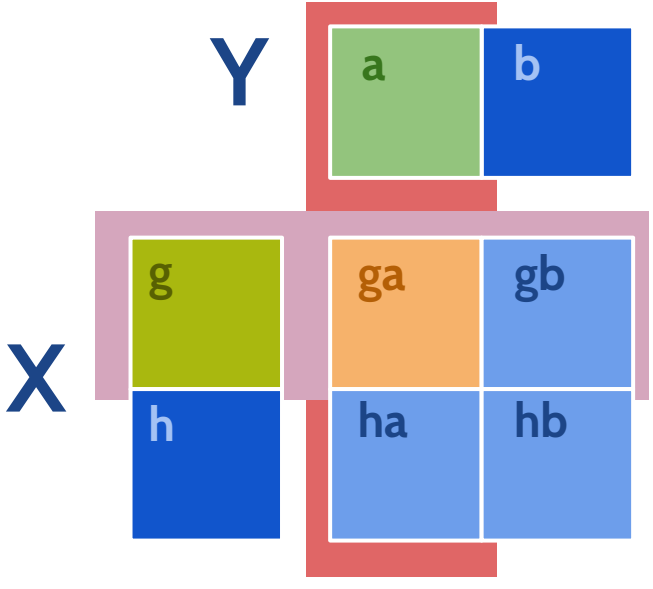
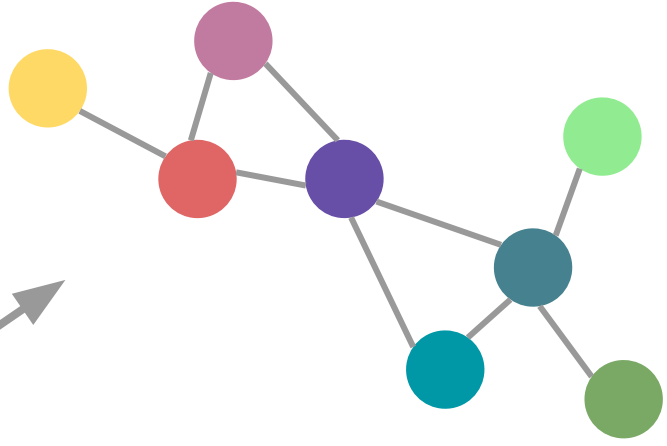
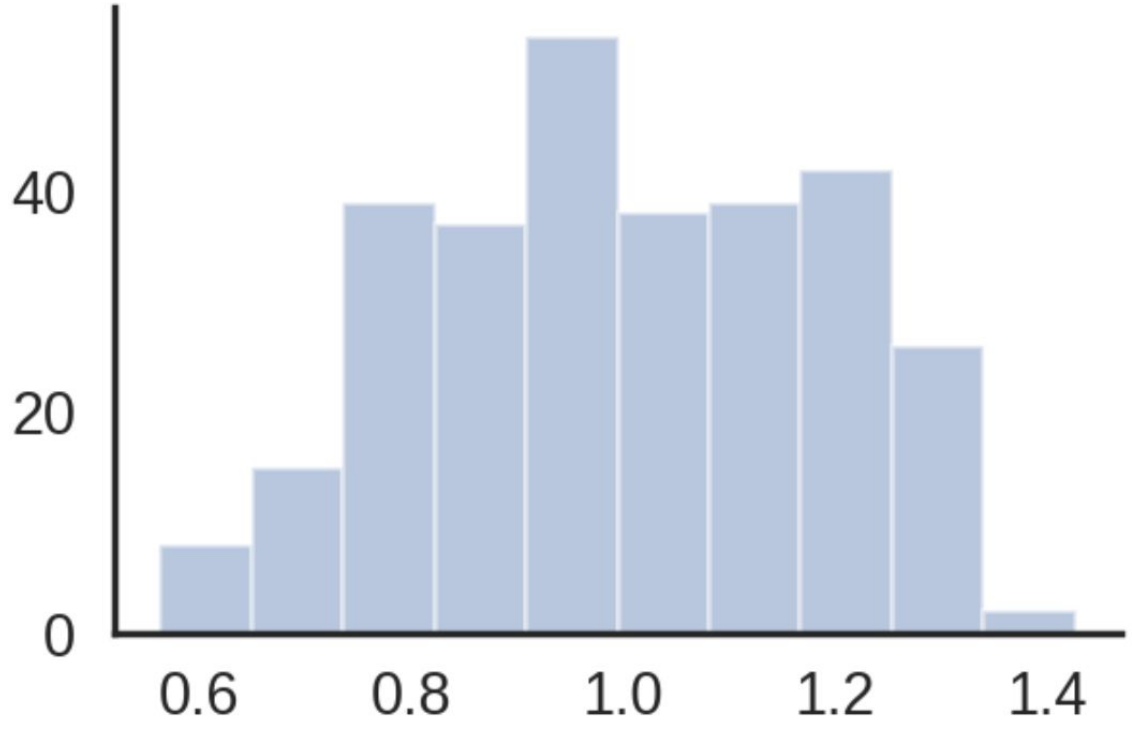
# Exploratory Data Visualization for High-dimensional Data



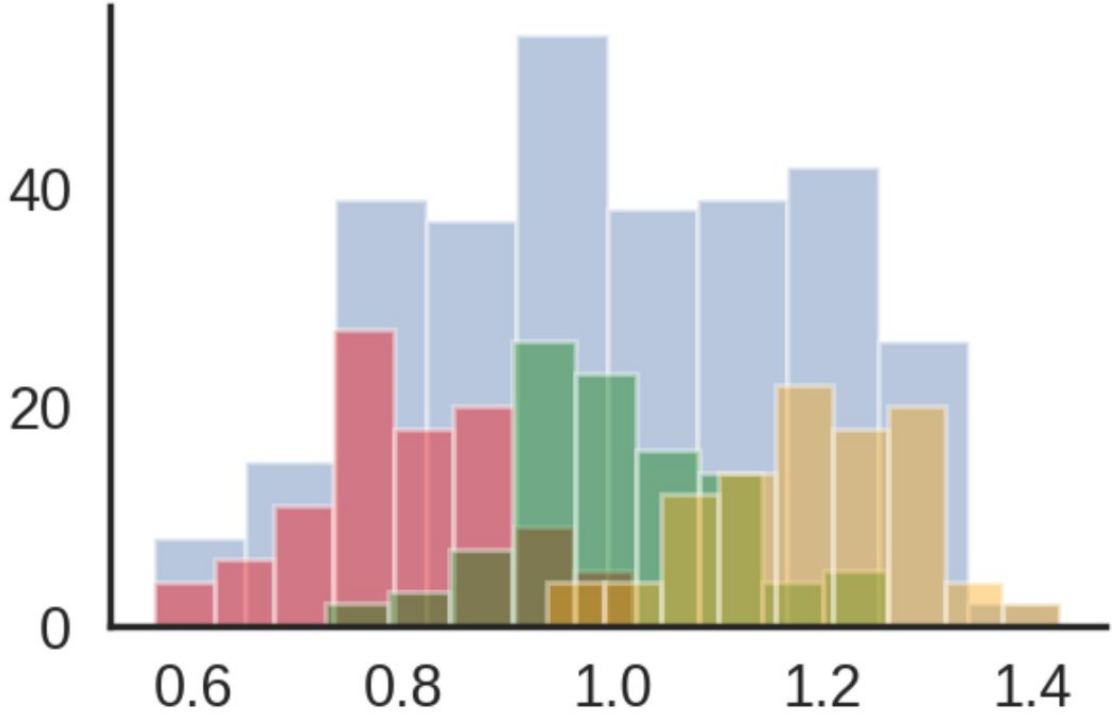



Jane Adams
Data Visualization Artist
Vermont Complex Systems Center
Tuesday March 30th 2021

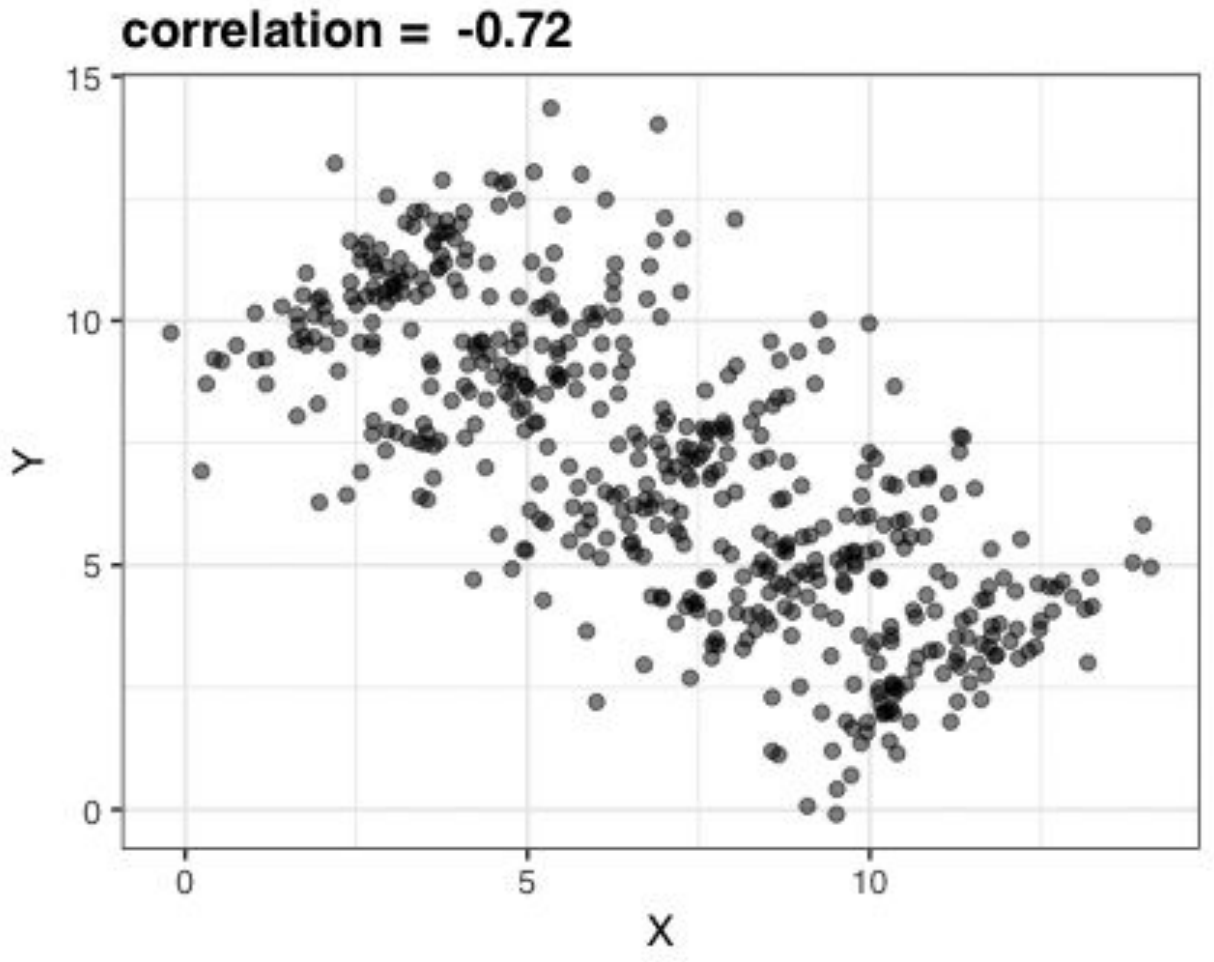
# Hidden stories...



# Hidden stories...



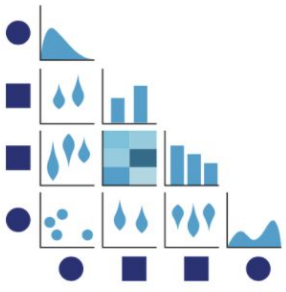
# Simpson's Paradox



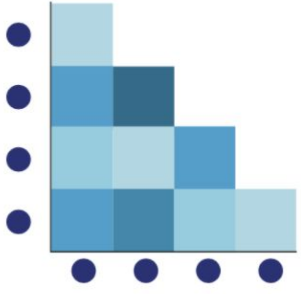

src: simplystatistics.org

# Sirius: Exploratory Analysis Python Package

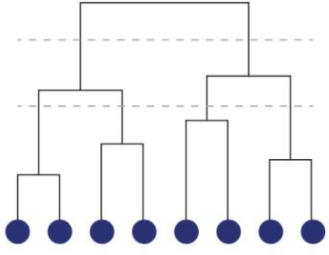




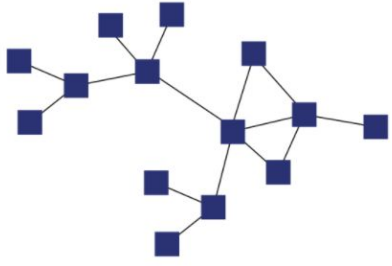
(A) Small Multiples



(B) Matrix

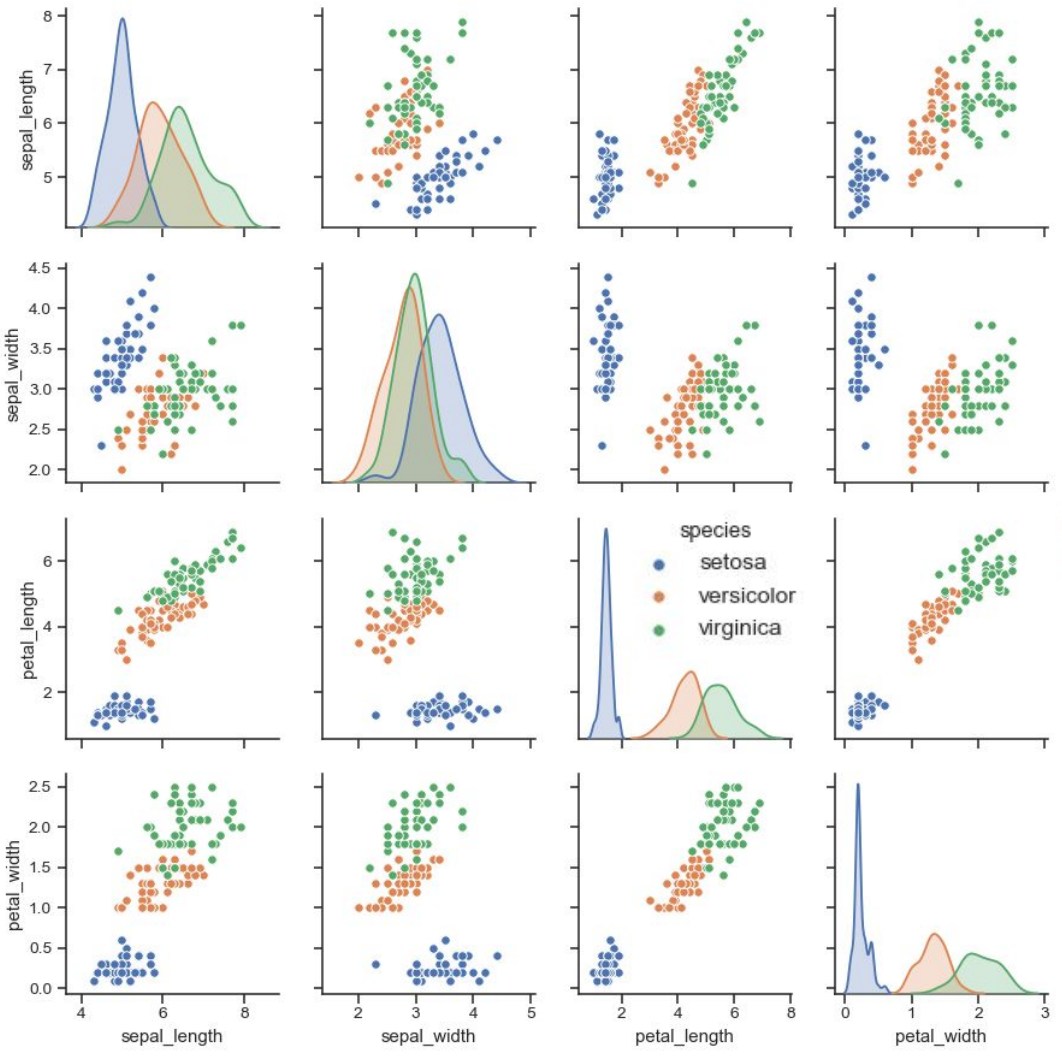


(C) Dendrogram



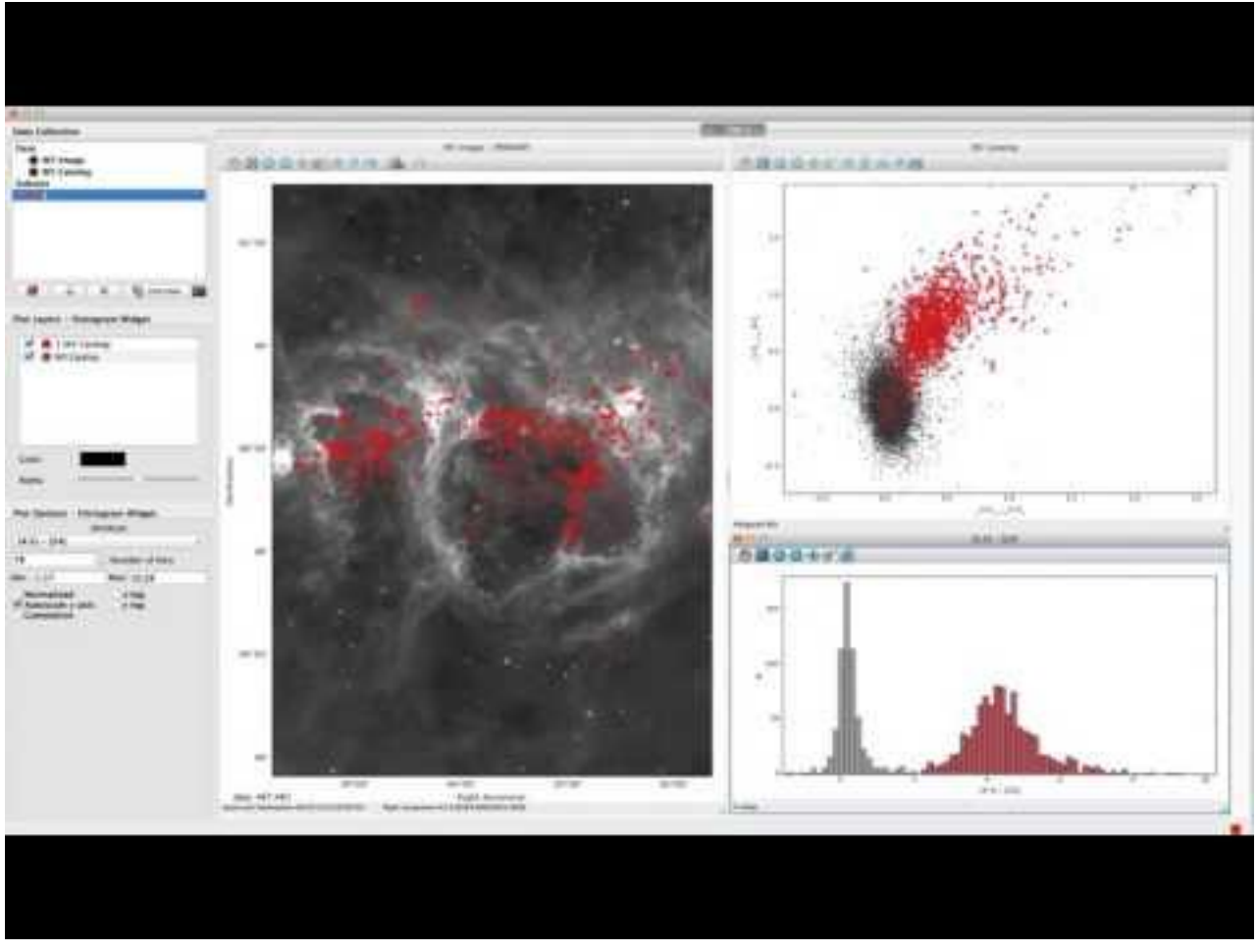
(D) Network

# Pair Plots

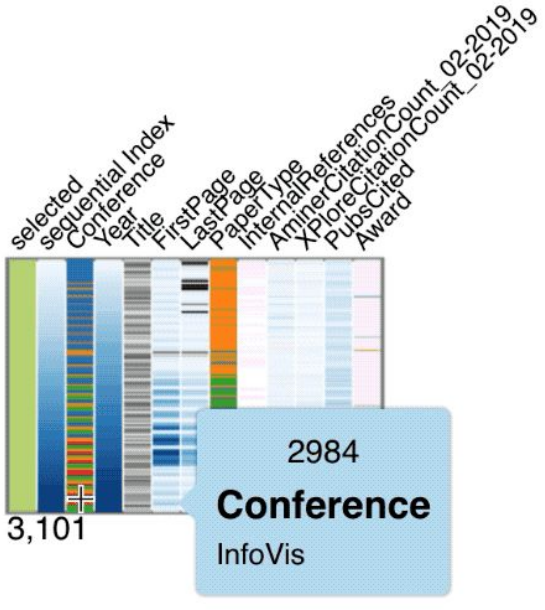



src: seaborn.pydata.org

# Brushing and Linking: GlueViz



# Navio: https://navio.dev/



# Comparing Navio to a Scatterplot Matrix

## Scatterplot matrix 370 rows, 7 attributes

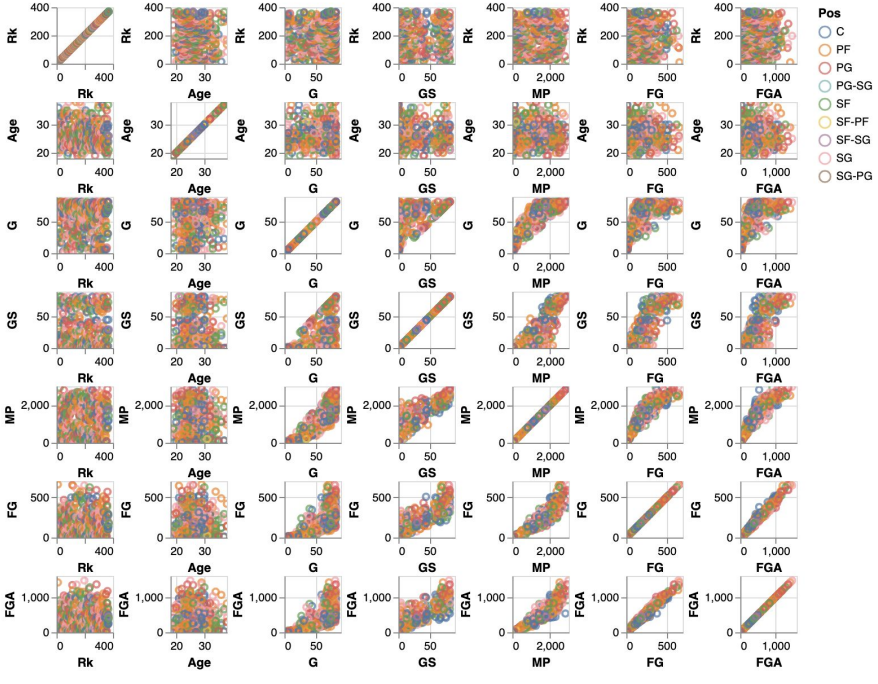
Unfortunately you can only show quantitative values on a scatterplot matrix, and we have to restrict the number of rows and attributes to show for avoiding breaking the notebook. Play with the parameters below

370 — *How many rows to show in the matrix*

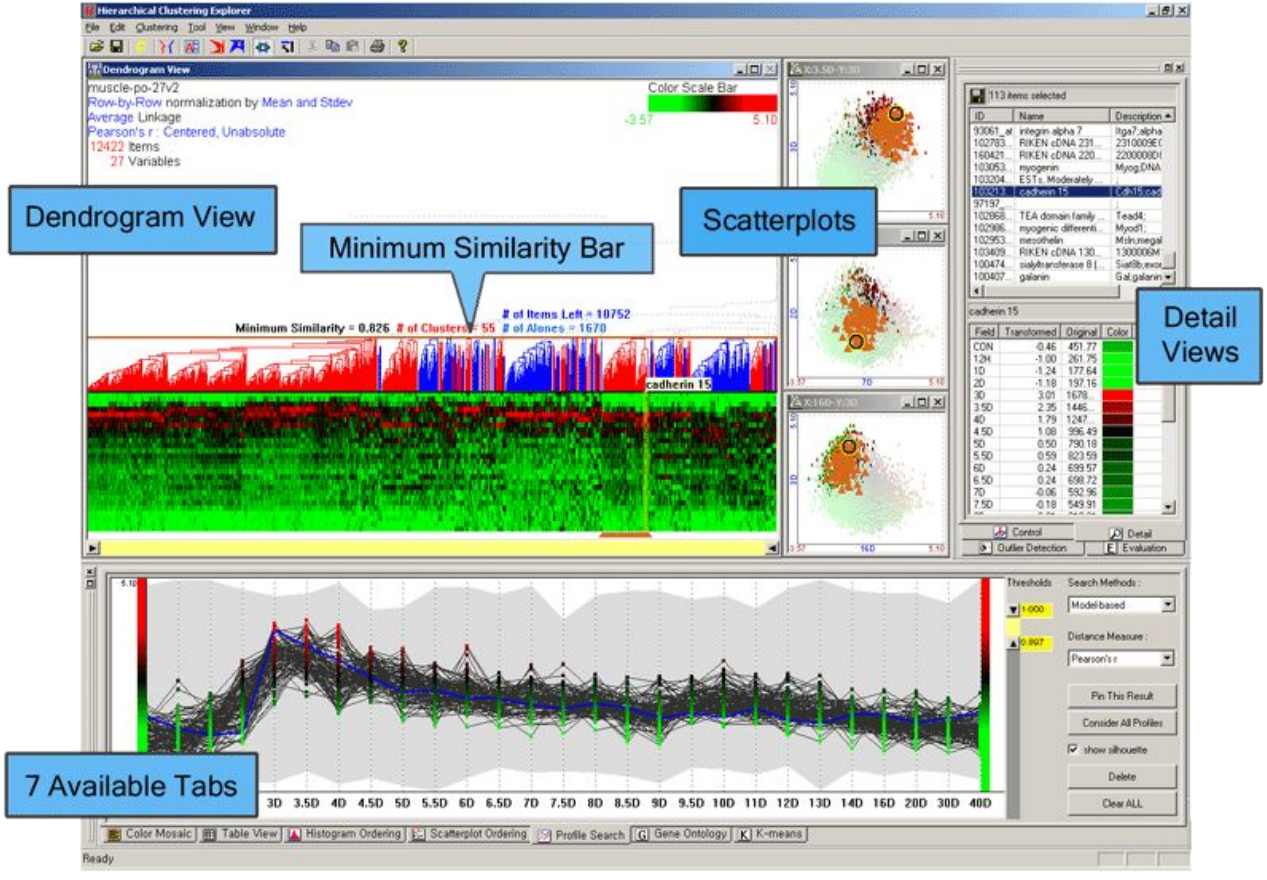
7 — *How many attributes to include in the matrix*

