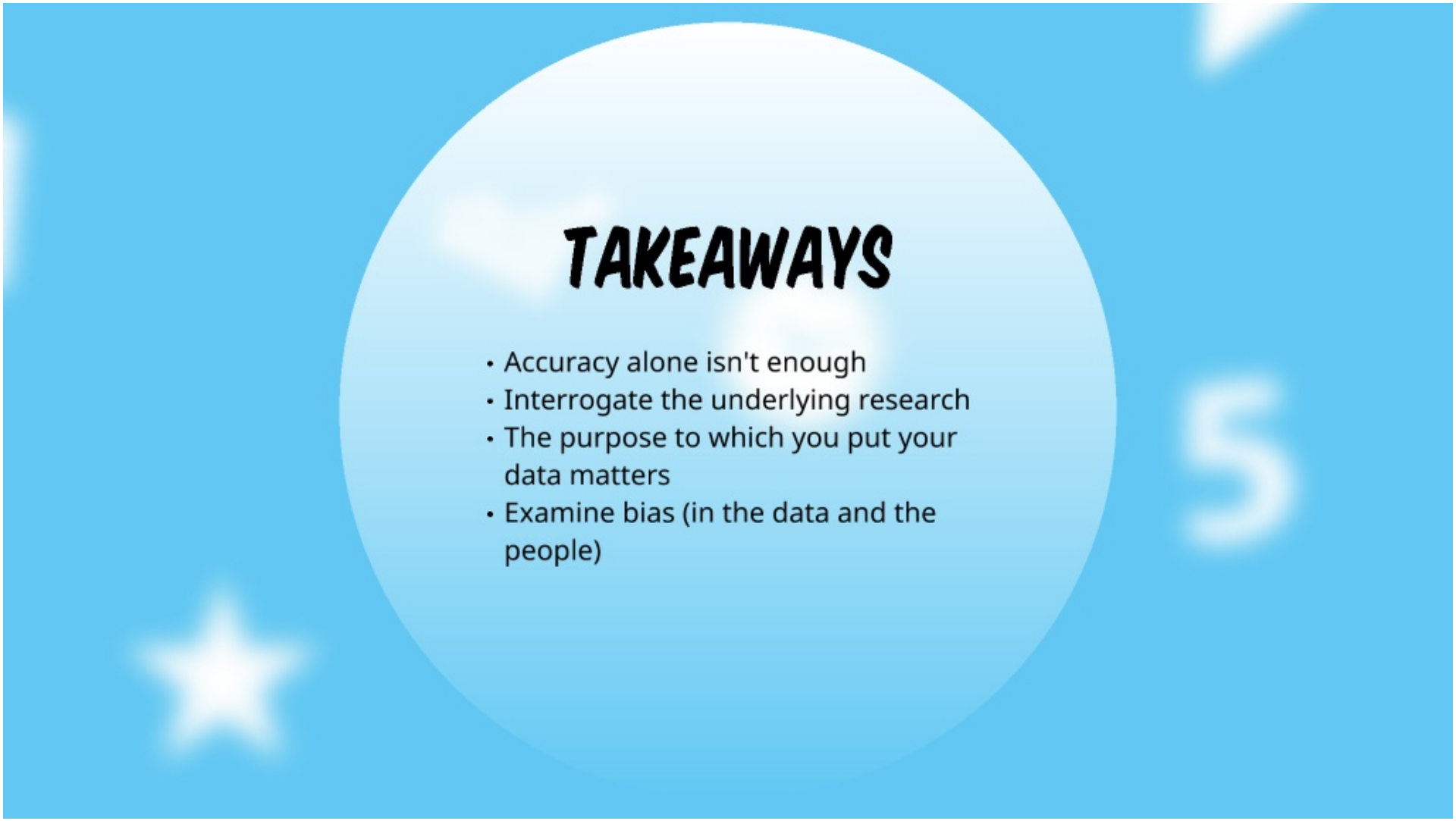

RANDOM FOREST DECISION TREES
ACCURACY: 0.970
AVERAGE ERROR: 0.03
• In all models, ecstasy was the top discriminator (followed by benzos and legal highs)
• Tuned hyperparameters with randomized search
• Performance improved from training to testing (RMSE dropped from 0.190 to 0.03)
TREE
TREE
TREE