**color_attrib**: Pos

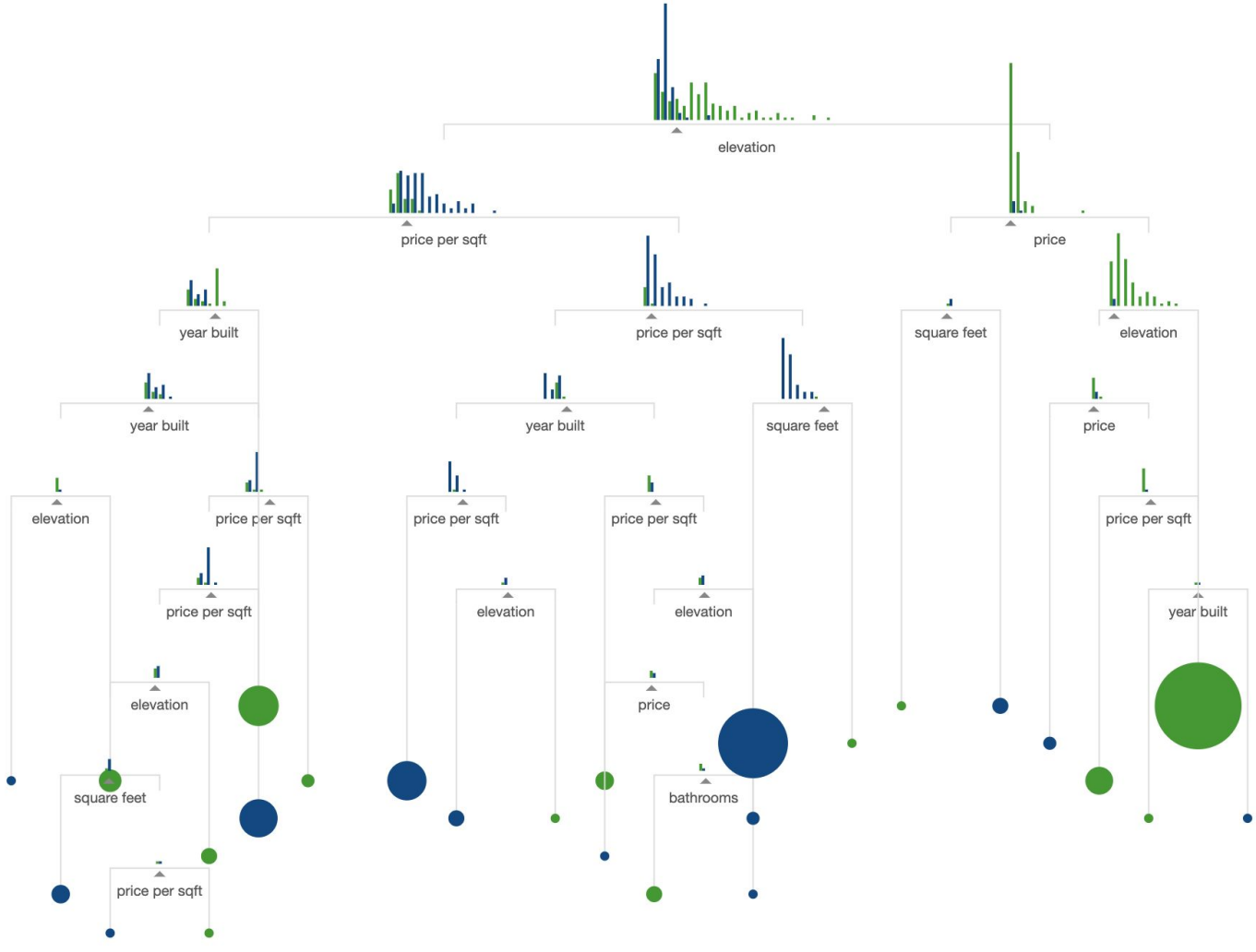
**color_type**: nominal



# Hierarchical Clustering Explorer (HCE)

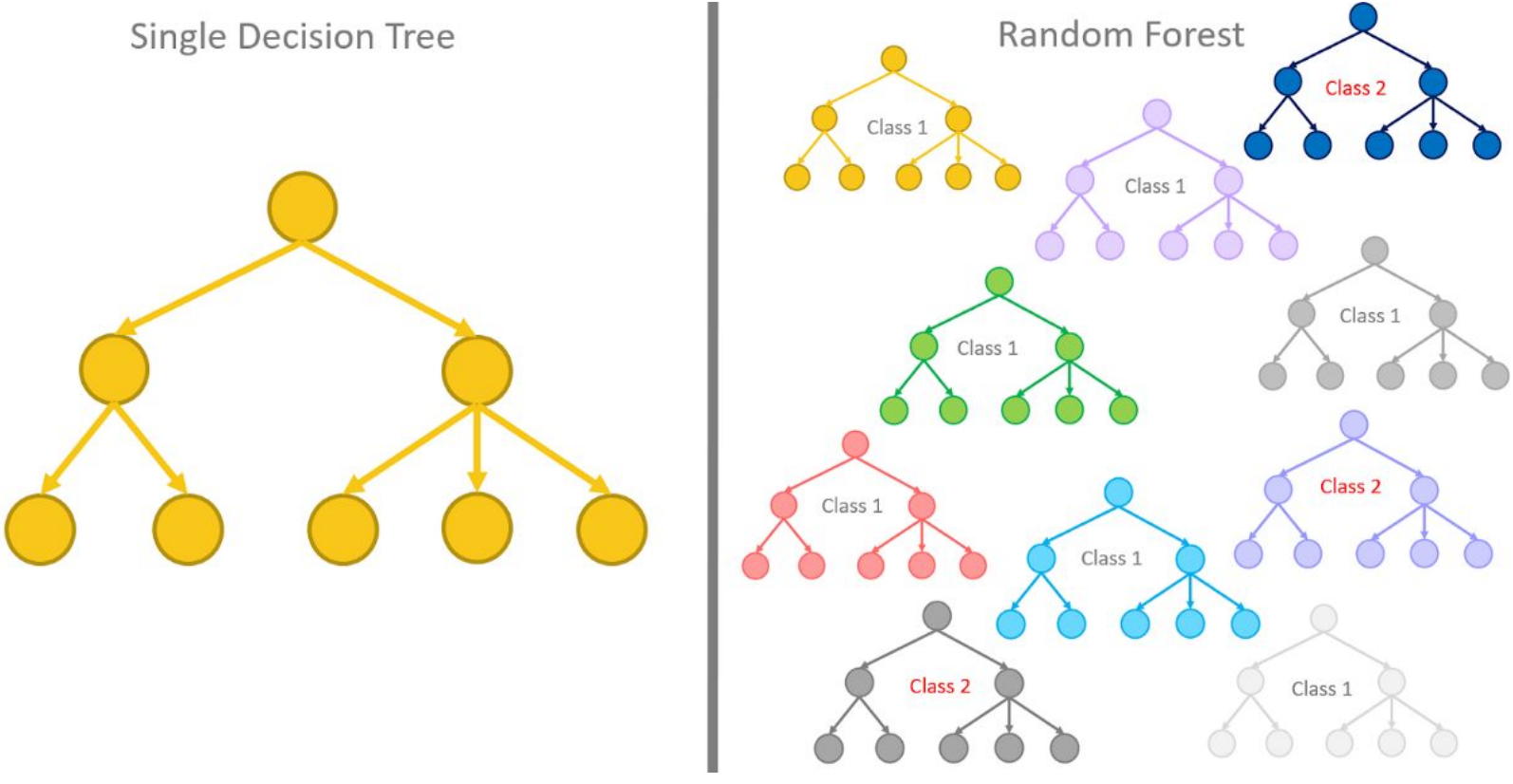


# Decision Trees



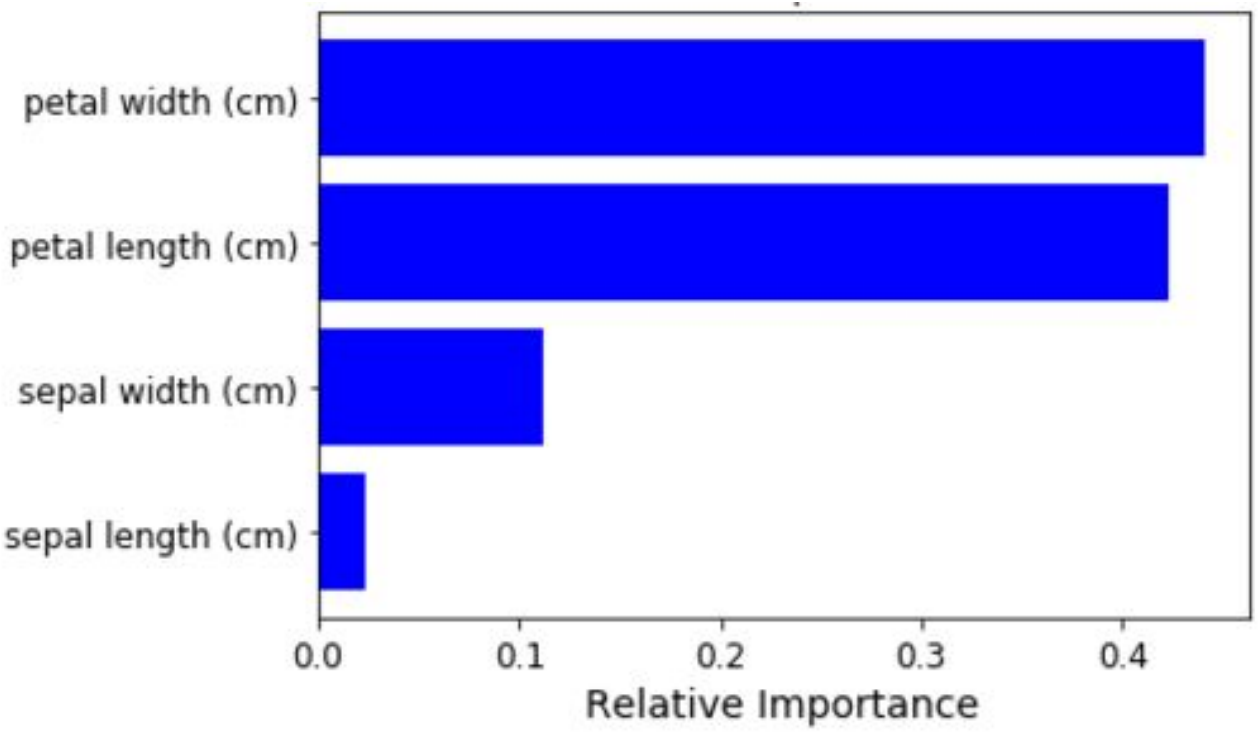

src: r2d3.us

# Random Forests



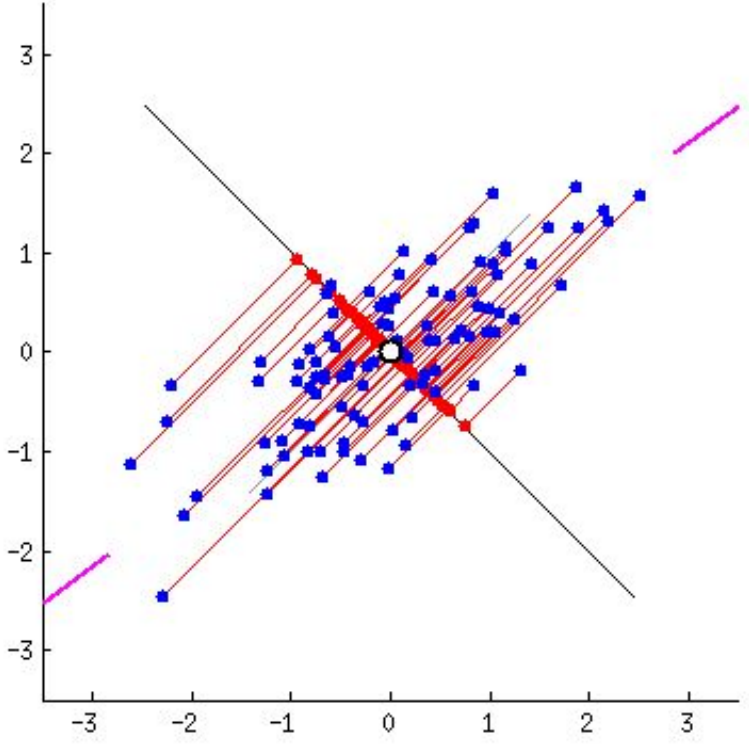

src: [Rosaria Silipo and Kathrin Melcher]

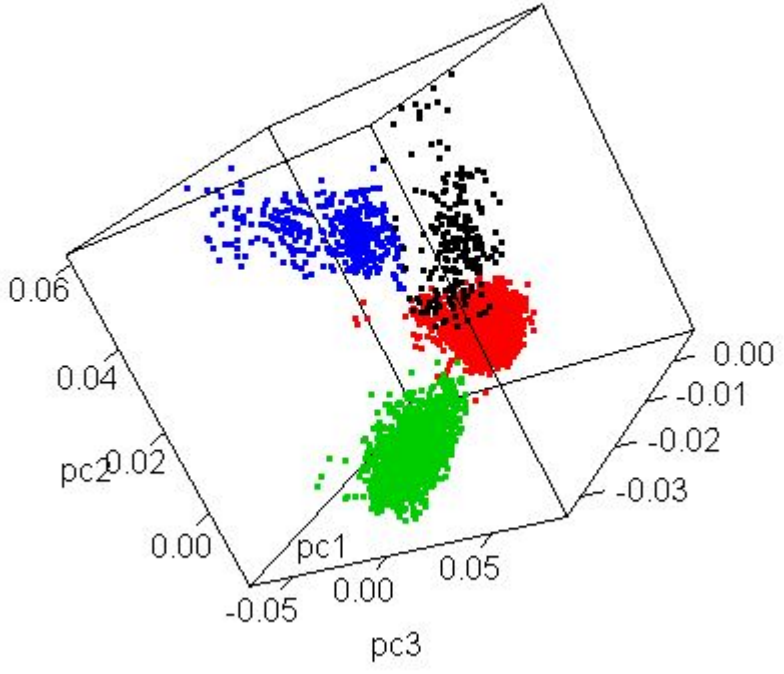
# Feature Importances



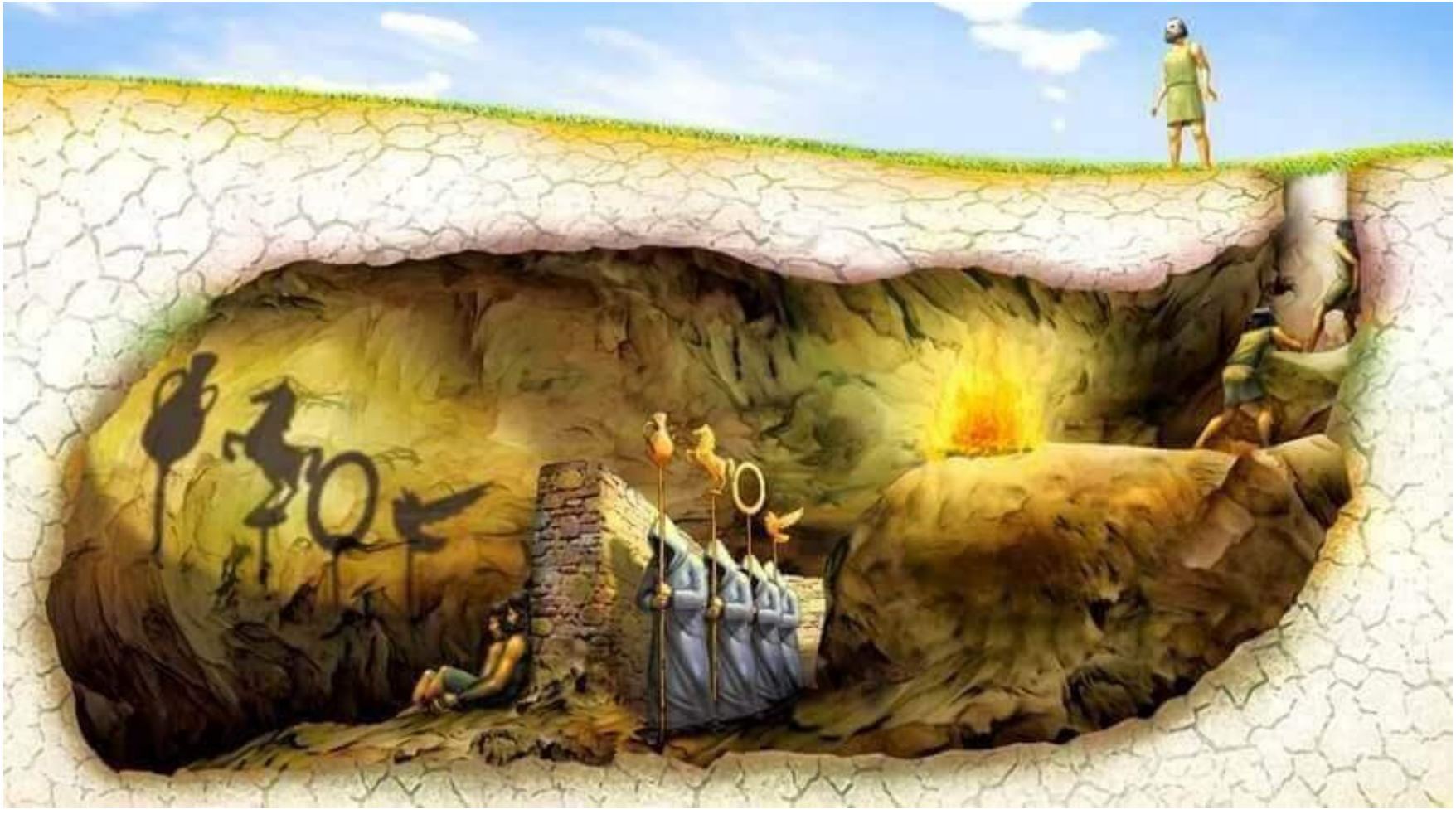

src: [Stack Overflow]

# Principal Component Analysis



# Plato's Cave

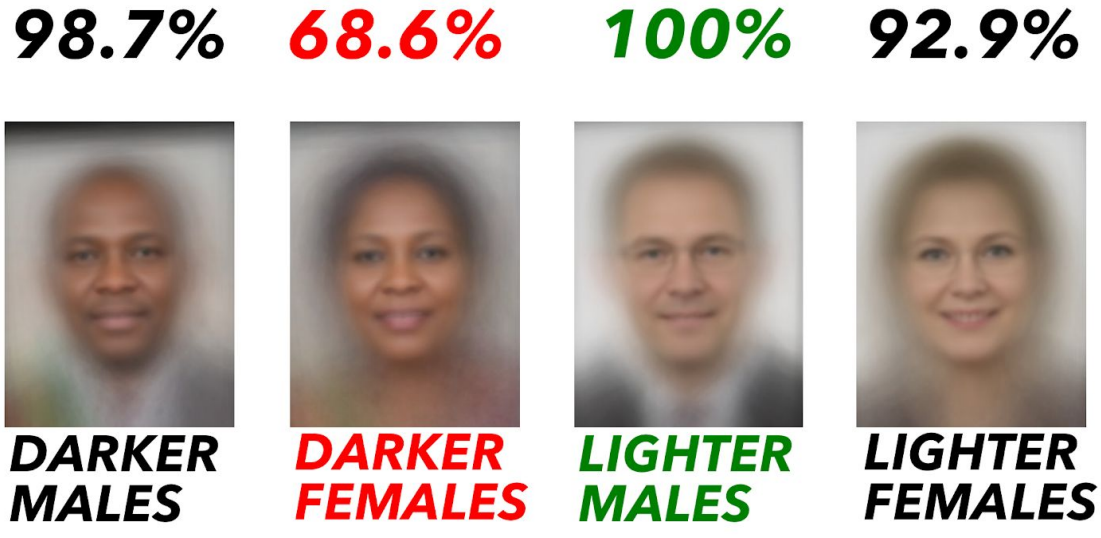


# Plato’s Cave: Modern Version

# Algorithmic Bias

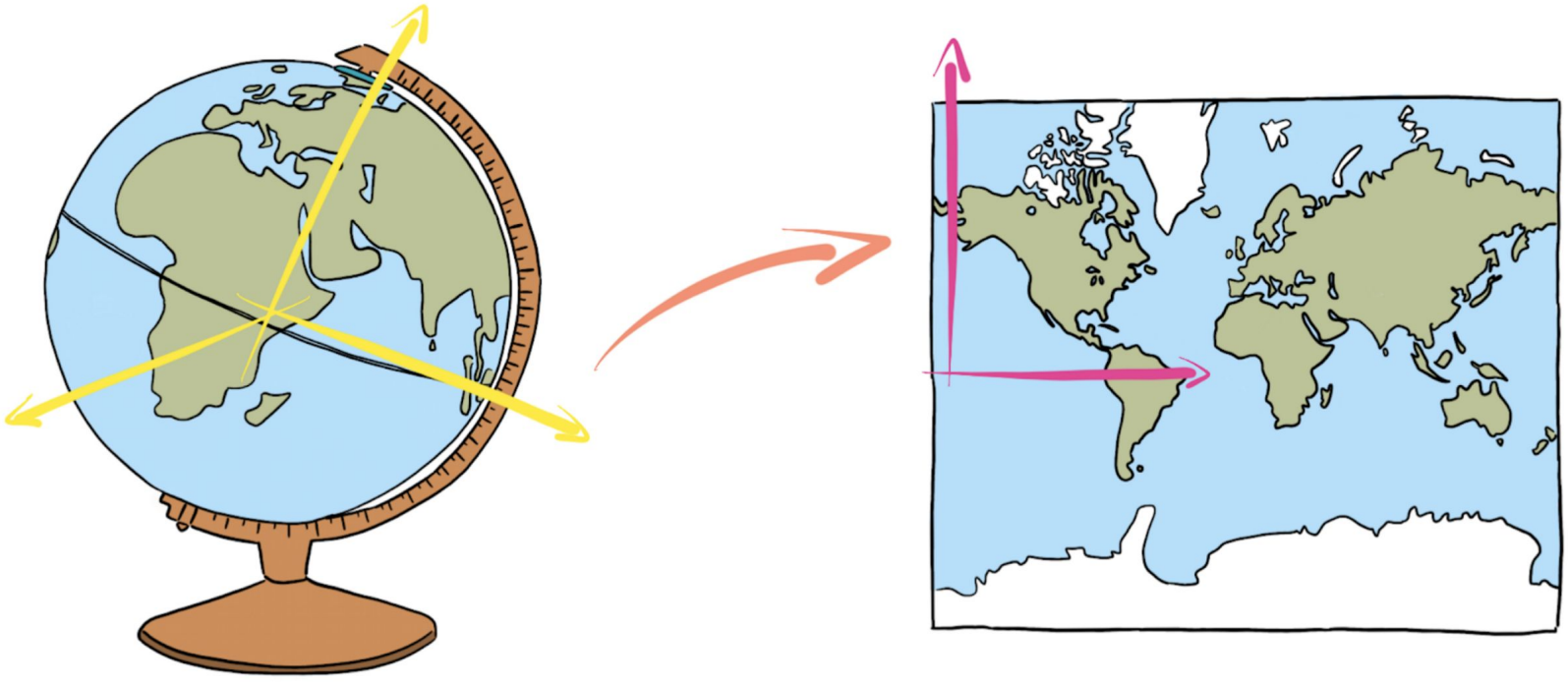
amazon

Rekognition
Performance on
Gender Classification



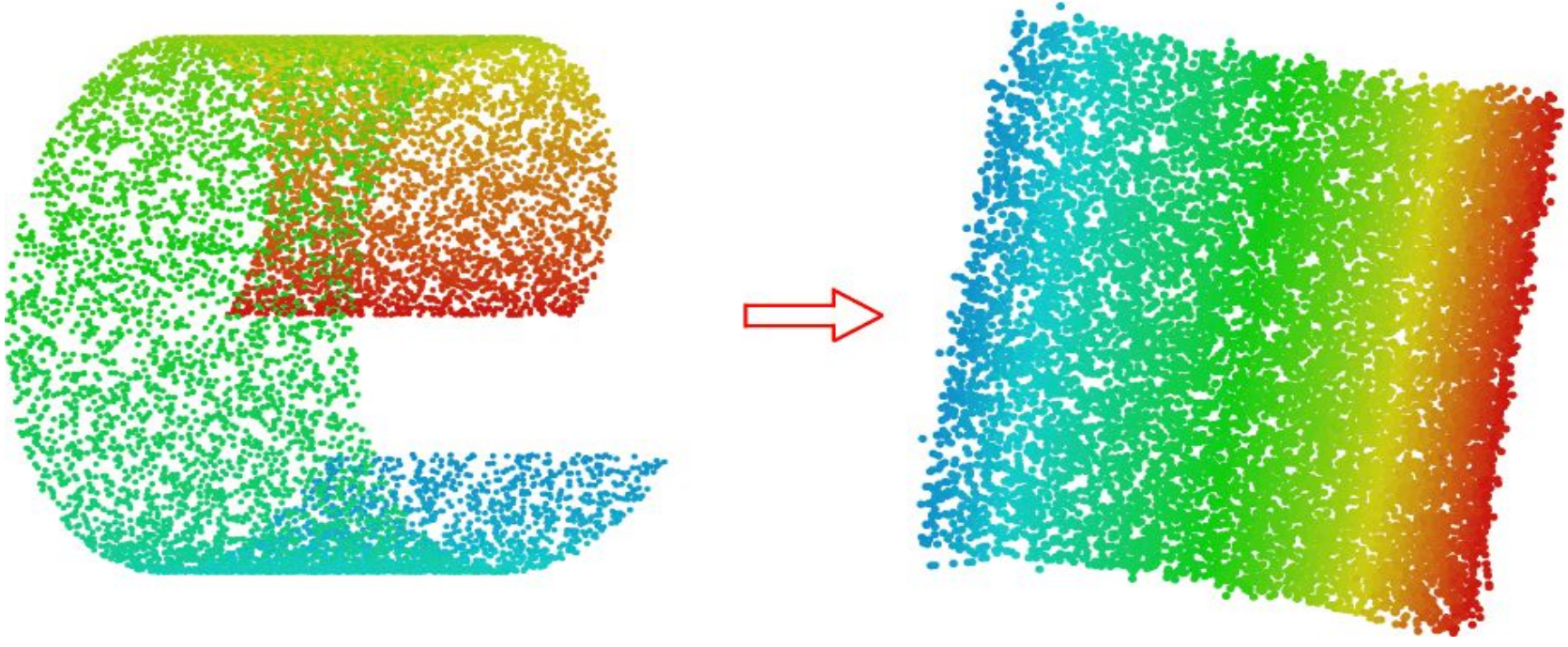

src: Joy Buolamwini

# The curse of dimensionality...



src: Fanny Kasapian, “How to Build a Non-Geographical Map”

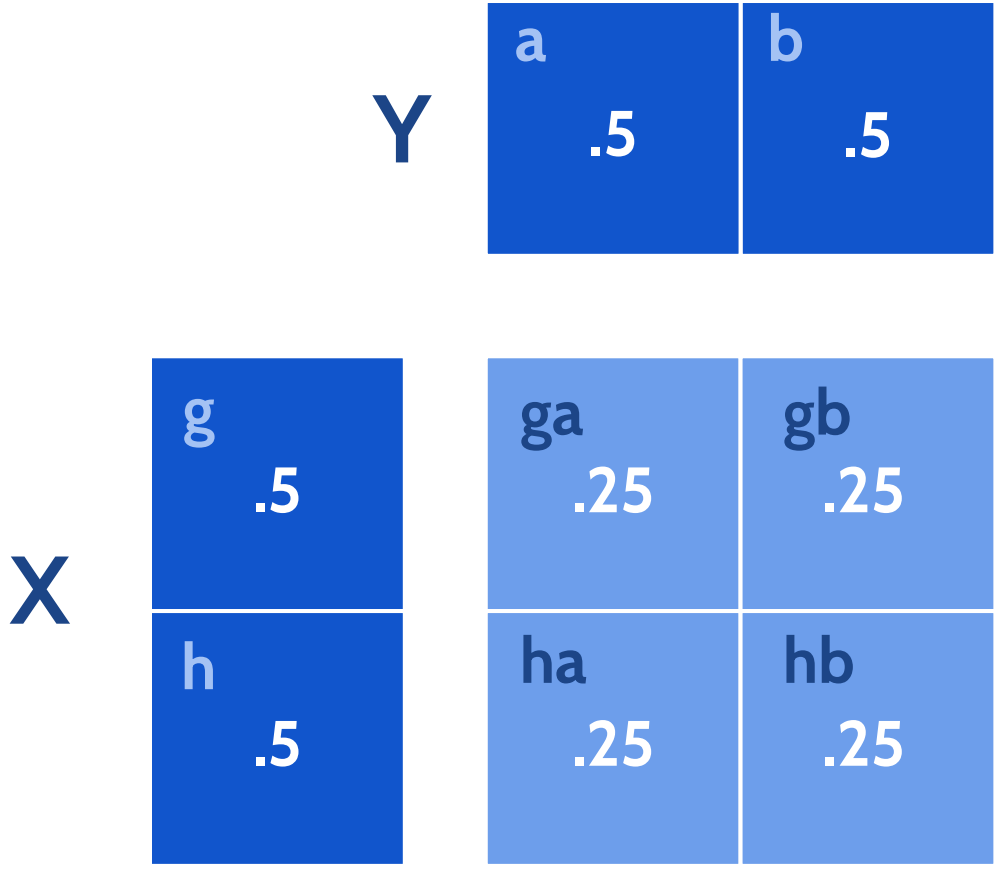
# Locally Linear Mapping



Source: “Nonlinear Dimensionality Reduction via Path-Based Isometric Mapping”, Amir Najafi et. al.
https://www.computer.org/csdl/journal/tp/2016/07/07293680/13rRUwfZC1K

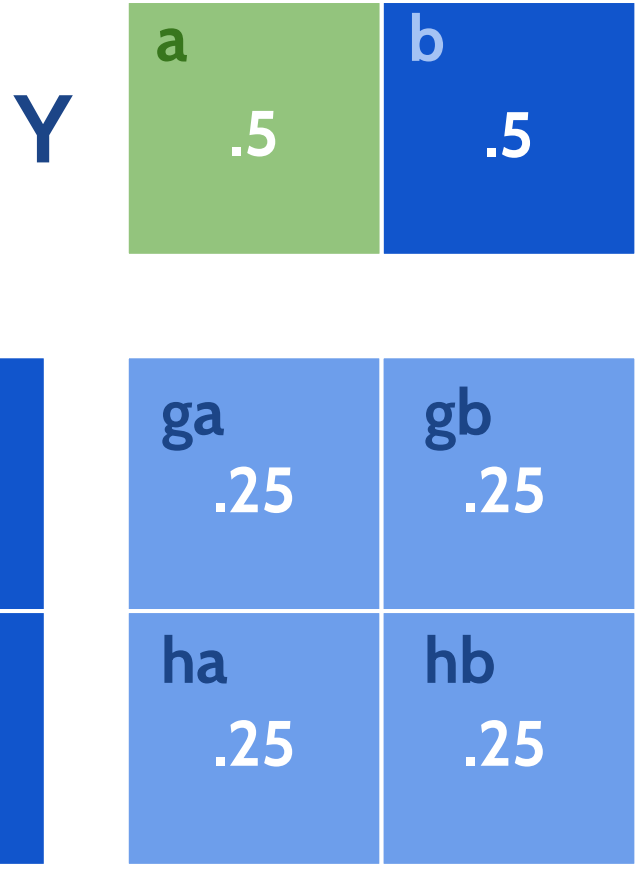
**How do we narrow the exploration space to combinations of variables that are ‘most interesting’?**

# Probability



# Probability

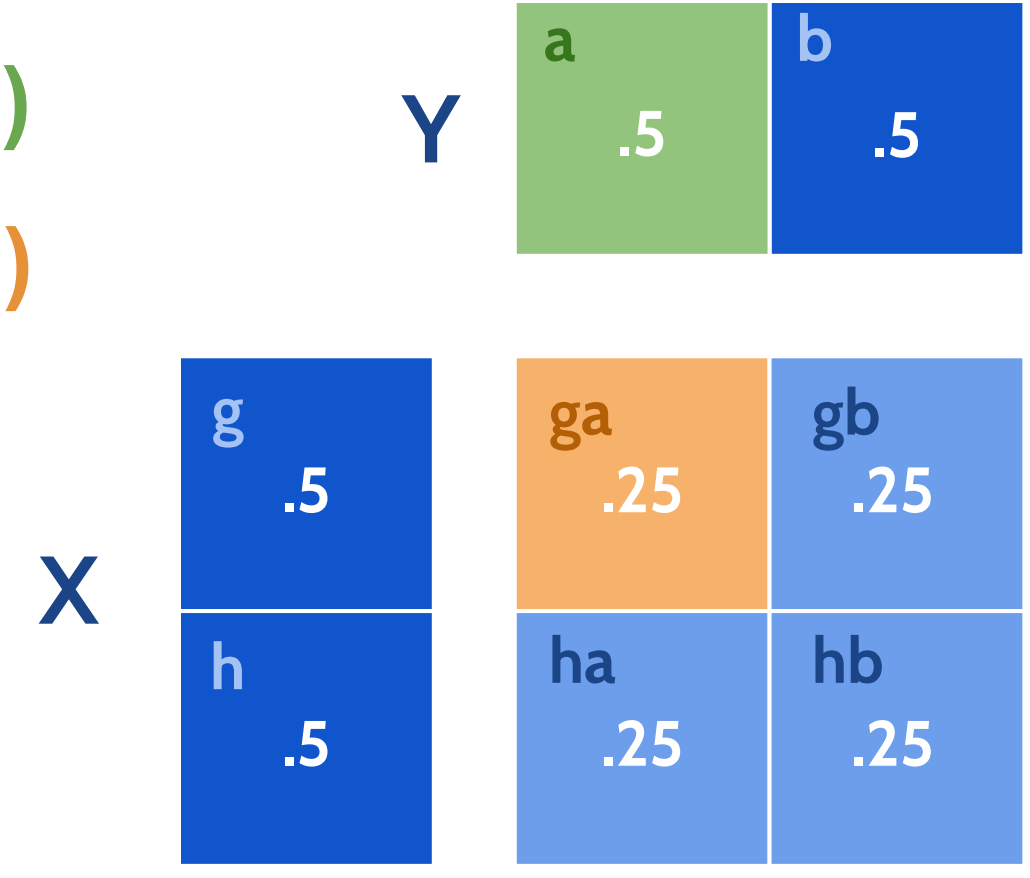
## Marginal: P( Ya )



# Probability

Marginal: P( Ya )

Joint: P( Xg , Ya )



# Groceries: Association Networks of Binary Features

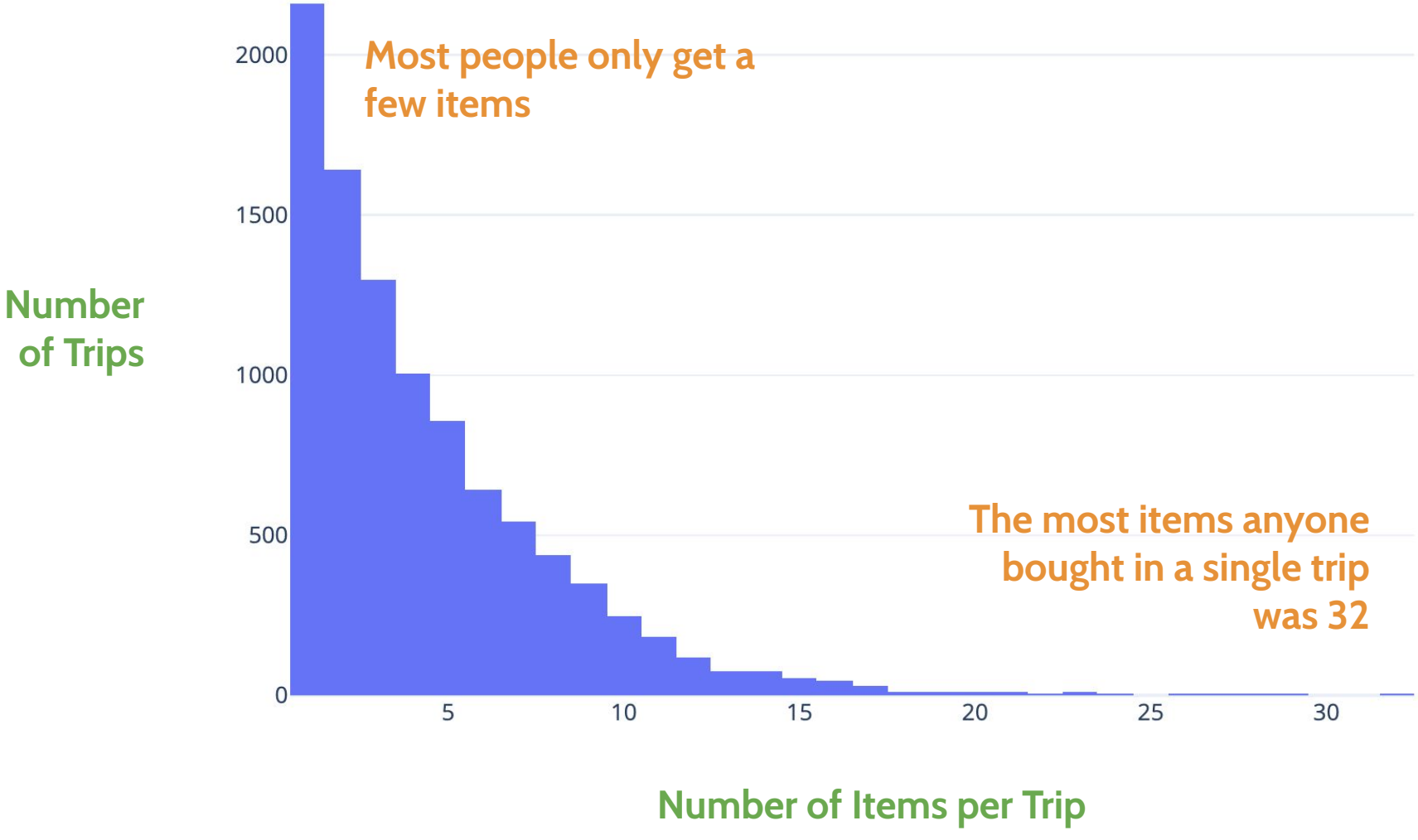
## 170 products (items, features)
## 10k customers (trips)

| beef | hamburger meat | fish | citrus fruit | tropical fruit | pip fruit | grapes | berries | nuts/prunes | root vegetables | onions | herbs | other vegetables |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

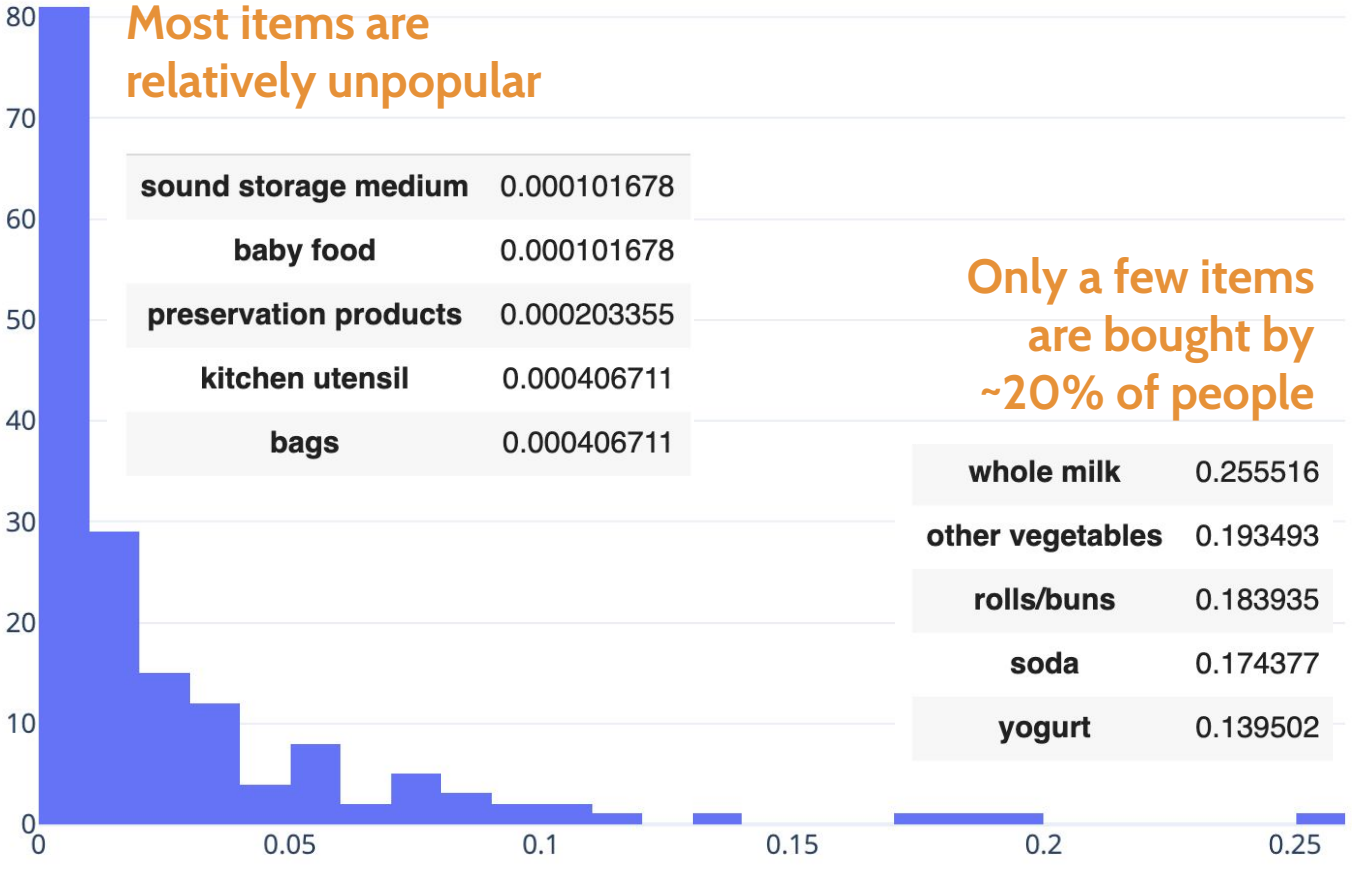
Items

Trips


flower soil/fertilizer
light bulbs
kitchen utensil
rubbing alcohol
make up remover
toilet cleaner
softener
specialty bar
cake bar
popcorn
white wine
whisky
fruit/vegetable juice
tea
cat food
jam
pudding powder
sauces
ketchup
specialty fat
rice
semi-finished bread
rolls/buns
frozen chicken
tidbits
spread cheese
soft cheese
whipped/sour cream
butter
onions
pip fruit
beef
finished products
frankfurter
8k
6k
4k
2k
0


Most people only get a few items
The most items anyone bought in a single trip was 32
Number of Trips
Number of Items per Trip
2000
1500
1000
500
0
5
10
15
20
25
30

Number of items




Most items are relatively unpopular

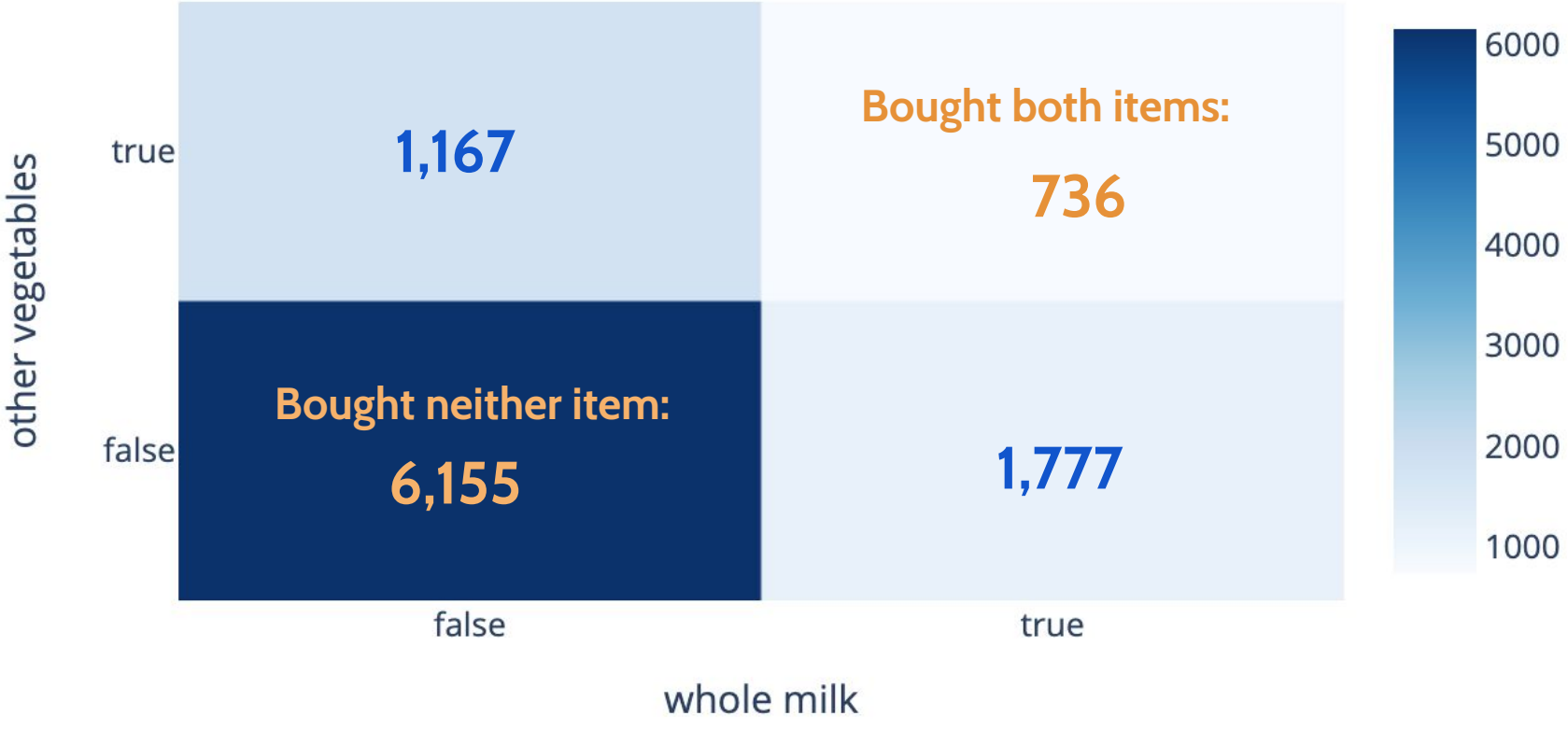
| | |
|---|---|
| sound storage medium | 0.000101678 |
| baby food | 0.000101678 |
| preservation products | 0.000203355 |
| kitchen utensil | 0.000406711 |
| bags | 0.000406711 |

Only a few items are bought by ~20% of people

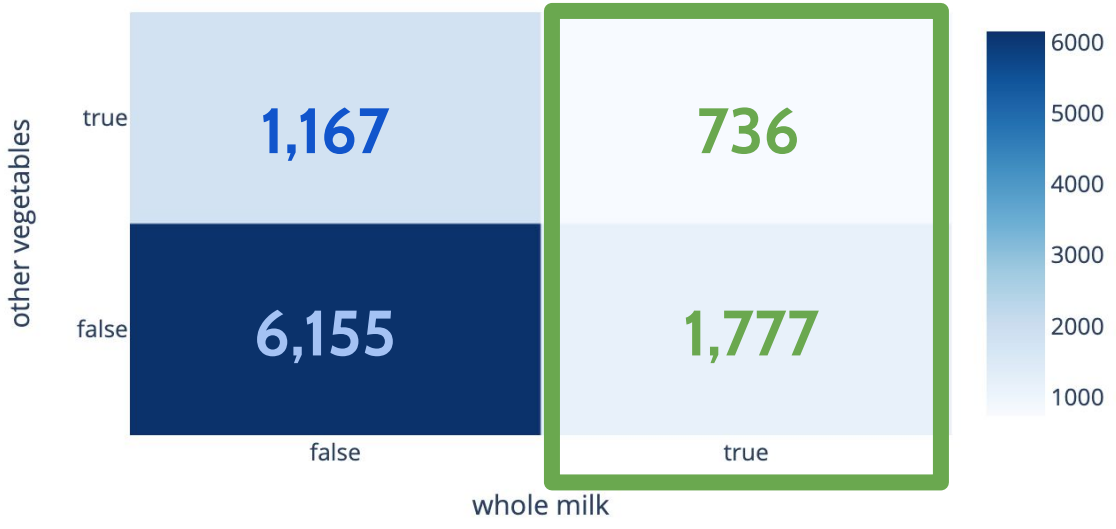
| | |
|---|---|
| whole milk | 0.255516 |
| other vegetables | 0.193493 |
| rolls/buns | 0.183935 |
| soda | 0.174377 |
| yogurt | 0.139502 |

Support (% of people buying this item)

# Groceries

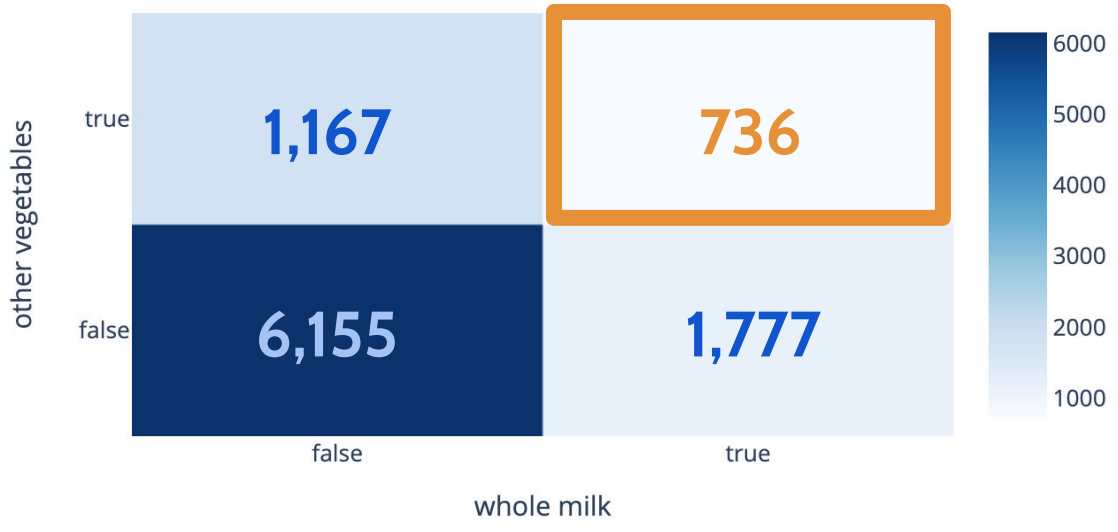

other vegetables
true
false
1,167
Bought both items:
736
Bought neither item:
6,155
1,777
false
true
whole milk
6000
5000
4000
3000
2000
1000

# Calculating Marginal Probability




$$\frac{736 + 1{,}777}{736 + 1{,}777 + 1{,}167 + 6{,}155} \approx 25.5\%$$

# Calculating Joint Probability



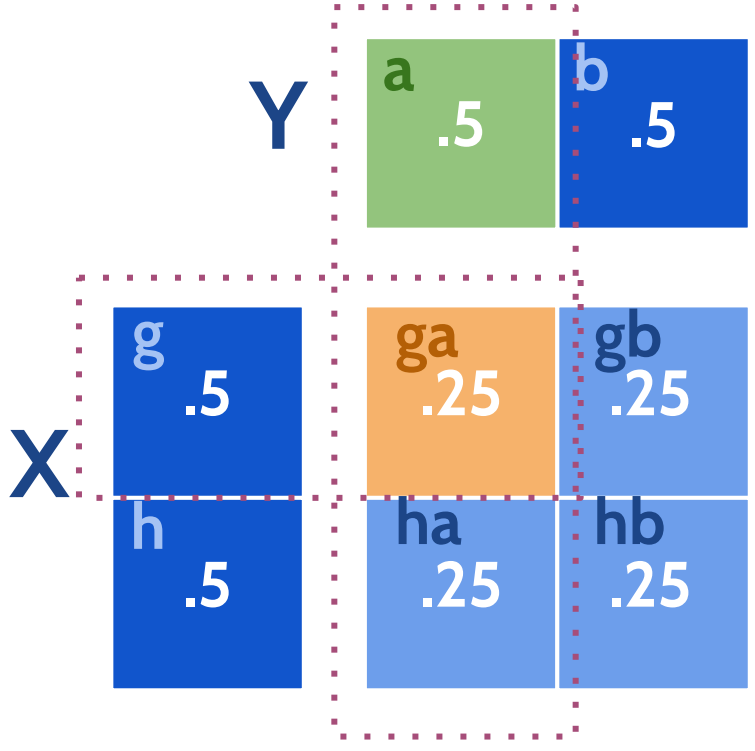

$$\frac{736}{736 + 1{,}167 + 6{,}155 + 1{,}777} \approx 7.5\%$$

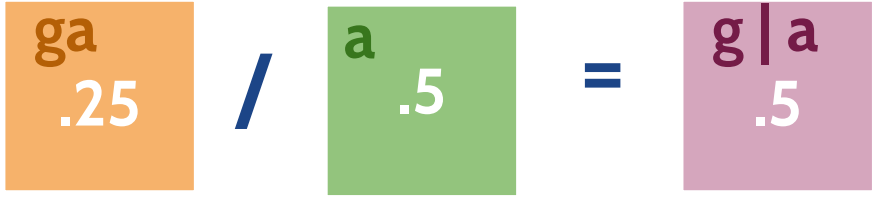
# Calculating Conditional Probability

Marginal: P( Ya )