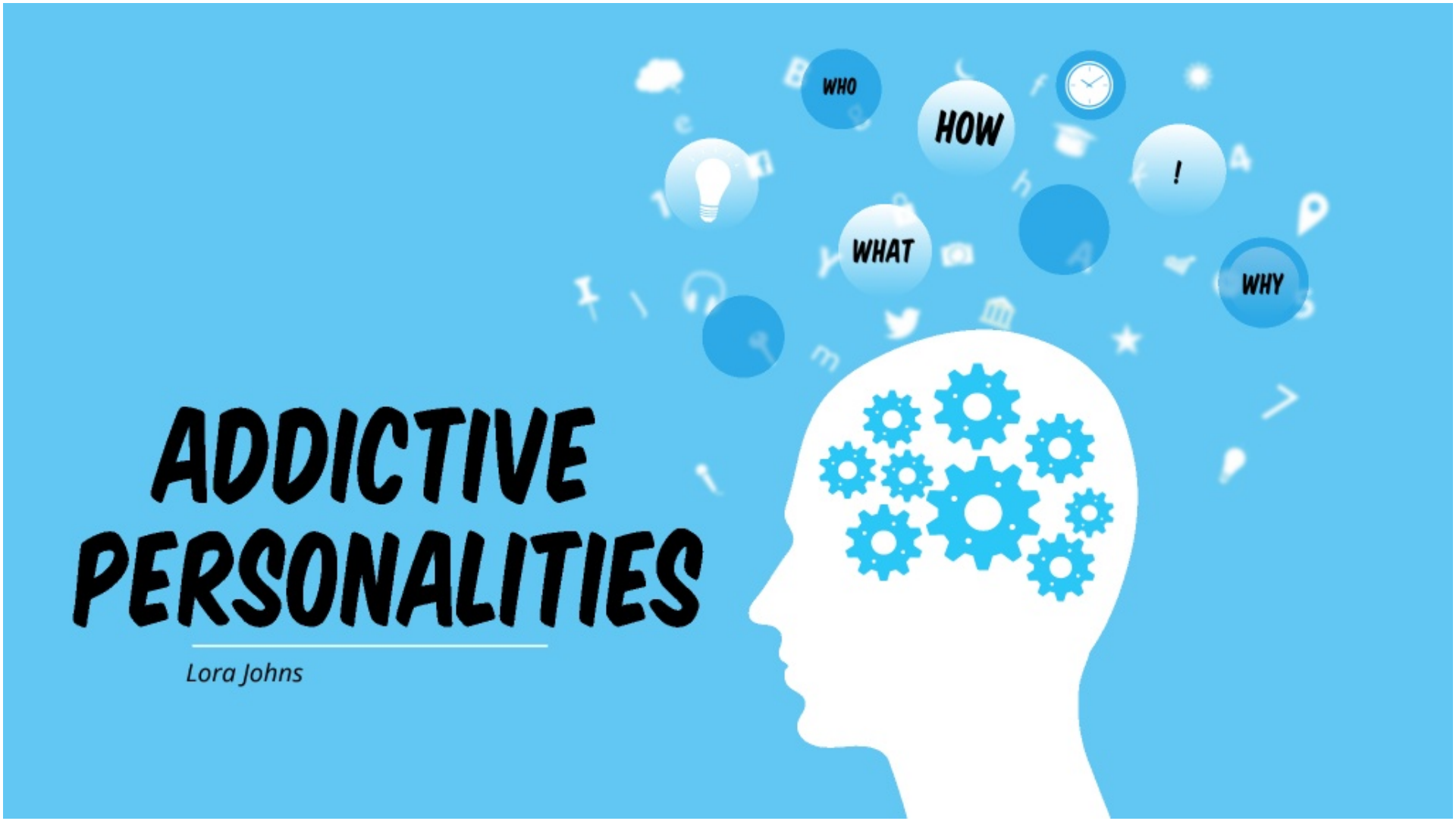

Ecstasy_1 ≤ 0.5
gini = 0.429
samples = 1413
value = [972, 441]
True
False
Amphet_1 ≤ 0.5
gini = 0.284
samples = 1029
value = [853, 176]
Benzos_1 ≤ 0.5
gini = 0.428
samples = 384
value = [119, 265]
Cannabis_1 ≤ 0.5
gini = 0.2
samples = 913
value = [810, 103]
Benzos_1 ≤ 0.5
gini = 0.467
samples = 116
value = [43, 73]
Legalh_1 ≤ 0.5
gini = 0.49
samples = 196
value = [84, 112]
Legalh_1 ≤ 0.5
gini = 0.303
samples = 188
value = [35, 153]
gini = 0.087
samples = 636
value = [607, 29]
gini = 0.392
samples = 277
value = [203, 74]
gini = 0.499
samples = 44
value = [23, 21]
gini = 0.401
samples = 72
value = [20, 52]
gini = 0.489
samples = 82
value = [47, 35]
gini = 0.438
samples = 114
value = [37, 77]
gini = 0.087
samples = 44
value = [2, 42]
gini = 0.353
samples = 144
value = [33, 111]


Ecstasy_1 ≤ 0.5
gini = 0.429
samples = 1413
value = [972, 441]
True
False
Amphet_1 ≤ 0.5
gini = 0.284
samples = 1029
value = [853, 176]
Benzos_1 ≤ 0.5
gini = 0.428
samples = 384
value = [119, 265]
Cannabis_1 ≤ 0.5
gini = 0.2
samples = 913
value = [810, 103]
Benzos_1 ≤ 0.5
gini = 0.467
samples = 116
value = [43, 73]
Legalh_1 ≤ 0.5
gini = 0.49
samples = 196
value = [84, 112]
Legalh_1 ≤ 0.5
gini = 0.303
samples = 188
value = [35, 153]
Country_USA ≤ 0.5
gini = 0.087
samples = 636
value = [607, 29]
Benzos_1 ≤ 0.5
gini = 0.392