Joint: P( Xg , Ya )

Conditional: P( Xg | Ya )

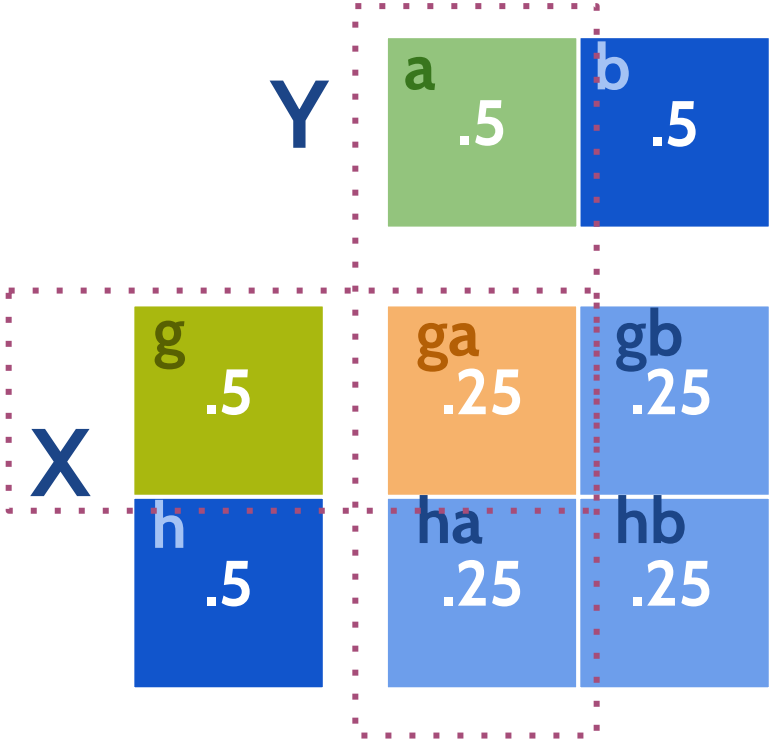


# Calculating “Lift”

ga .25 / a .5 = g | a .5

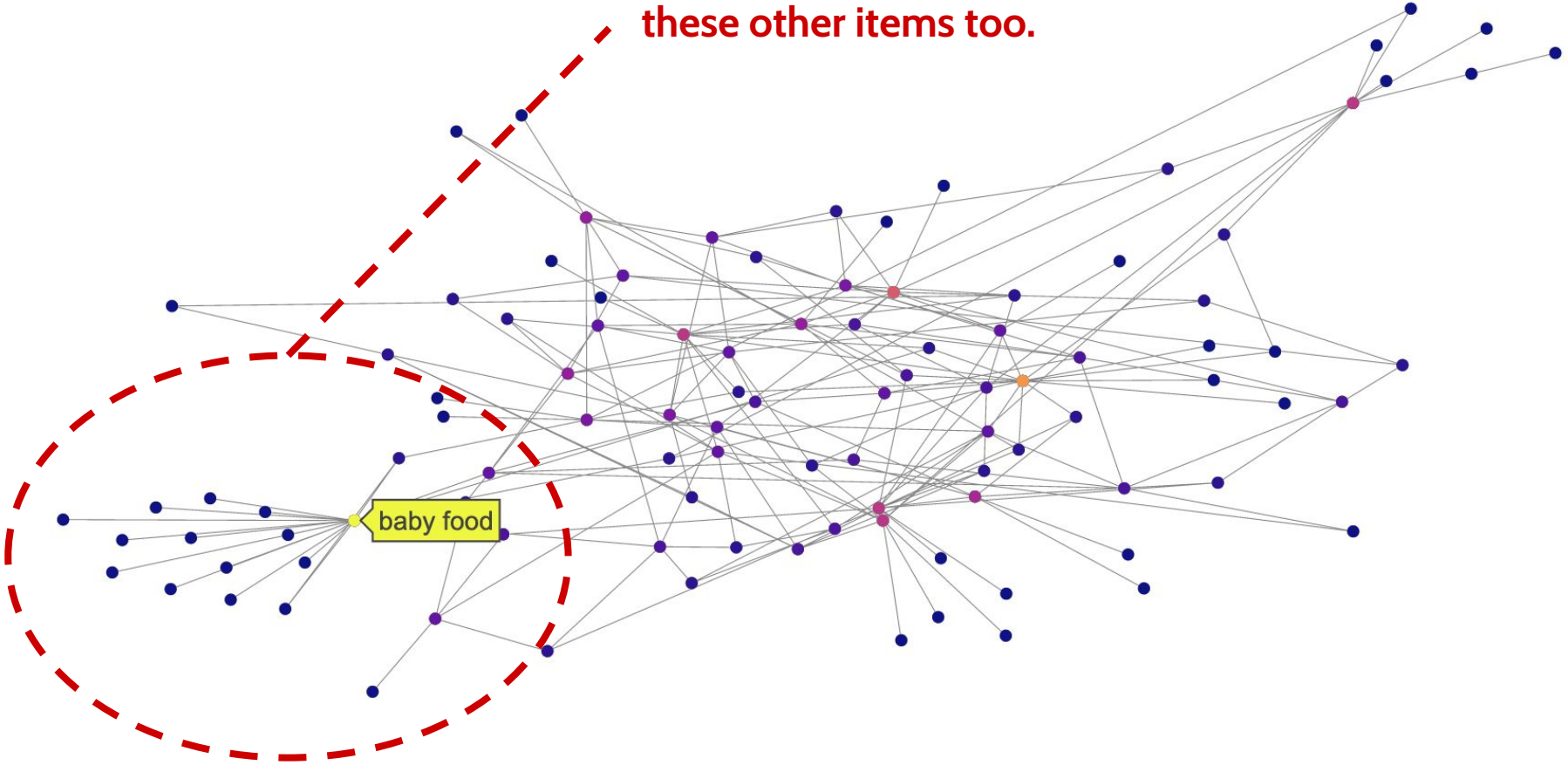
g | a .5 / g .5 = 1




“It is 1 times more likely that X = g if I know that Y = a .”

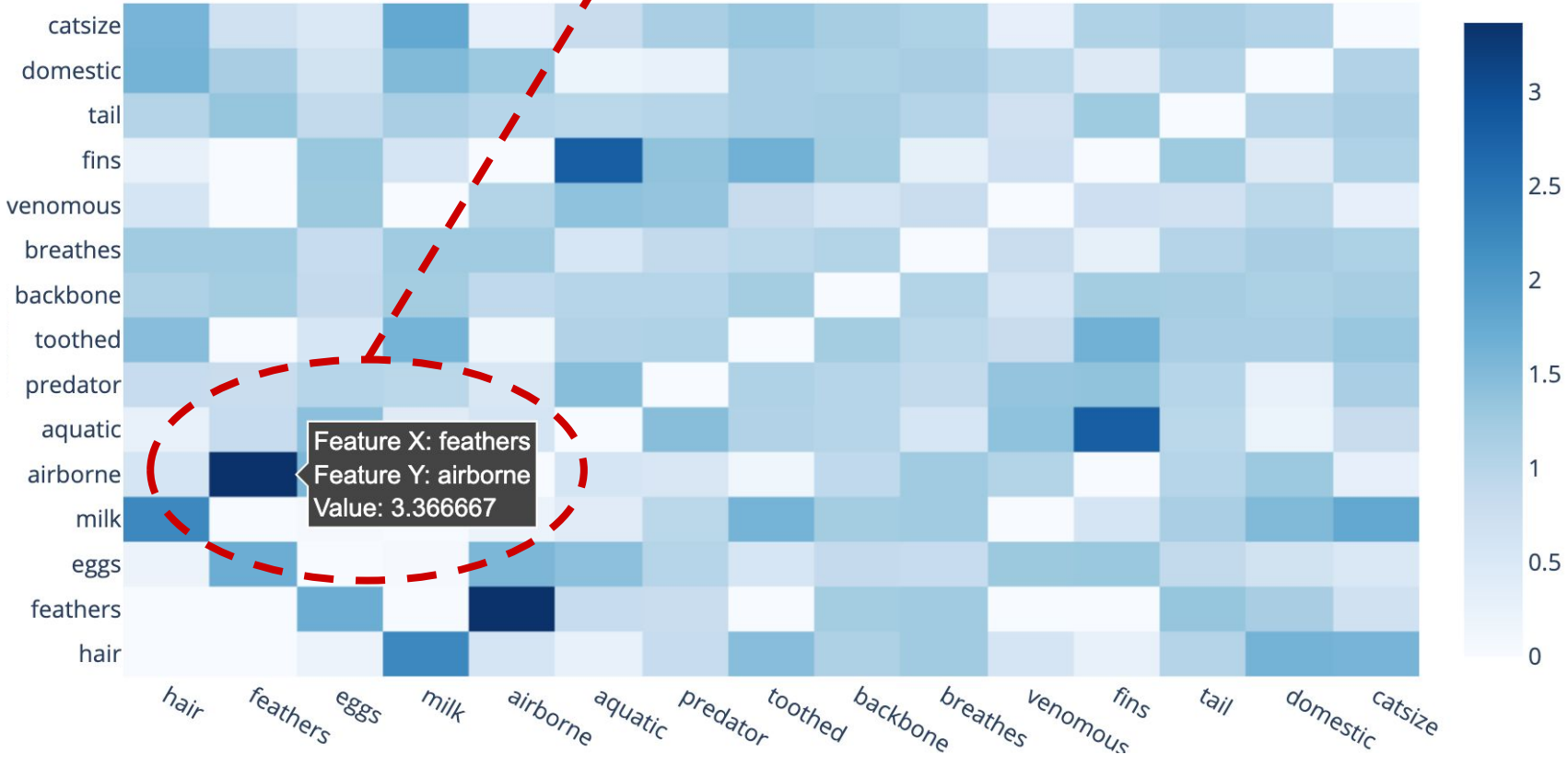
# Association Network

If someone bought baby food, it's much more likely that they bought these other items too.

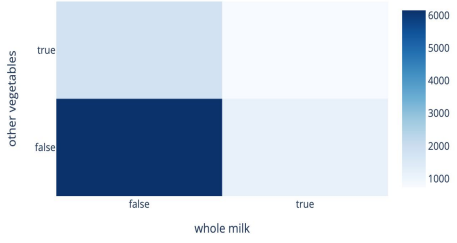


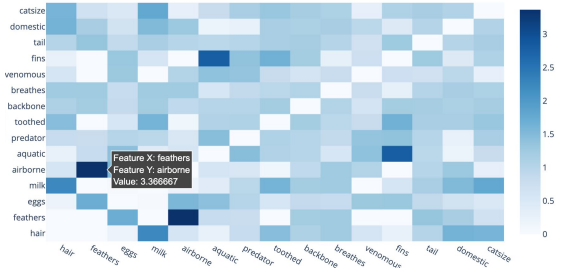
# Association Network

**If I know that an animal has feathers, I'm 3x more likely to know that it can be airborne.**

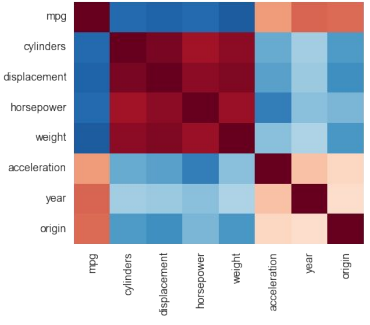


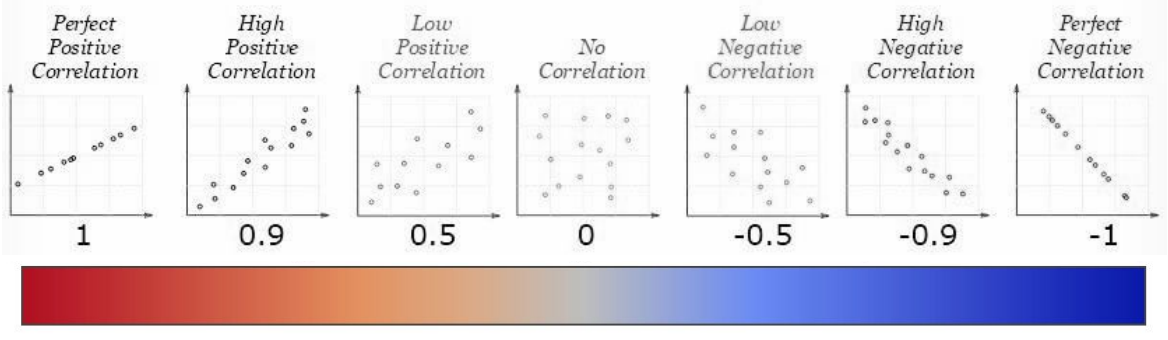
# Association Network For Binary Features




# Correlation Network For Continuous Features

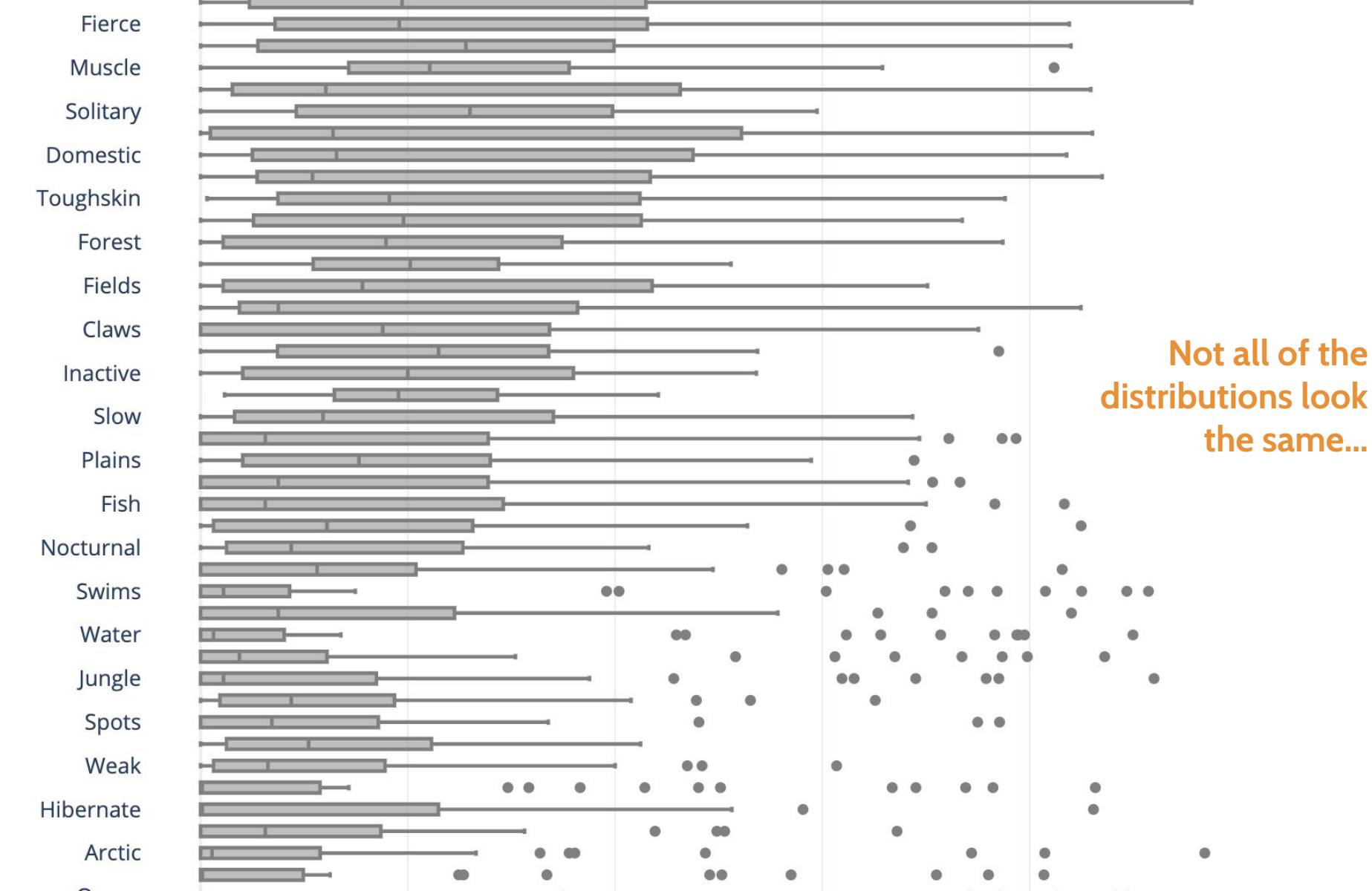


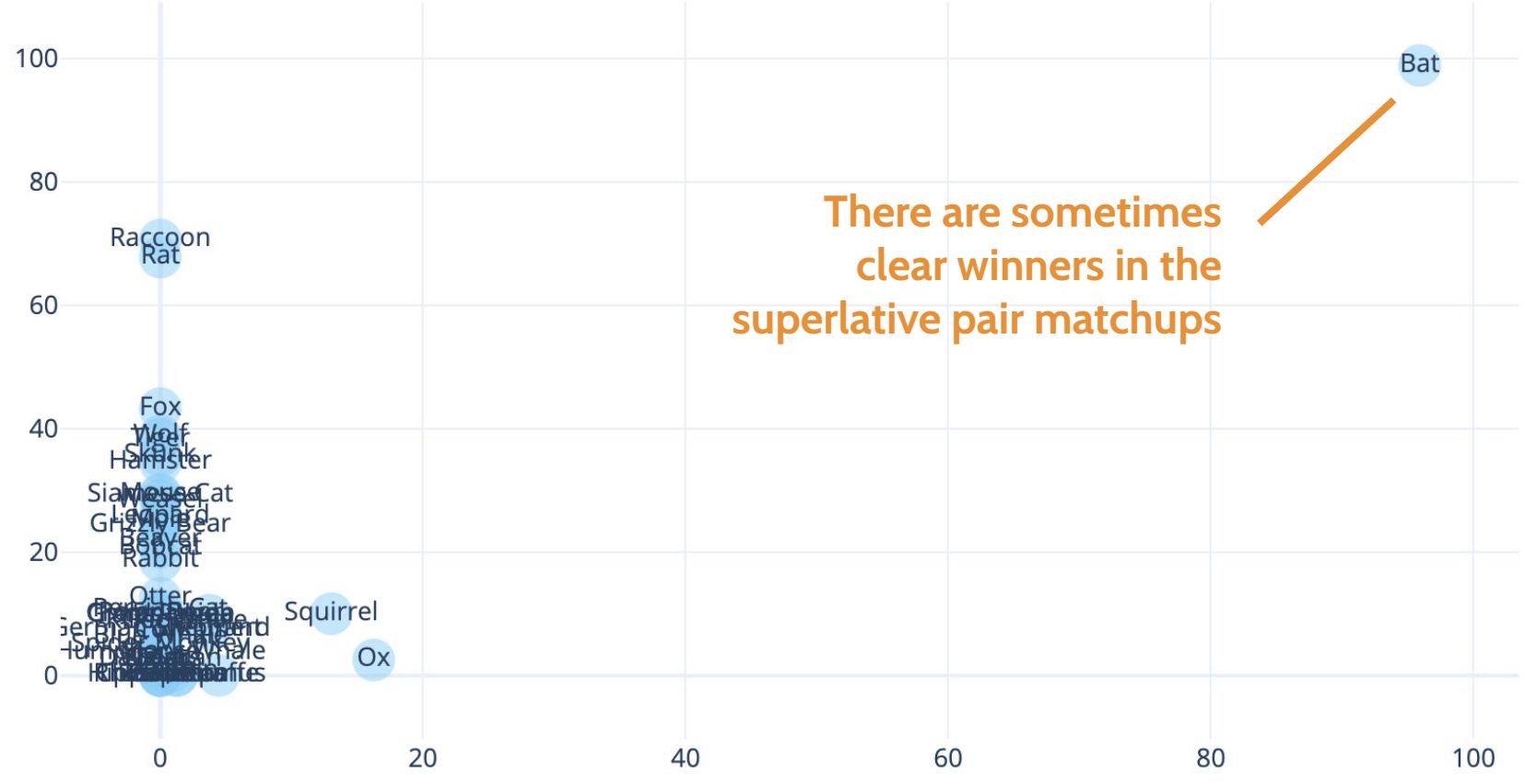
# Animals: Correlation Networks of Quantitative Features

## 85 characteristics (features)
## 49 animals (observations)

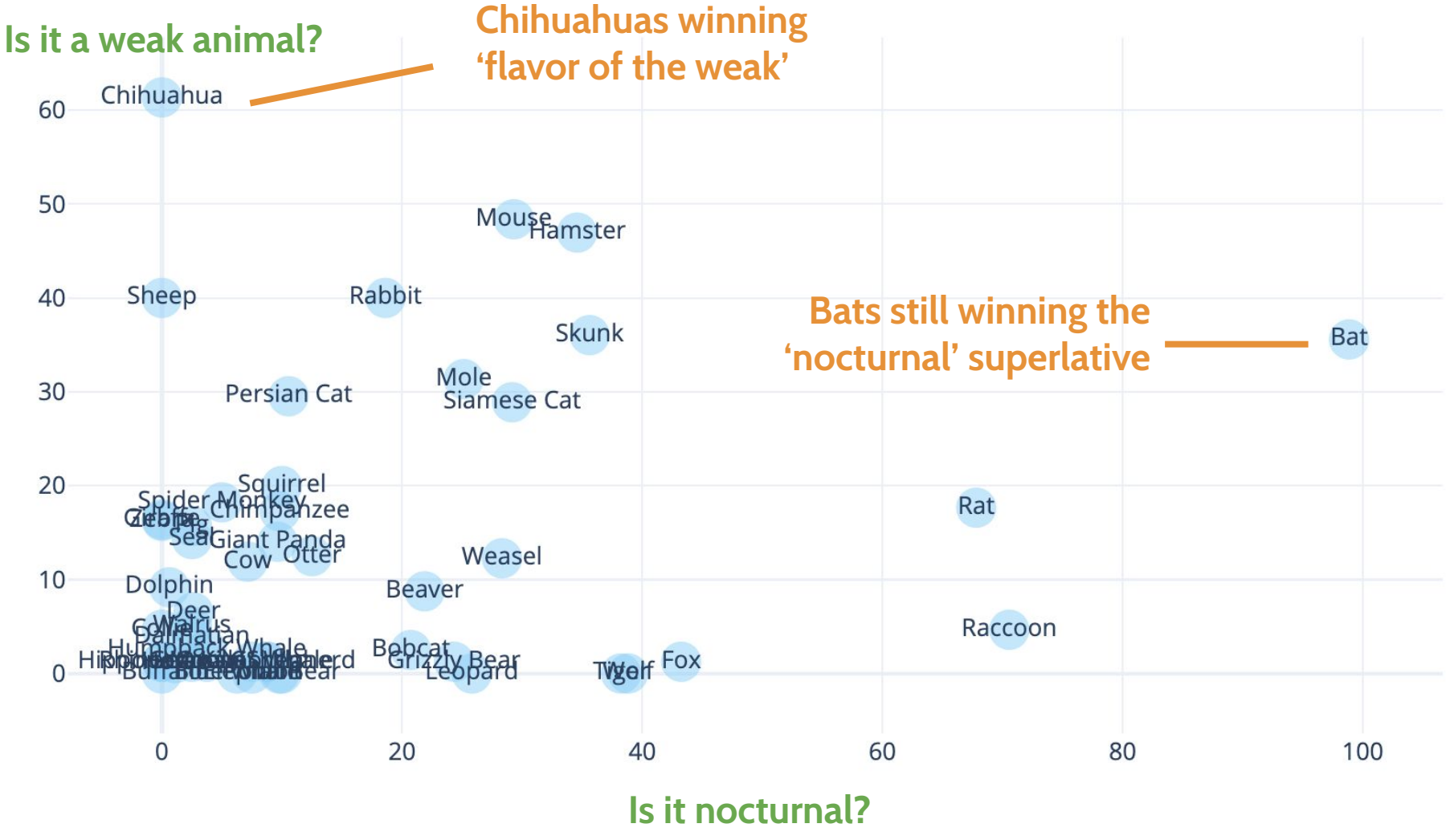
| animal | Black | White | Blue | Brown | Gray | Orange | Red | Yellow | Patches | Spots | Stripes | Furry |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grizzly Bear | 39.25 | 1.39 | 0.00 | 74.14 | 3.75 | 0.00 | 0.0 | 0.00 | 1.25 | 0.00 | 0.00 | 82.37 |
| Killer Whale | 83.40 | 64.79 | 0.00 | 0.00 | 1.25 | 0.00 | 0.0 | 0.00 | 68.49 | 32.69 | 0.00 | 1.25 |
| Beaver | 19.38 | 0.00 | 0.00 | 87.81 | 7.50 | 0.00 | 0.0 | 0.00 | 0.00 | 7.50 | 0.00 | 46.25 |
| Dalmatian | 69.58 | 73.33 | 0.00 | 6.39 | 0.00 | 0.00 | 0.0 | 0.00 | 37.08 | 100.00 | 0.00 | 27.15 |
| Persian Cat | 19.38 | 50.09 | 29.44 | 8.98 | 38.19 | 0.00 | 0.0 | 0.00 | 17.93 | 6.25 | 6.25 | 90.19 |
| Horse | 44.90 | 42.91 | 4.44 | 69.41 | 35.94 | 0.00 | 0.0 | 0.00 | 22.29 | 15.80 | 0.00 | 40.58 |
| German Shepherd | 43.54 | 15.88 | 5.00 | 54.16 | 26.82 | 3.12 | 2.5 | 0.38 | 48.78 | 11.59 | 1.56 | 66.05 |

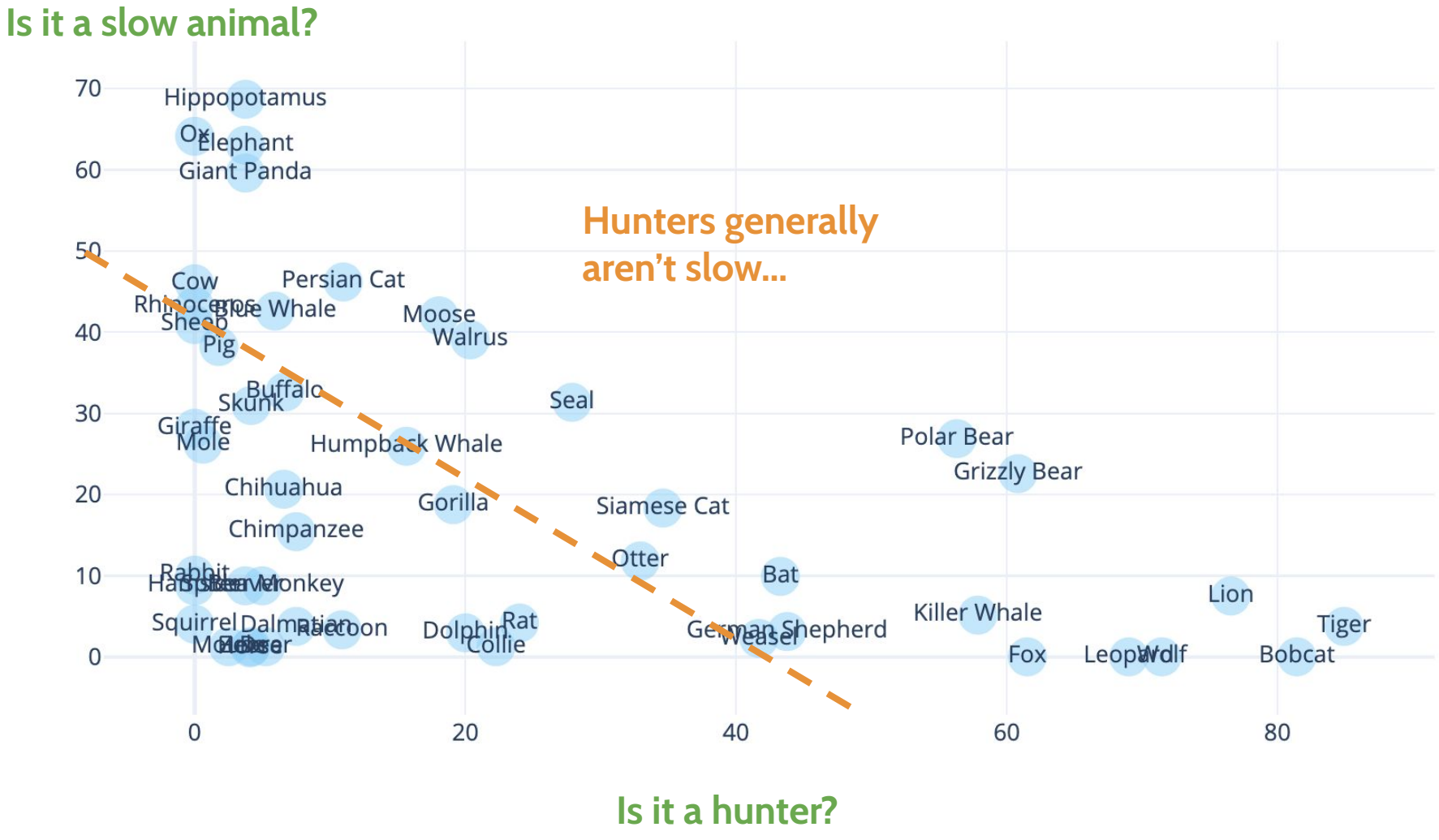

Fierce
Muscle
Solitary
Domestic
Toughskin
Forest
Fields
Claws
Inactive
Slow
Plains
Fish
Nocturnal
Swims
Water
Jungle
Spots
Weak
Hibernate
Arctic
Not all of the distributions look the same...

Is it nocturnal?




Does it fly?

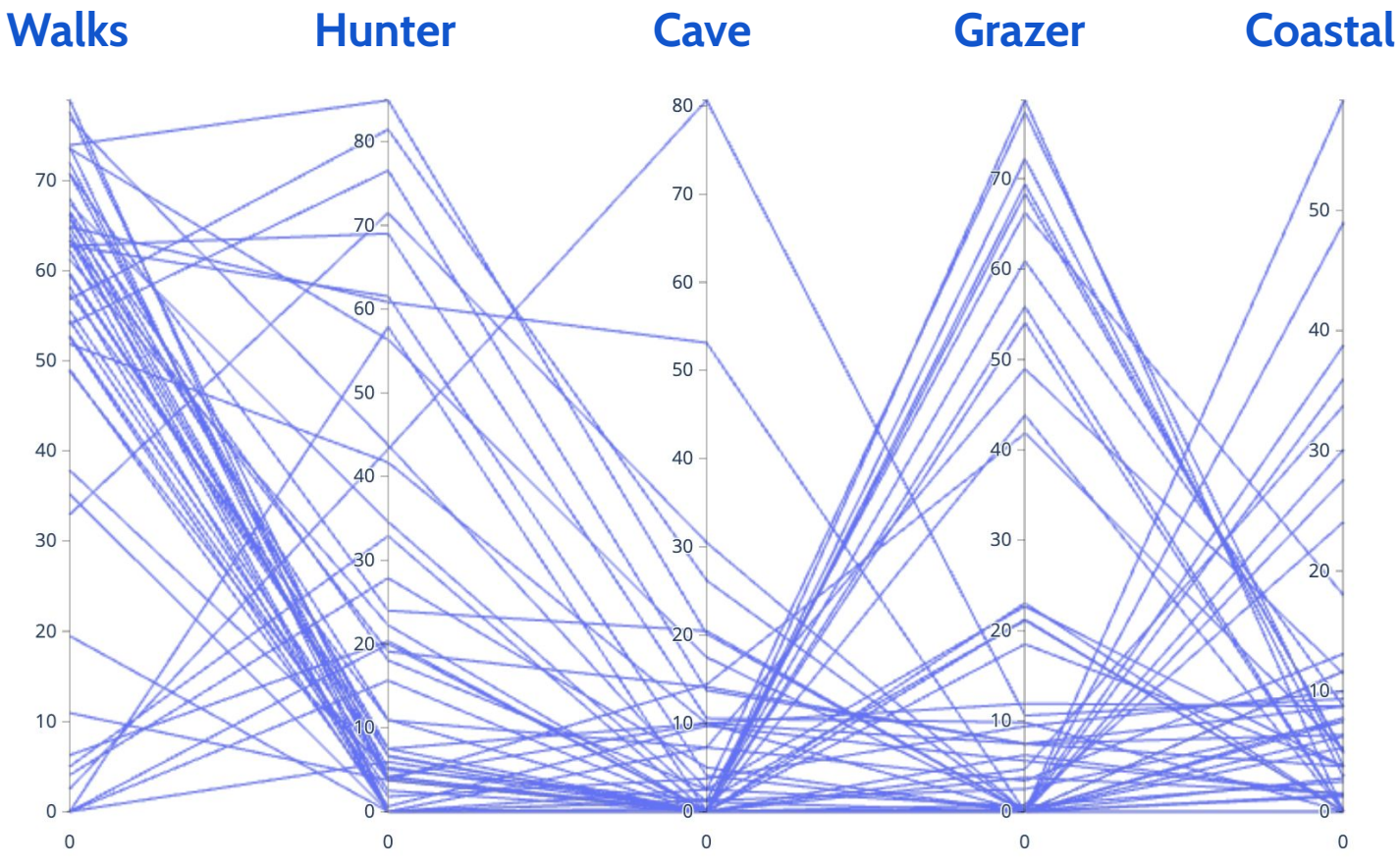

Is it a weak animal?
Chihuahuas winning
‘flavor of the weak’
Bats still winning the
‘nocturnal’ superlative
Is it nocturnal?
Chihuahua
Mouse
Hamster
Sheep
Rabbit
Skunk
Bat
Mole
Persian Cat
Siamese Cat
Squirrel
Spider Monkey
Chimpanzee
Rat
Giant Panda
Cow
Otter
Weasel
Dolphin
Beaver
Deer
Raccoon
Bobcat
Grizzly Bear
Fox
Leopard
Tiger
Wolf
60
50
40
30
20
10
0
0
20
40
60
80
100


Is it a slow animal?
Hippopotamus
Ox
Elephant
Giant Panda
Cow
Persian Cat
Blue Whale
Moose
Sheep
Walrus
Pig
Buffalo
Seal
Skunk
Giraffe
Mole
Humpback Whale
Polar Bear
Grizzly Bear
Chihuahua
Gorilla
Siamese Cat
Chimpanzee
Otter
Bat
Rabbit
Monkey
Lion
Killer Whale
Squirrel
Dalmatian
Raccoon
Dolphin
Rat
Collie
German Shepherd
Weasel
Tiger
Fox
Leopard
Wolf
Bobcat
Hunters generally aren't slow...
70
60
50
40
30
20
10
0
0
20
40
60
80
Is it a hunter?

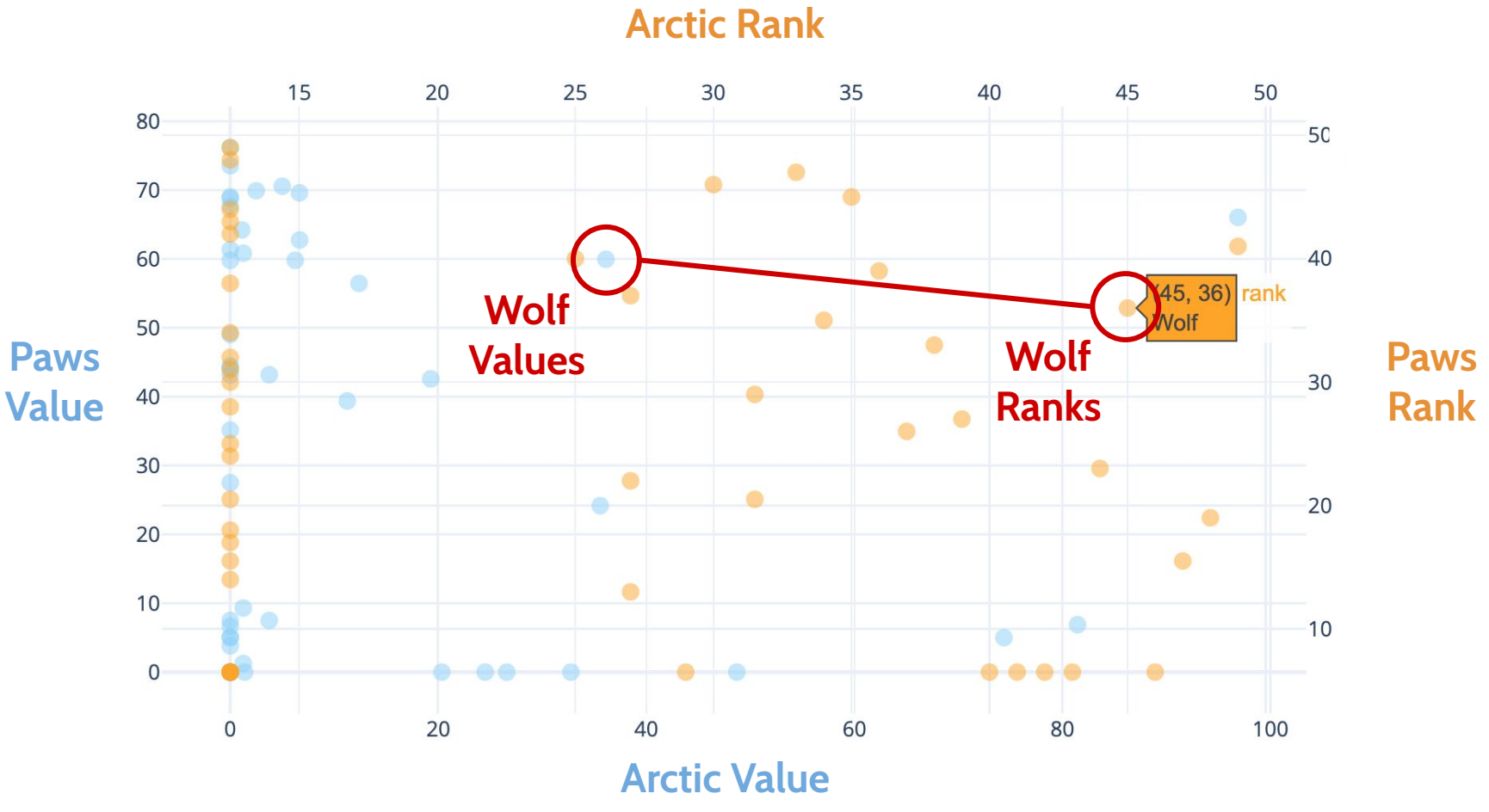
Is it a slow animal?


Is it a hunter?

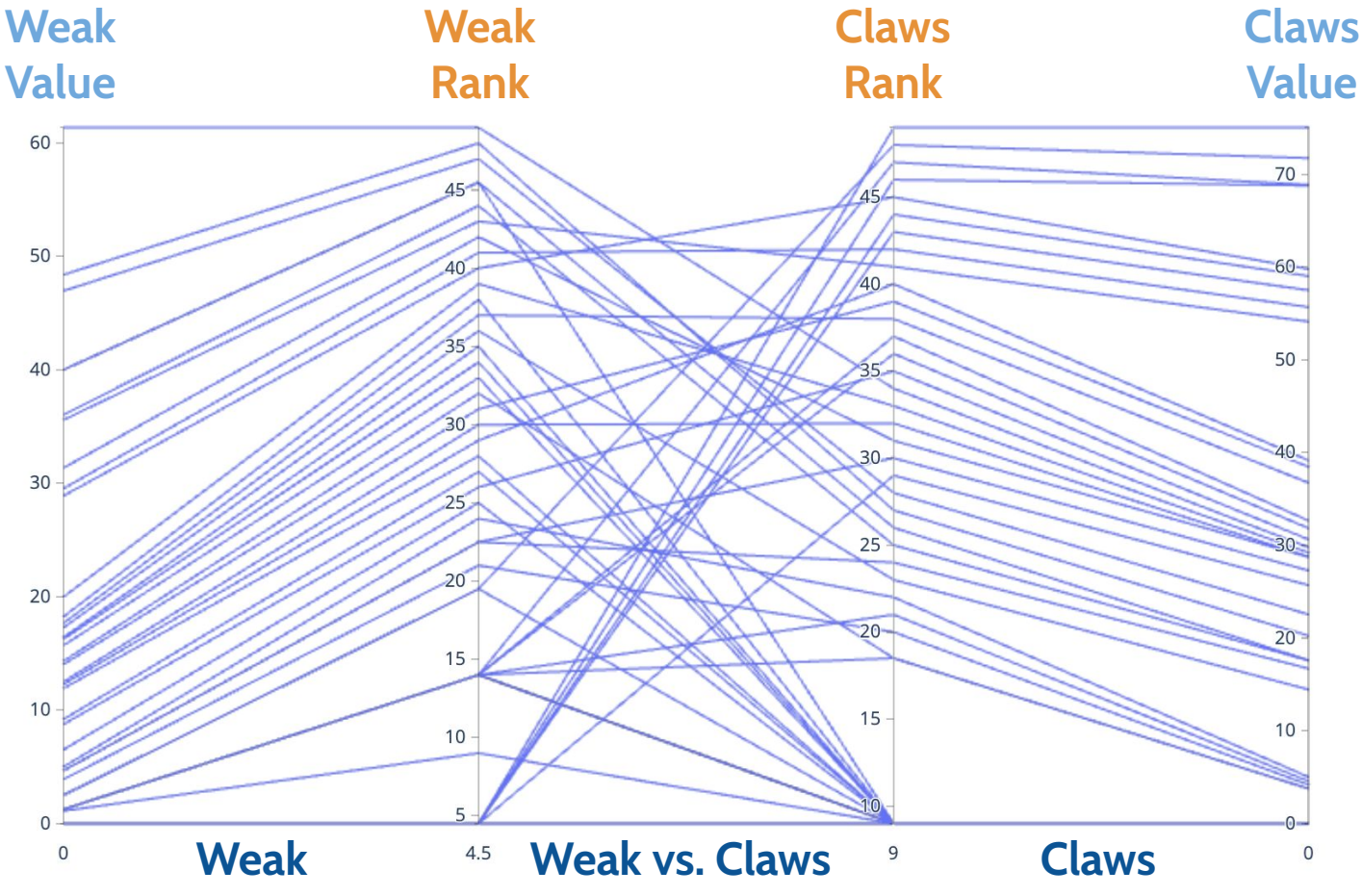
# Parallel Coordinate Plot



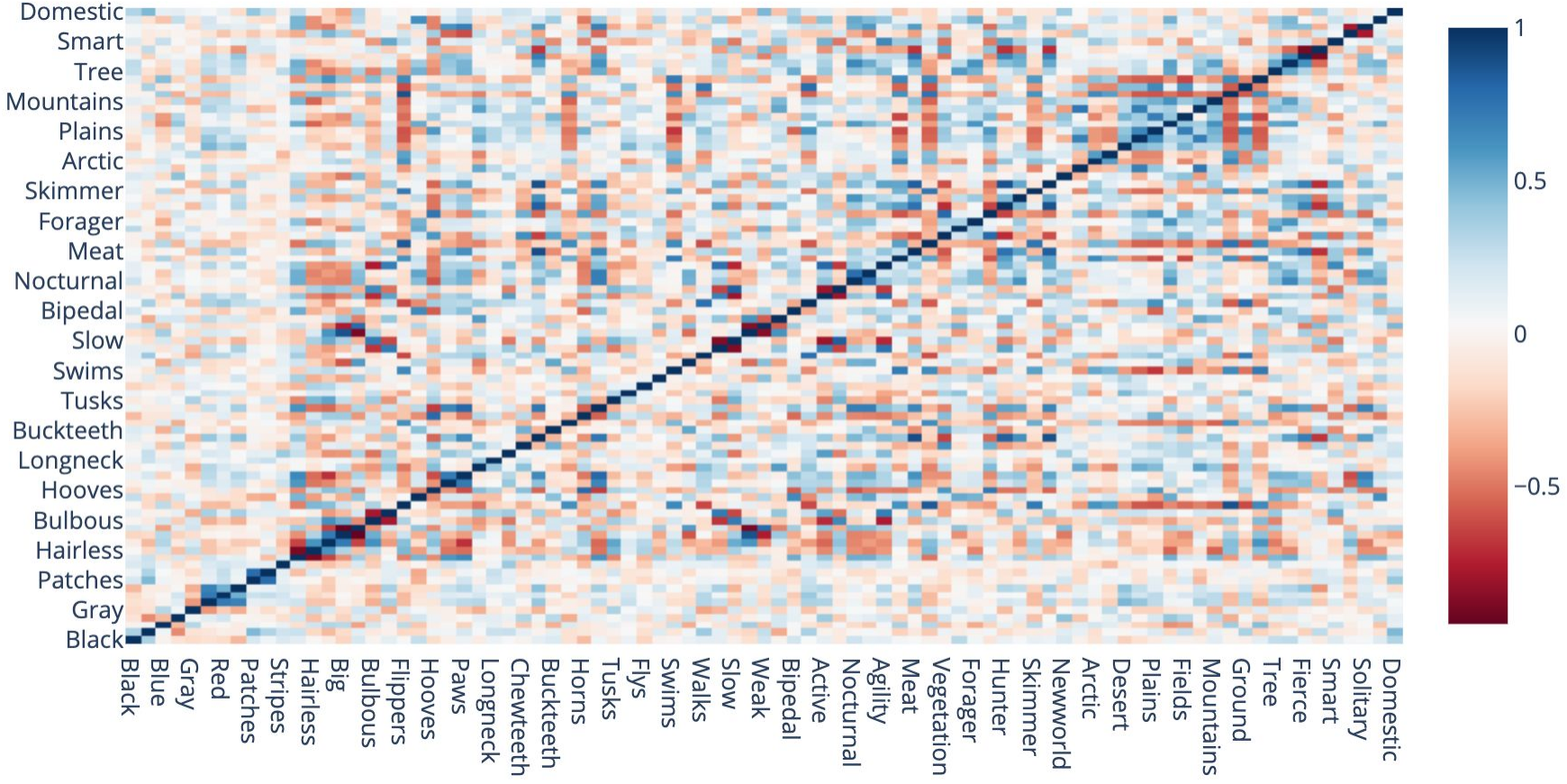
# Rank


Arctic Rank
15
20
25
30
35
40
45
50
80
70
60
50
40
30
20
10
0
50
40
30
20
10
(45, 36) rank
Wolf
Wolf Values
Wolf Ranks
Paws Value
Paws Rank
0
20
40
60
80
100
Arctic Value

# Parallel Coordinate Plot

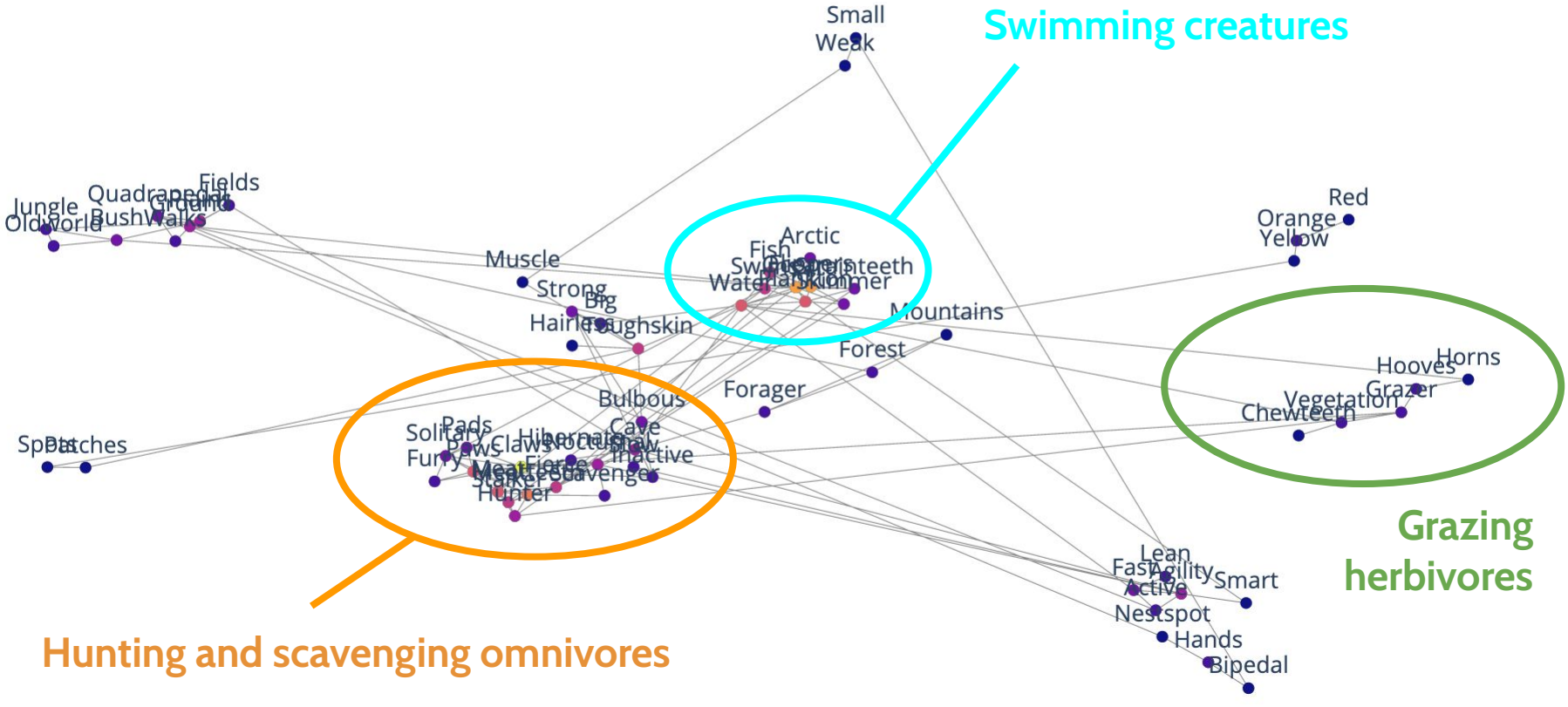


# Correlation Matrix (Heatmap)



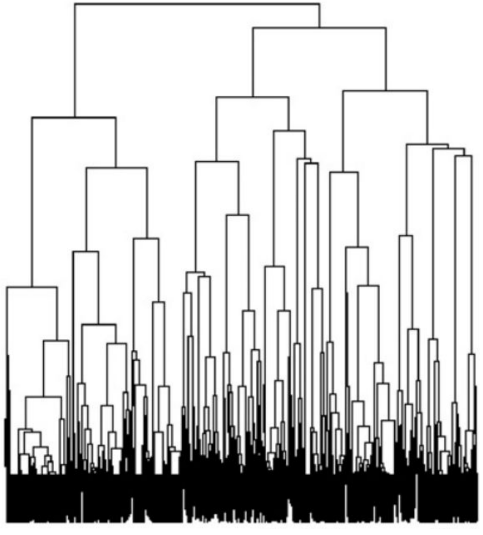
# Correlation Network

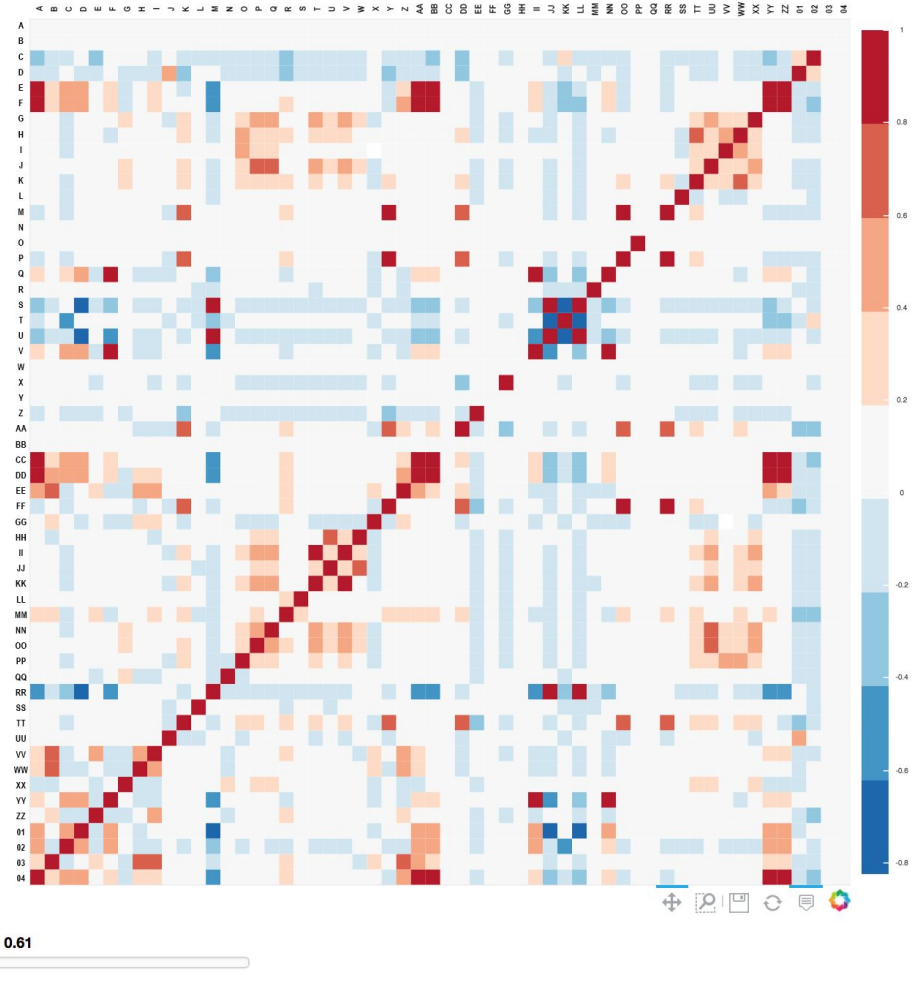
Showing edges where correlation exceeds 0.6