samples = 277
value = [203, 74]
Legalh_1 ≤ 0.5
gini = 0.499
samples = 44
value = [23, 21]
Age_Age_35-44 ≤ 0.5
gini = 0.401
samples = 72
value = [20, 52]
Oscore ≤ 0.514
gini = 0.489
samples = 82
value = [47, 35]
Oscore ≤ 0.446
gini = 0.438
samples = 114
value = [37, 77]
Ascore ≤ 0.911
gini = 0.087
samples = 44
value = [2, 42]
VSA_0 ≤ 0.5
gini = 0.353
samples = 144
value = [33, 111]
gini = 0.062
samples = 597
value = [578, 19]
gini = 0.381
samples = 39
value = [29, 10]
gini = 0.31
samples = 203
value = [164, 39]
gini = 0.499
samples = 74
value = [39, 35]
gini = 0.375
samples = 16
value = [12, 4]
gini = 0.477
samples = 28
value = [11, 17]
gini = 0.355
samples = 65
value = [15, 50]
gini = 0.408
samples = 7
value = [5, 2]
gini = 0.475
samples = 31
value = [12, 19]
gini = 0.431
samples = 51
value = [35, 16]
gini = 0.0
samples = 14
value = [0, 14]
gini = 0.466
samples = 100
value = [37, 63]
gini = 0.045
samples = 43
value = [1, 42]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.409
samples = 101
value = [29, 72]
gini = 0.169
samples = 43
value = [4, 39]




WHO
HOW
!
WHAT
WHY
ADDICTIVE PERSONALITIES
Lora Johns


TAKEAWAYS
• Accuracy alone isn't enough
• Interrogate the underlying research
• The purpose to which you put your data matters
• Examine bias (in the data and the people)


WHO
HOW
!
WHAT
WHY
ADDICTIVE PERSONALITIES
Lora Johns