# Correlation Network

sorted using SciPy's hierarchical clustering dendrogram:

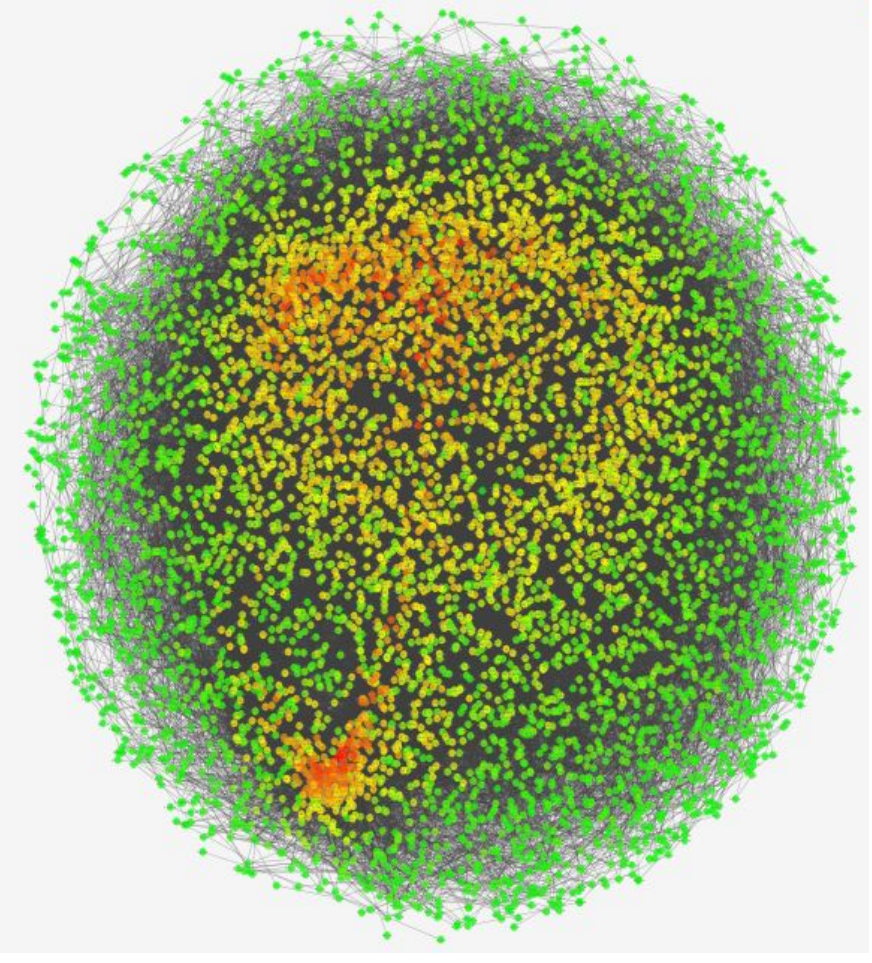


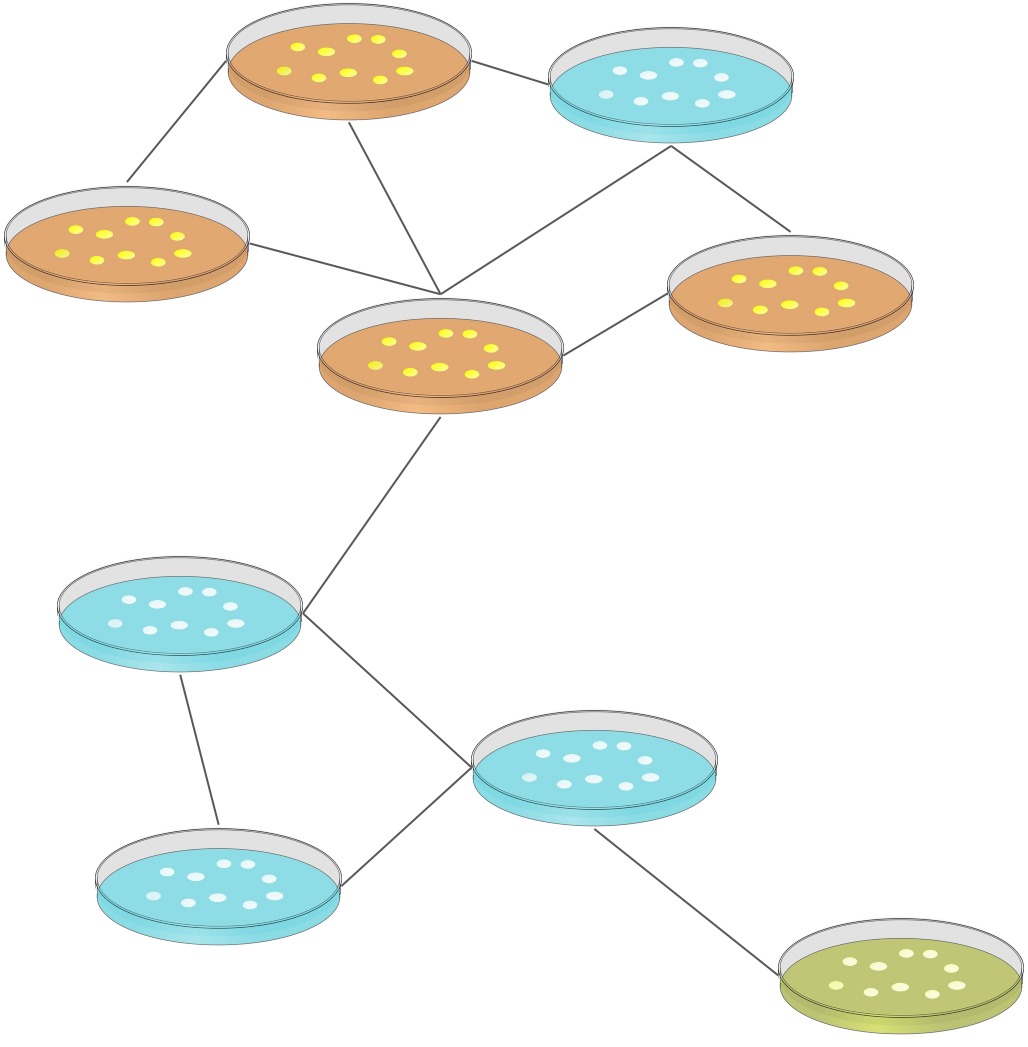



Support: **0.61**

# Gene Correlation Network

There is a lot of existing research about gene correlation / co-expression networks...

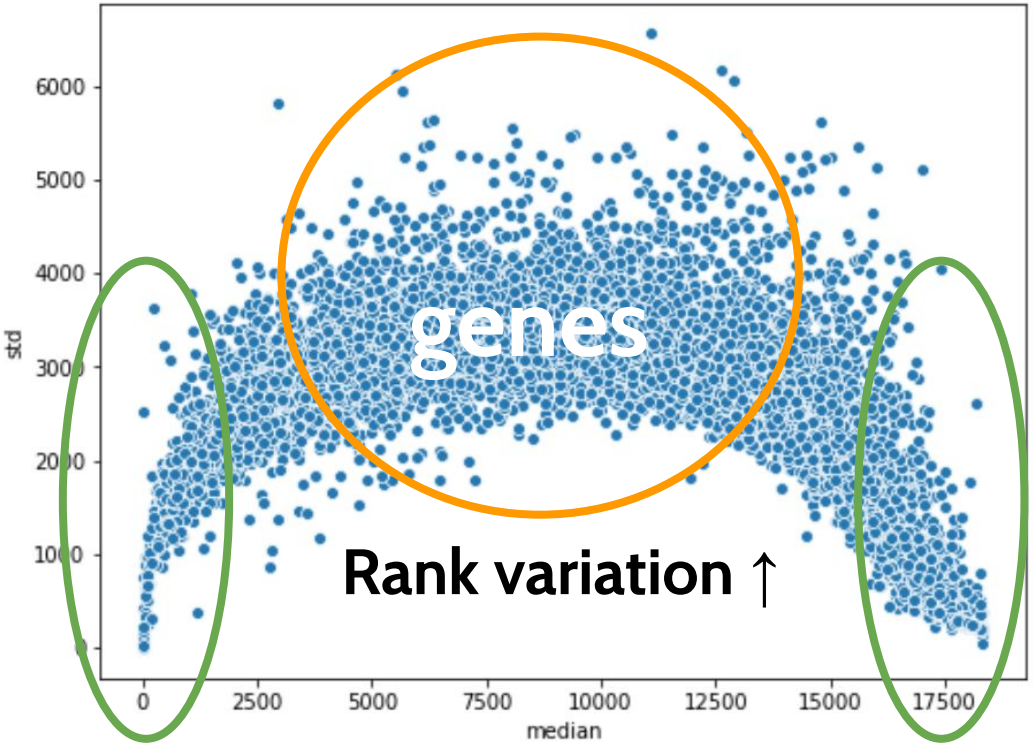




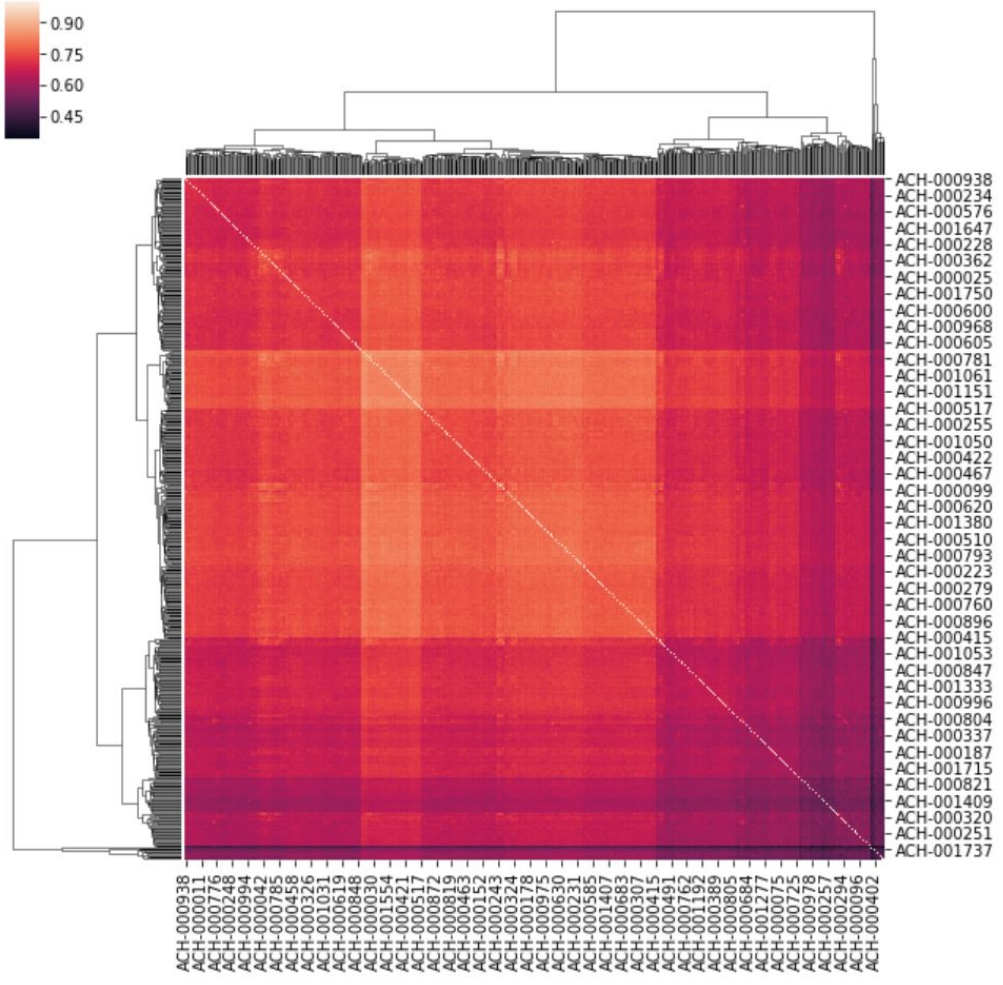
...but how similar are the cell lines of different types of cancer?

There is much agreement on which genes are most or least important...

...but there is a lot of rank variance in the middle gene dependence rankings

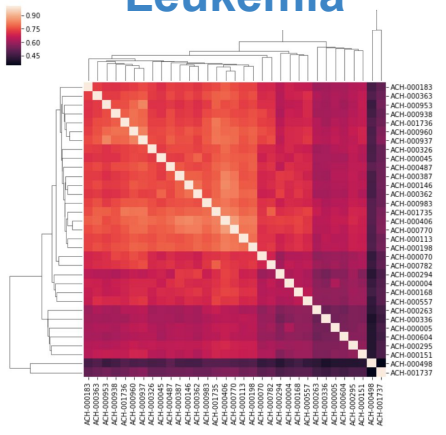


## Gene dependence rank correlations among all cell lines

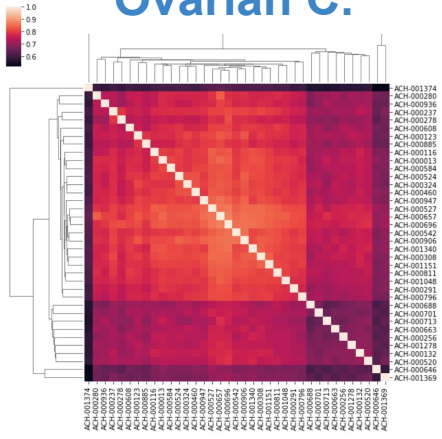


**Gene rank correlations between cell lines** by primary disease

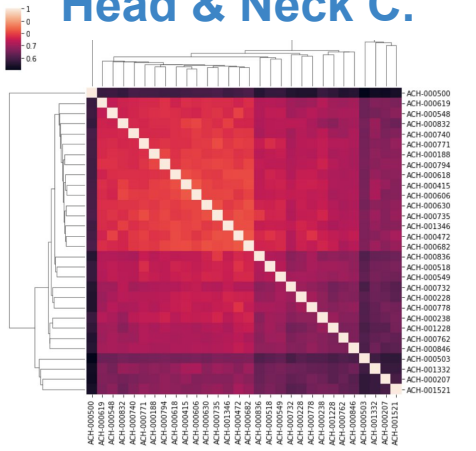
## Leukemia



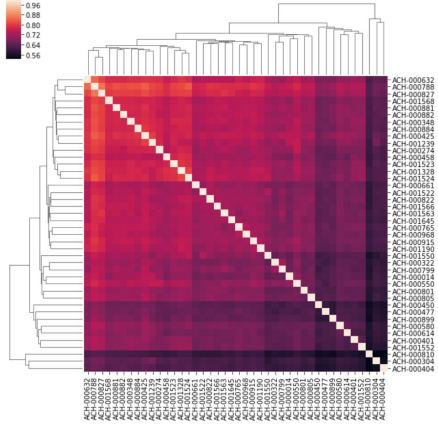
## Ovarian C.



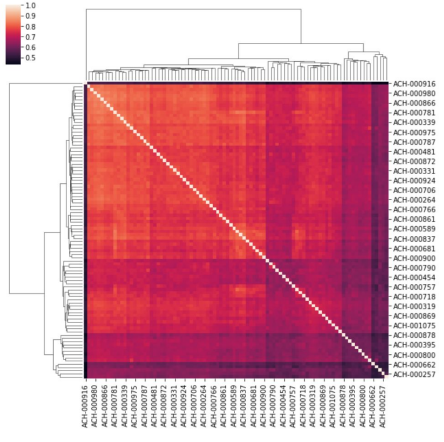
## Head & Neck C.



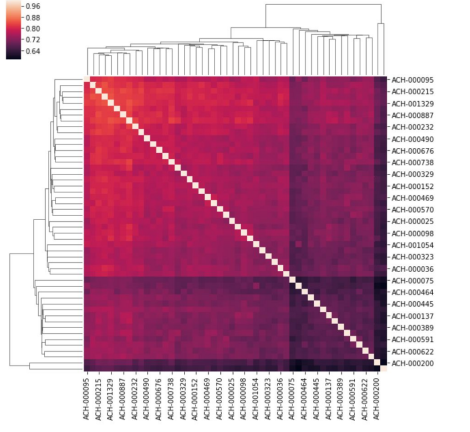
## Skin C.



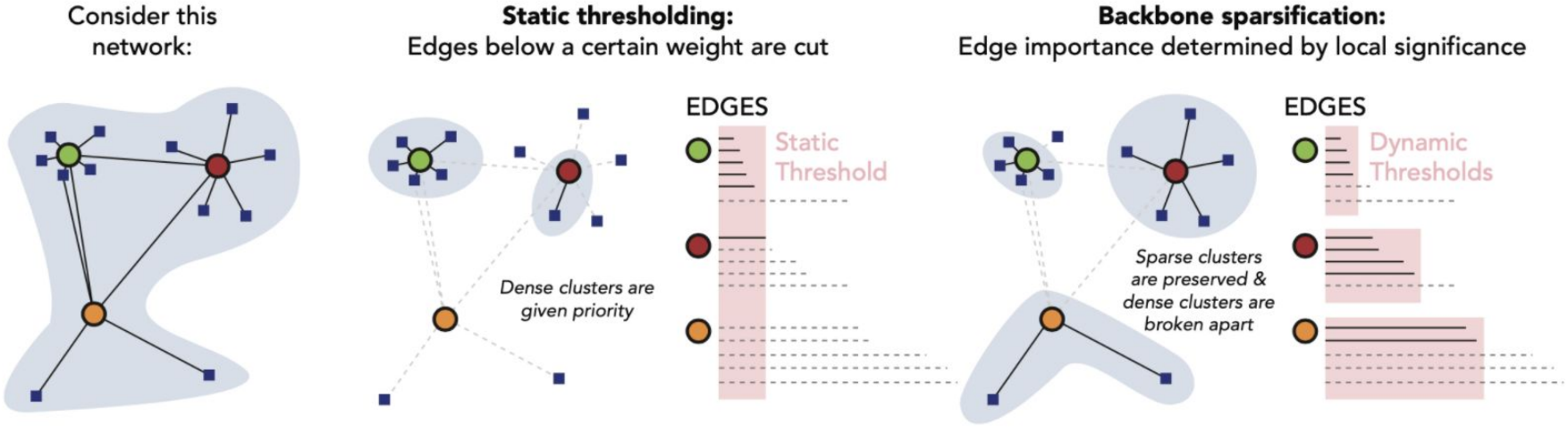
## Lung C.



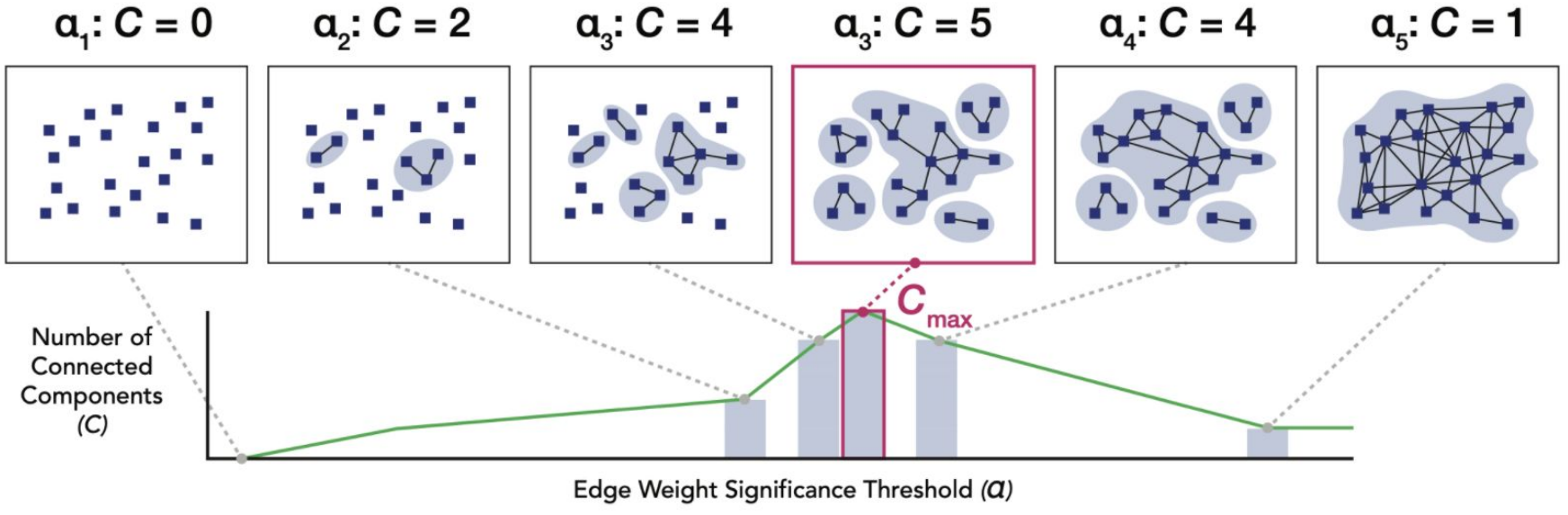
## Brain C.



# Edge Thresholding

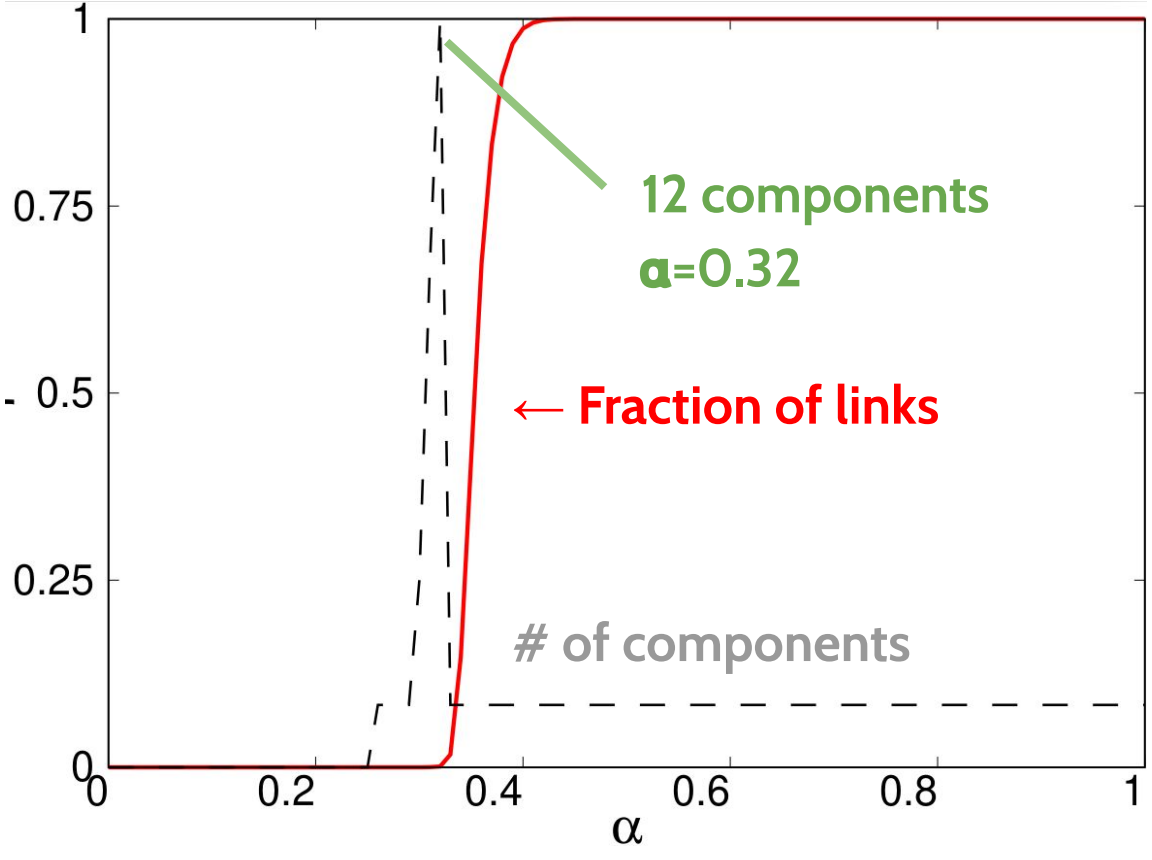


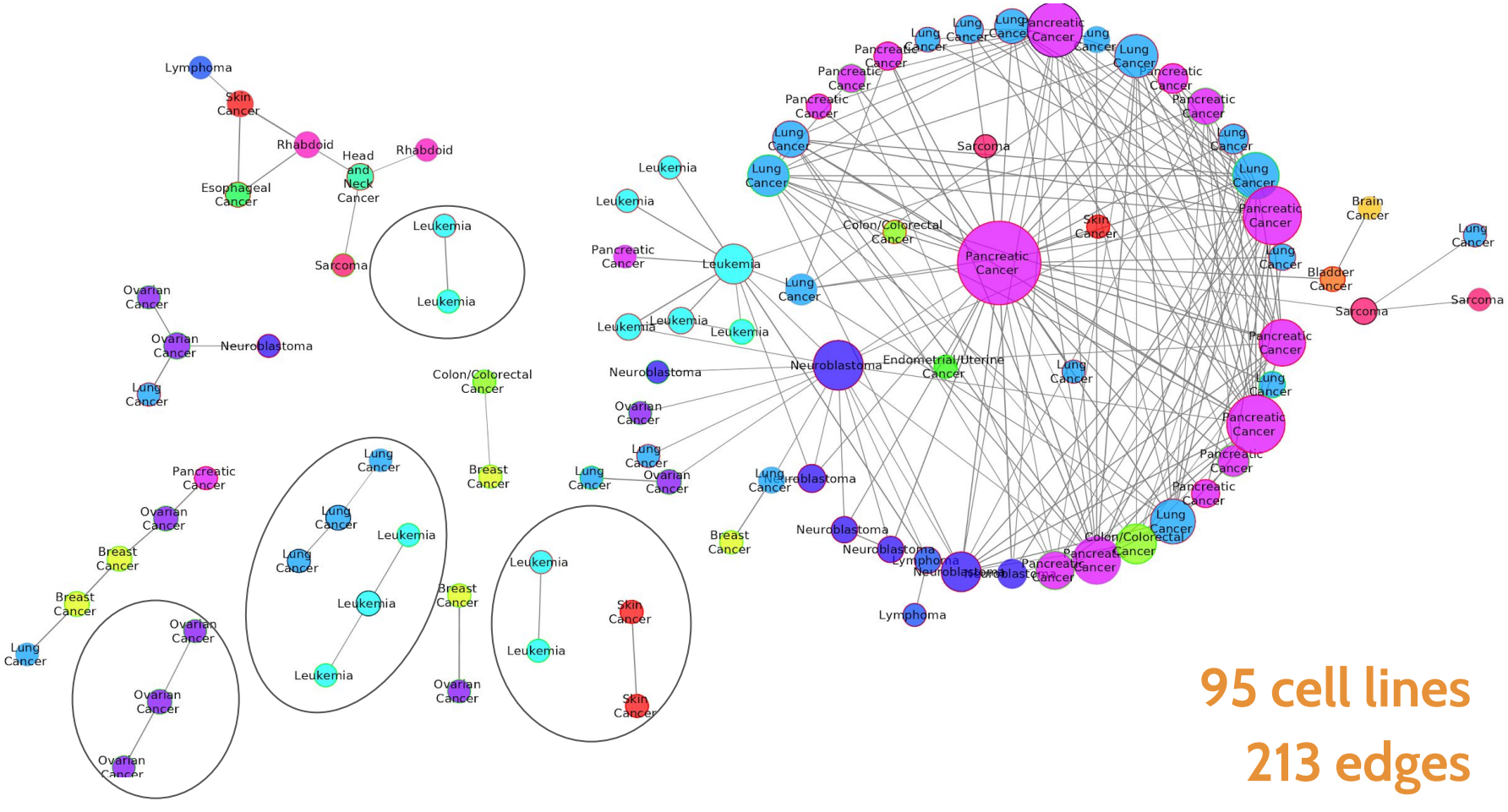
# Edge Thresholding



# Edge Thresholding

**Link weight thresholding** on the cell line rank correlation matrix



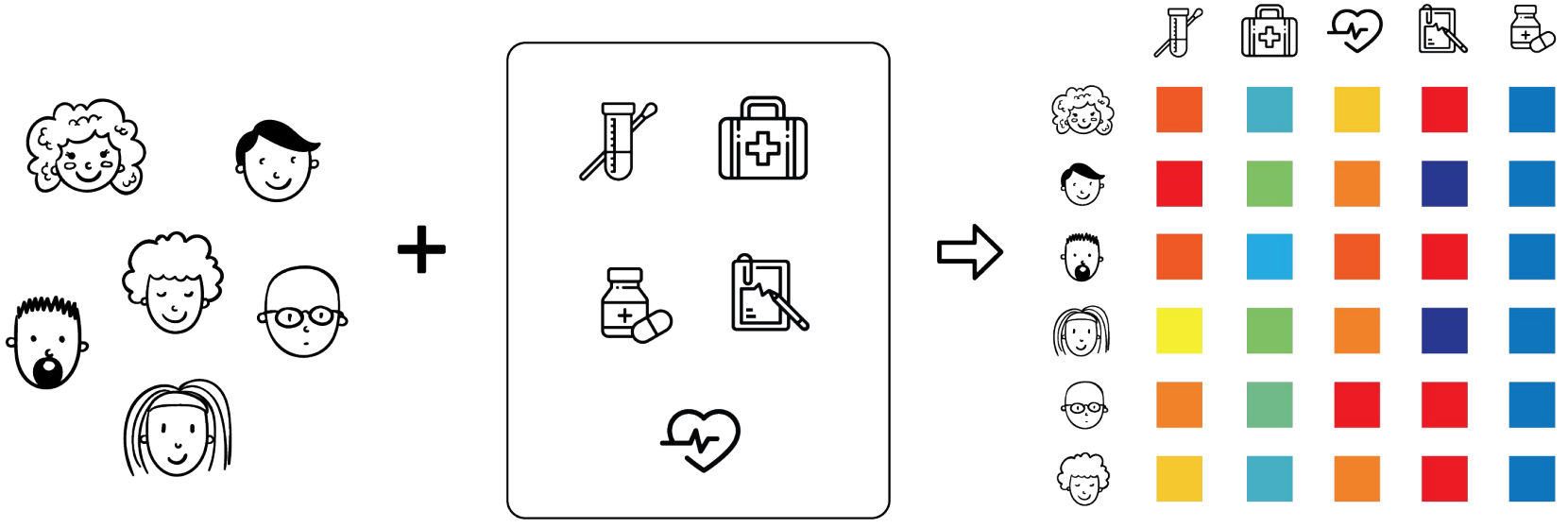



**95 cell lines**

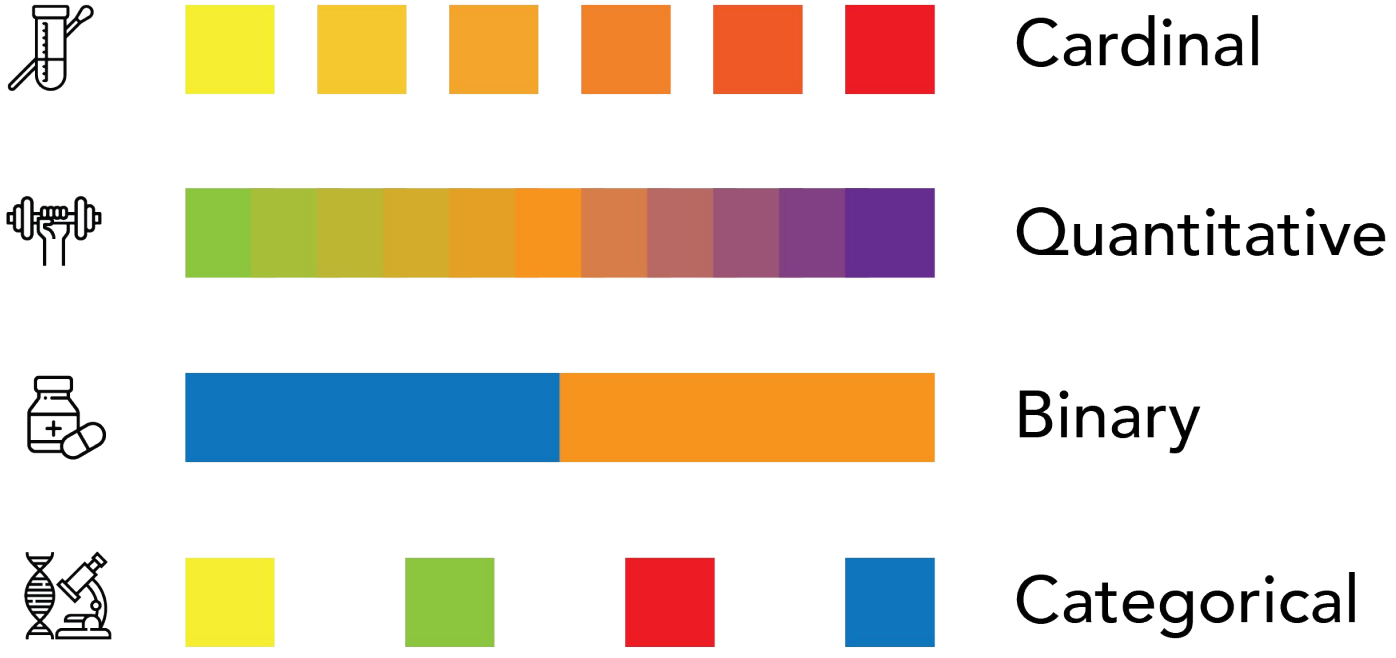
**213 edges**

Ok, we have a few ideas about measuring ‘interesting-ness’...

But what about when we have different kinds of variables?



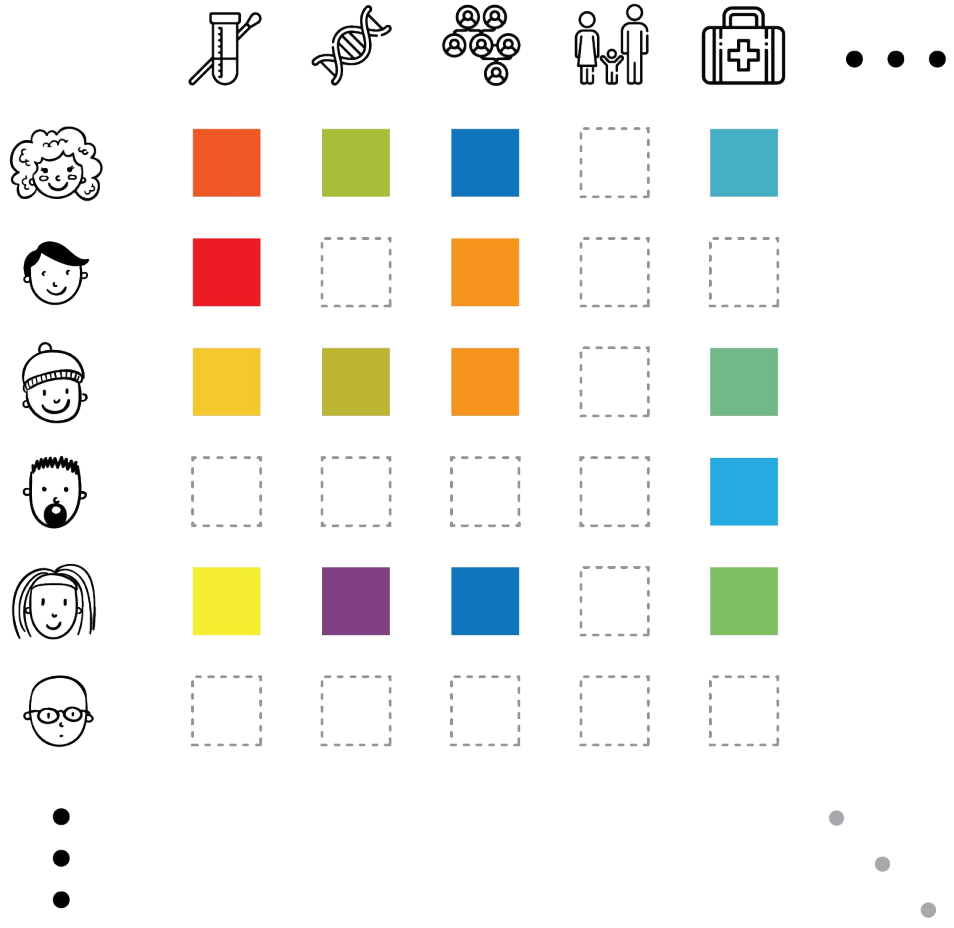
## We have lots of different data types...

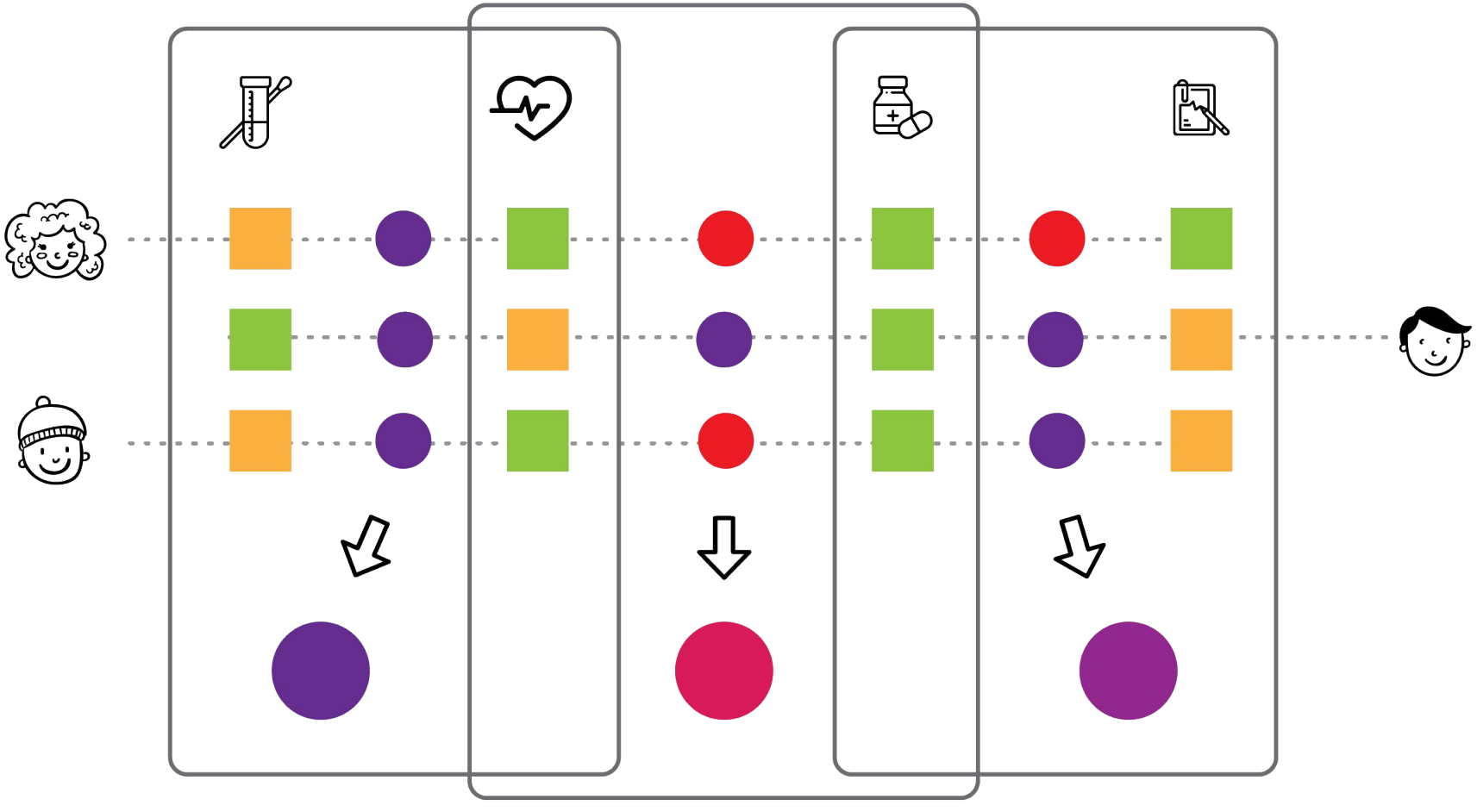


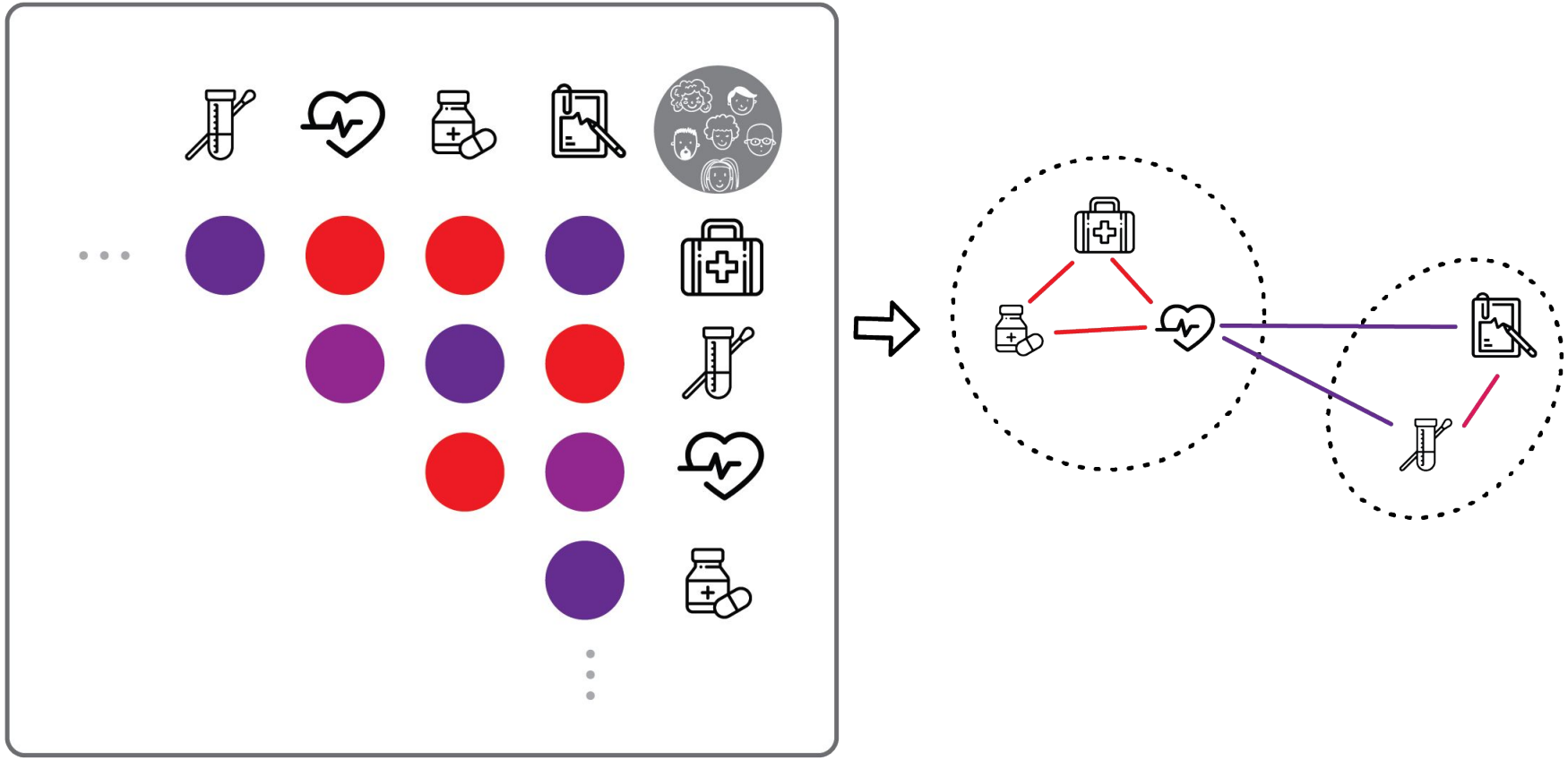
...and we have

a lot of

missing data





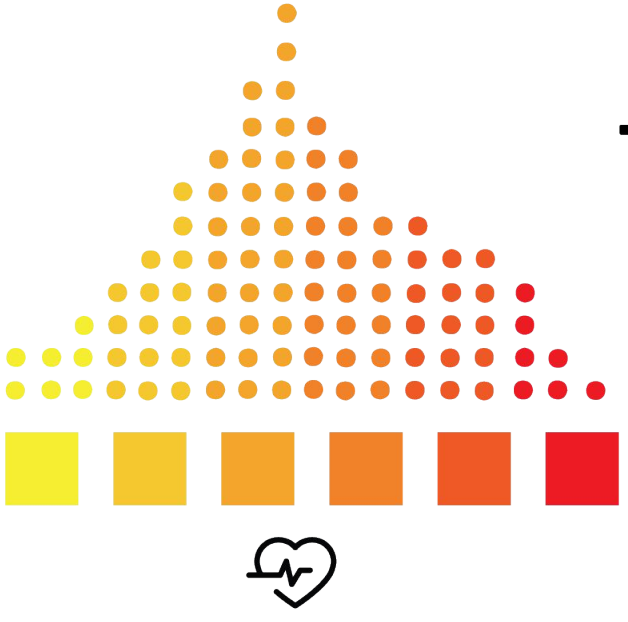


**We do this using** **mutual information** **:**

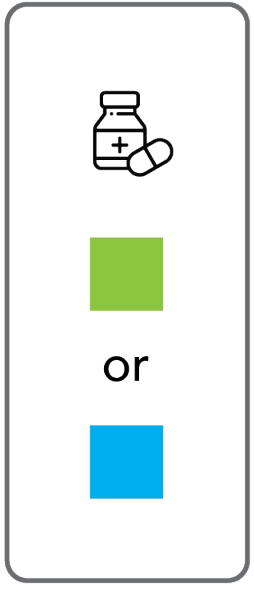
a measure of how much we learn

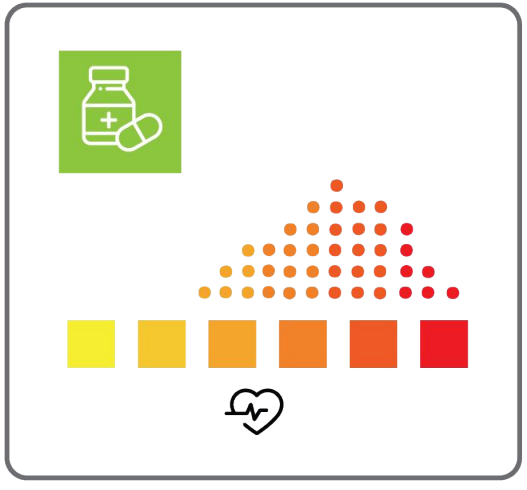
about one feature

from another

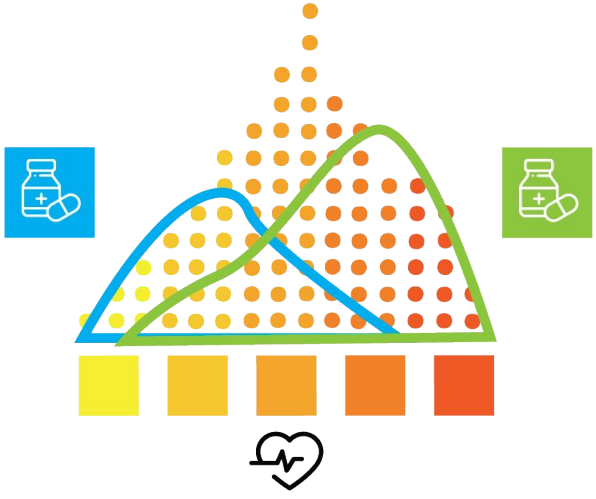


+

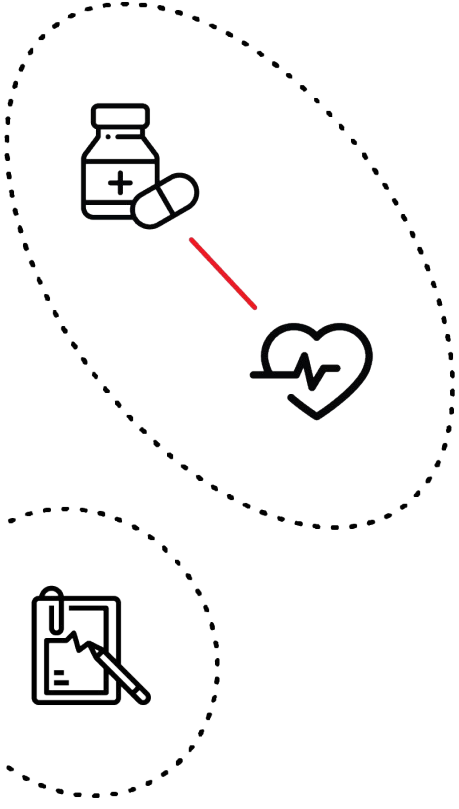


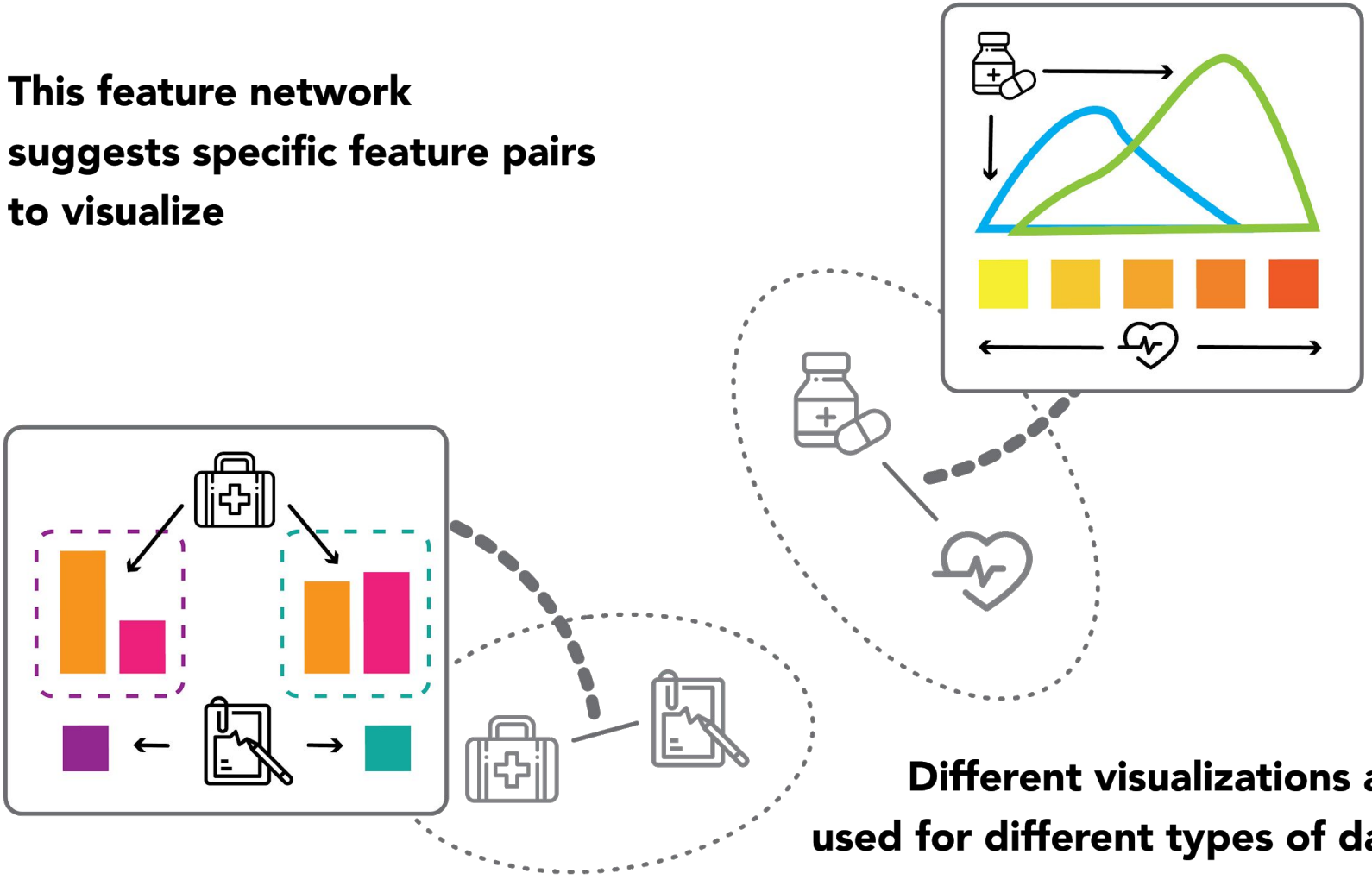
If a feature pair has high mutual information, we draw an edge between them in the graph



Feature pairs with low mutual information are not connected

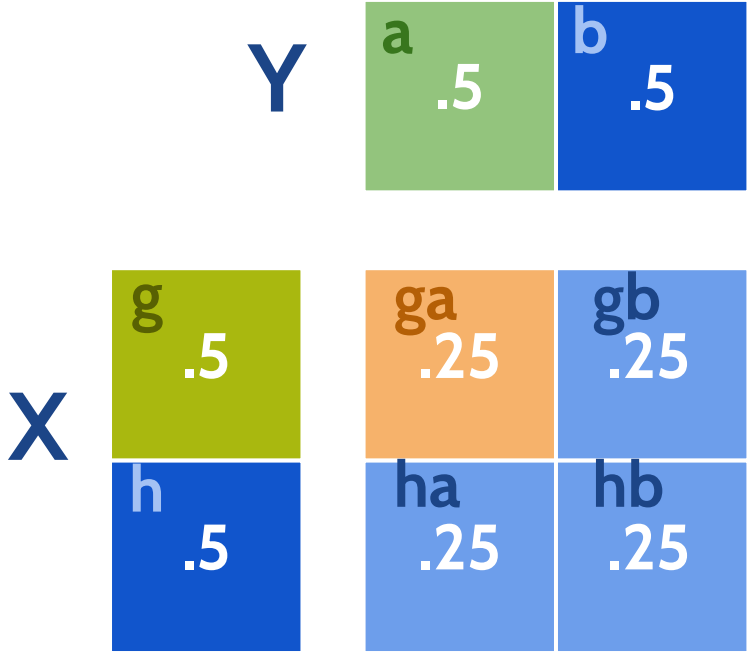


This feature network suggests specific feature pairs to visualize

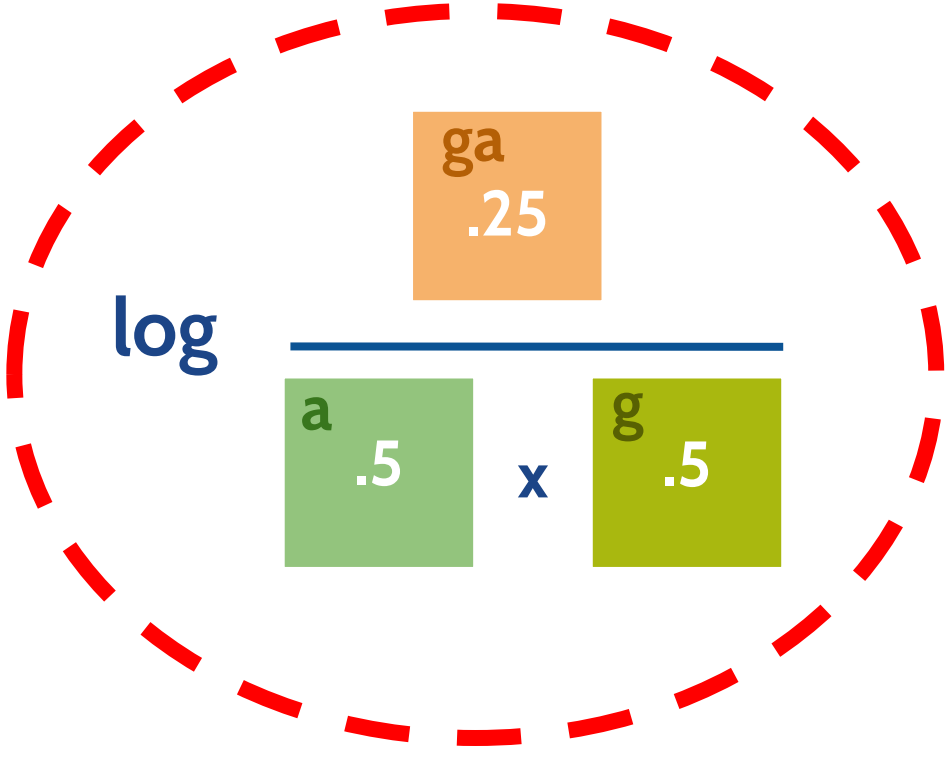


Different visualizations are used for different types of data

# Pointwise Mutual Information




## Discrete vs. Discrete

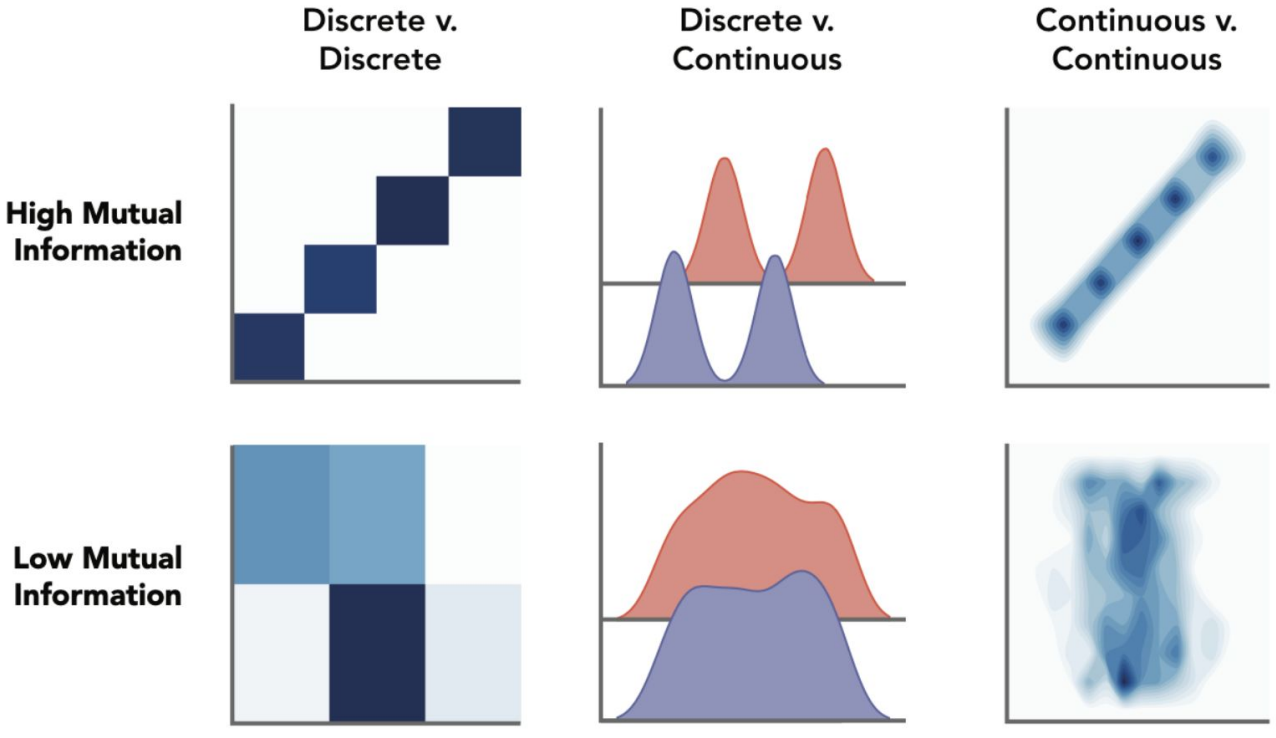


# Discrete Mutual Information

# Continuous Mutual Information

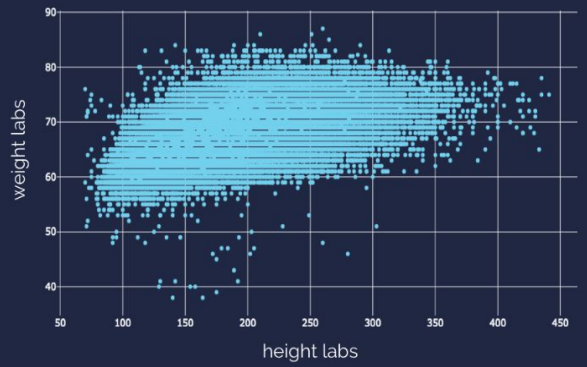
$$\int_{X:\ i_{min}-i_{max}} \int_{Y:\ j_{min}-j_{max}} XY \times \log \frac{XY}{X \times Y}$$

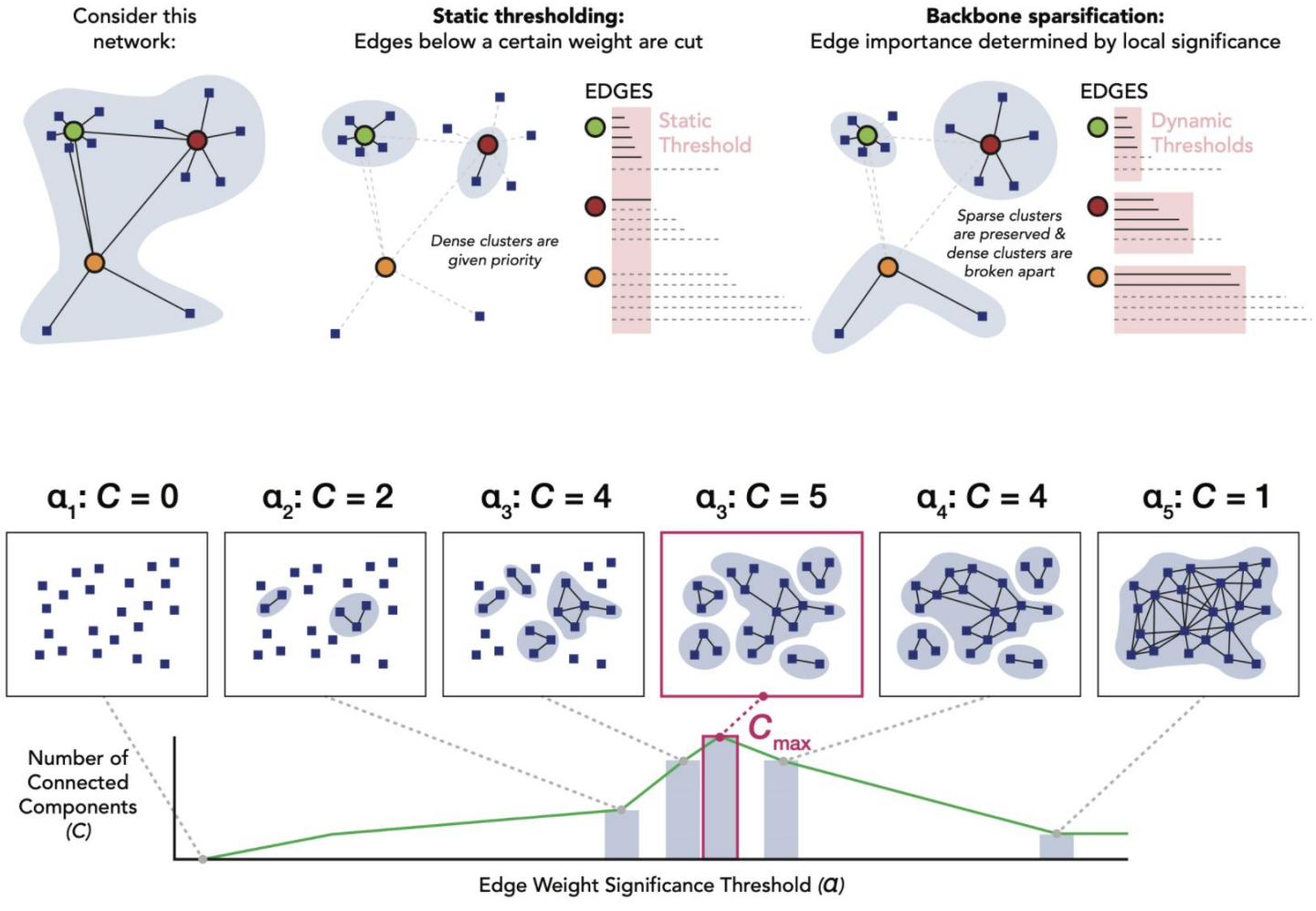
# Mutual Information

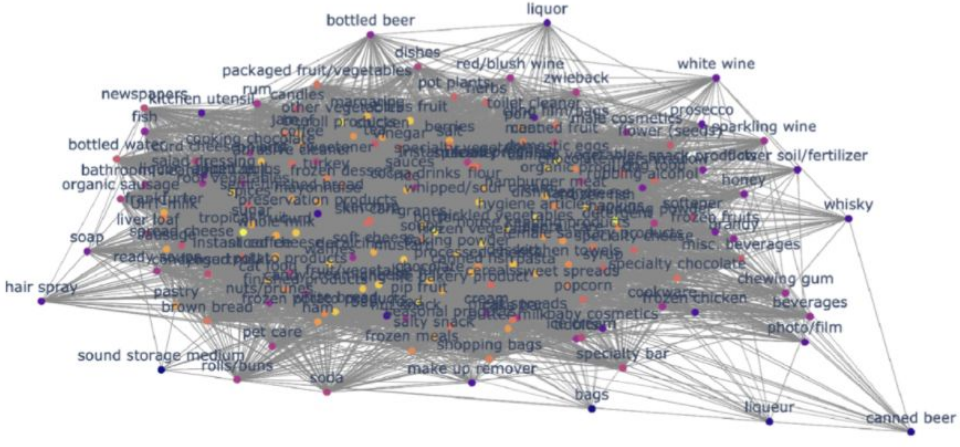


## Edge Details

### HEIGHT LABS VS WEIGHT LABS

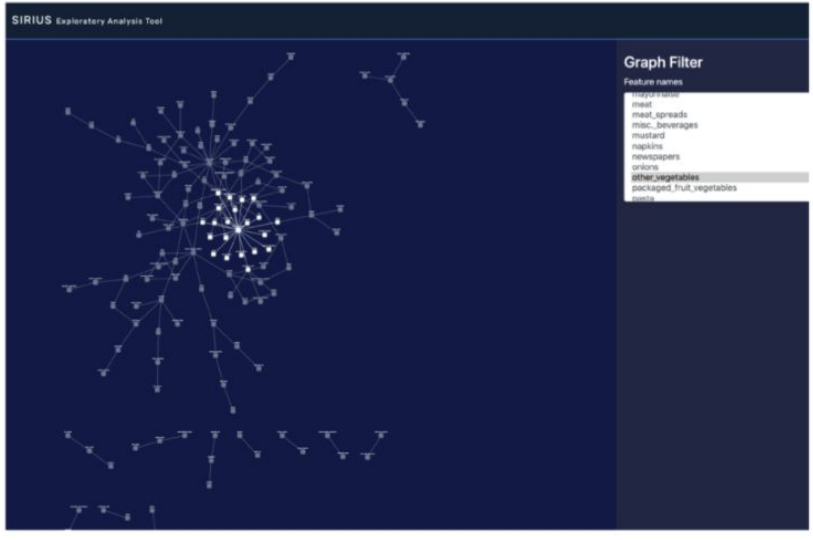



Consider this network:
Static thresholding:
Edges below a certain weight are cut
Dense clusters are given priority
EDGES
Static Threshold
Backbone sparsification:
Edge importance determined by local significance
Sparse clusters are preserved & dense clusters are broken apart
EDGES
Dynamic Thresholds
α1: C = 0
α2: C = 2
α3: C = 4
α3: C = 5
α4: C = 4
α5: C = 1
Cmax
Number of Connected Components (C)
Edge Weight Significance Threshold (α)




**Association Network**

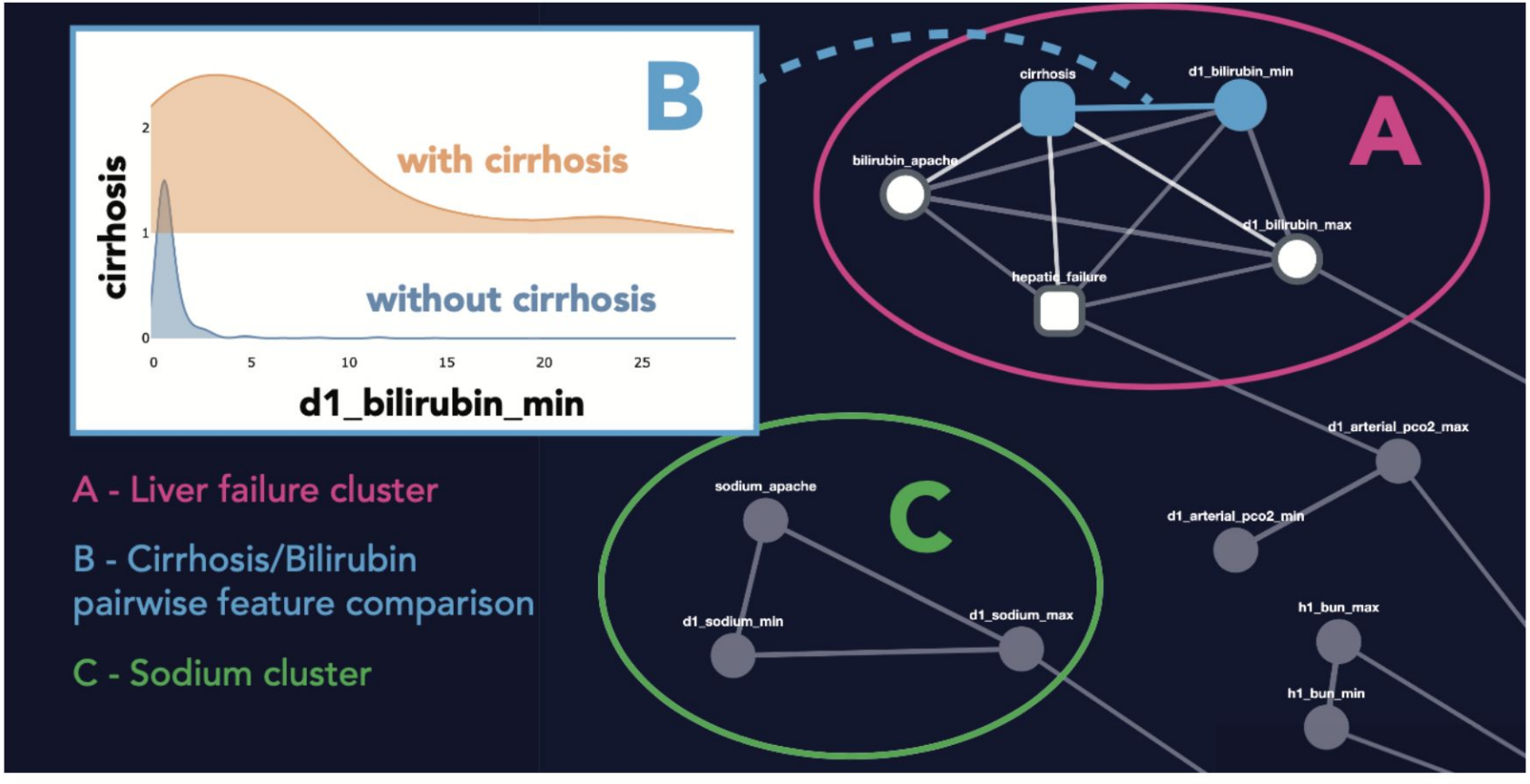
*Lift thresheld at static value, such that 50% of edges remain*



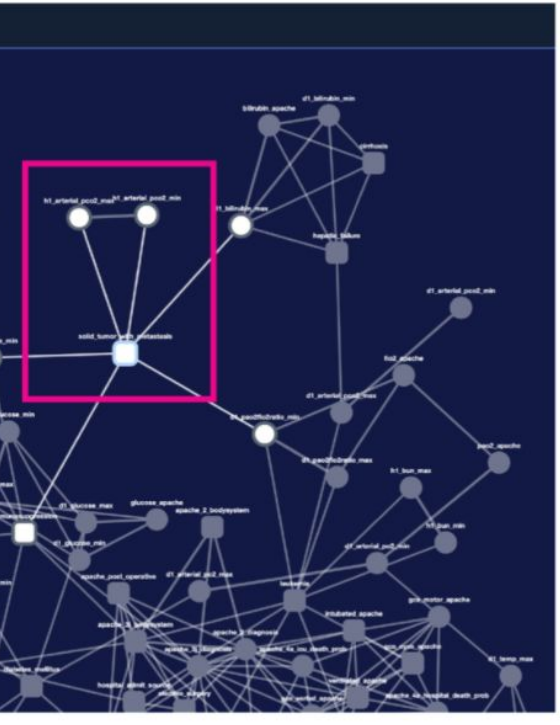

**Sirius Mutual Information Network**

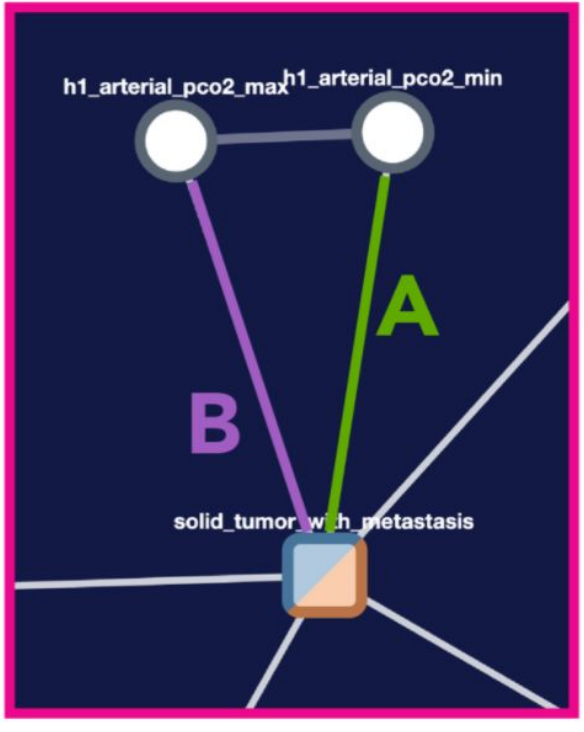
*Mutual information thresheld dynamically using alpha value from backbone sparsification*

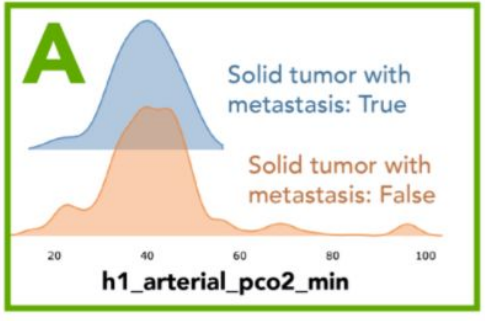
# Returning to Pairwise Comparisons

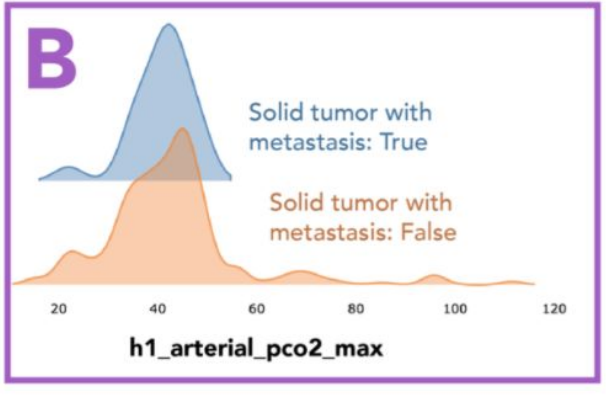


# Returning to Pairwise Comparisons

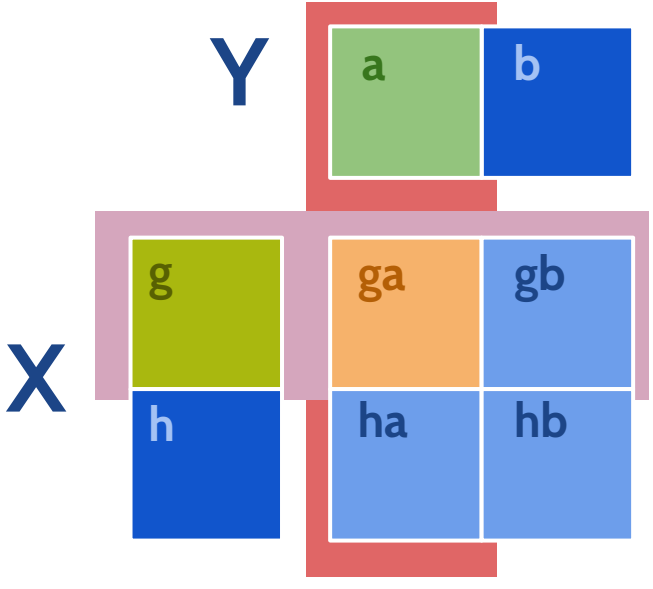


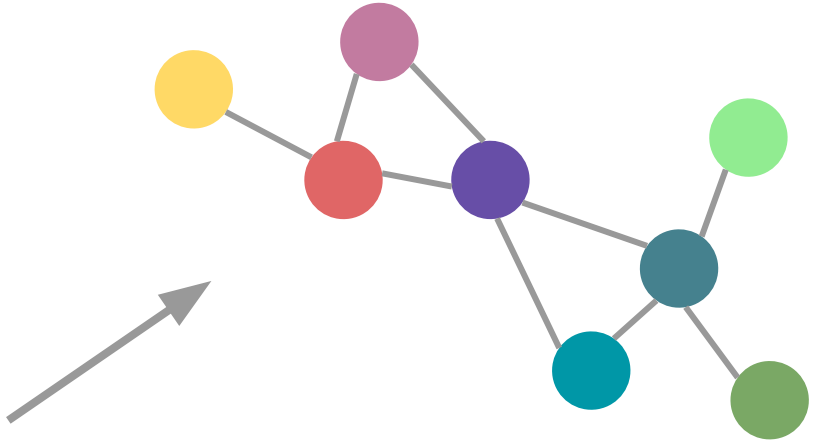


# Thanks!



twitter:
@artistjaneadams