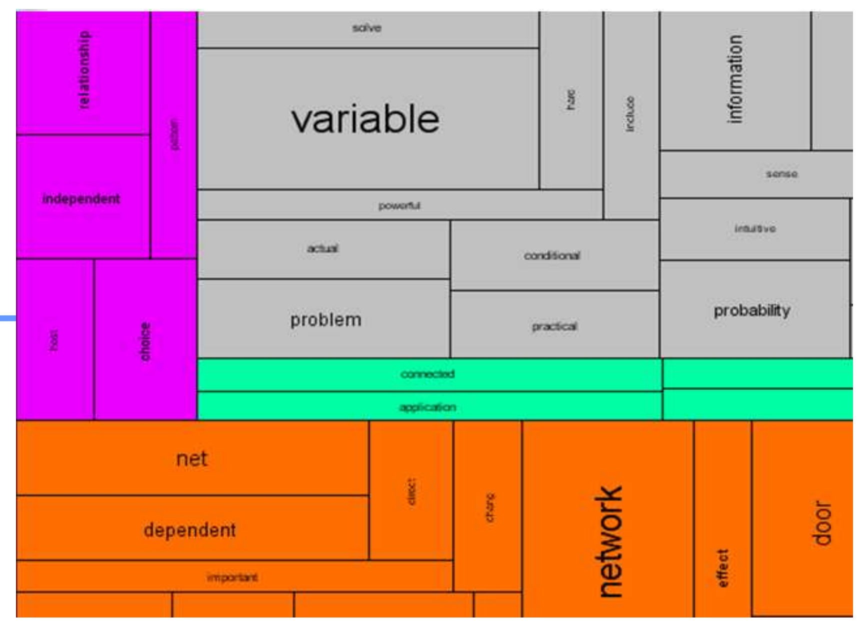
Text analytics

Untangling meanings in unstructured statements



Steven Struhl

Converge Analytic

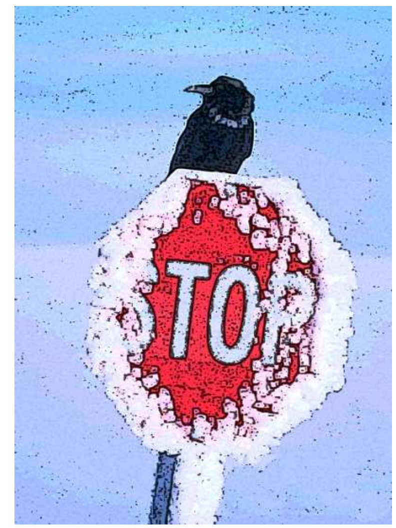
What is text analytics?

- Text analytics can be divided into two types of activities
  - **Predictive** or **model-based**
    - Text becomes a set of predictor variables used in a model
    - Models can have as the target variable (e.g.) overall ratings, use or purchases
  - **Descriptive** or **enumerative**
    - Probably the most common type of text analytics
    - Looks for frequencies of word groups, associations of words, proximities of words, etc.
- **Sentiment analysis** falls somewhere between these two
  - Some words or phrases are given a negative or positive valence
    - These positives and negatives are counted and a total score of positive or negative or a **sentiment** score is derived

Prefers the predictive

Going about text analytics: start with a collection of words

- We begin with a document—or unstructured collections of words
  - First, **stop words** (the, of, and, a, to,...) must be removed
    - Frequency of these words is so large that they can swamp the analysis
  - However, stop words should not be filtered when analyzing frequent phrases
    - Phrases can help to identify writers and provide other stylistic information
- Next words must be made **regular**
  - Spelling errors need to be corrected using a dictionary
  - Plurals must be singularized
  - Idioms need to be resolved
  - Tenses need to be made uniform so that the same word does not get diluted over minor variations
    - This is sometimes called **stemming**
- We may also look for **word pairs** (e.g., “not good” or “not bad”)
- Then we can begin



Not our type of stop

Predictive word analysis

- This could be carried out on individual words or phrases
  - But this gains the most power after the data has been **coded,** just as we would do for any questionnaire
- Models may need to eliminate variables as well as find relationships
- Two demonstrations—
  - Bayes Nets selecting variables and building a strong predictive model of intent to continue
  - Classification trees (CHAID) showing strong relationships between verbatim comments and top box overall ratings
- Both examples start with coded answers

Now for our next demonstration . . .

Bayes Nets find what is important and show how it fits together

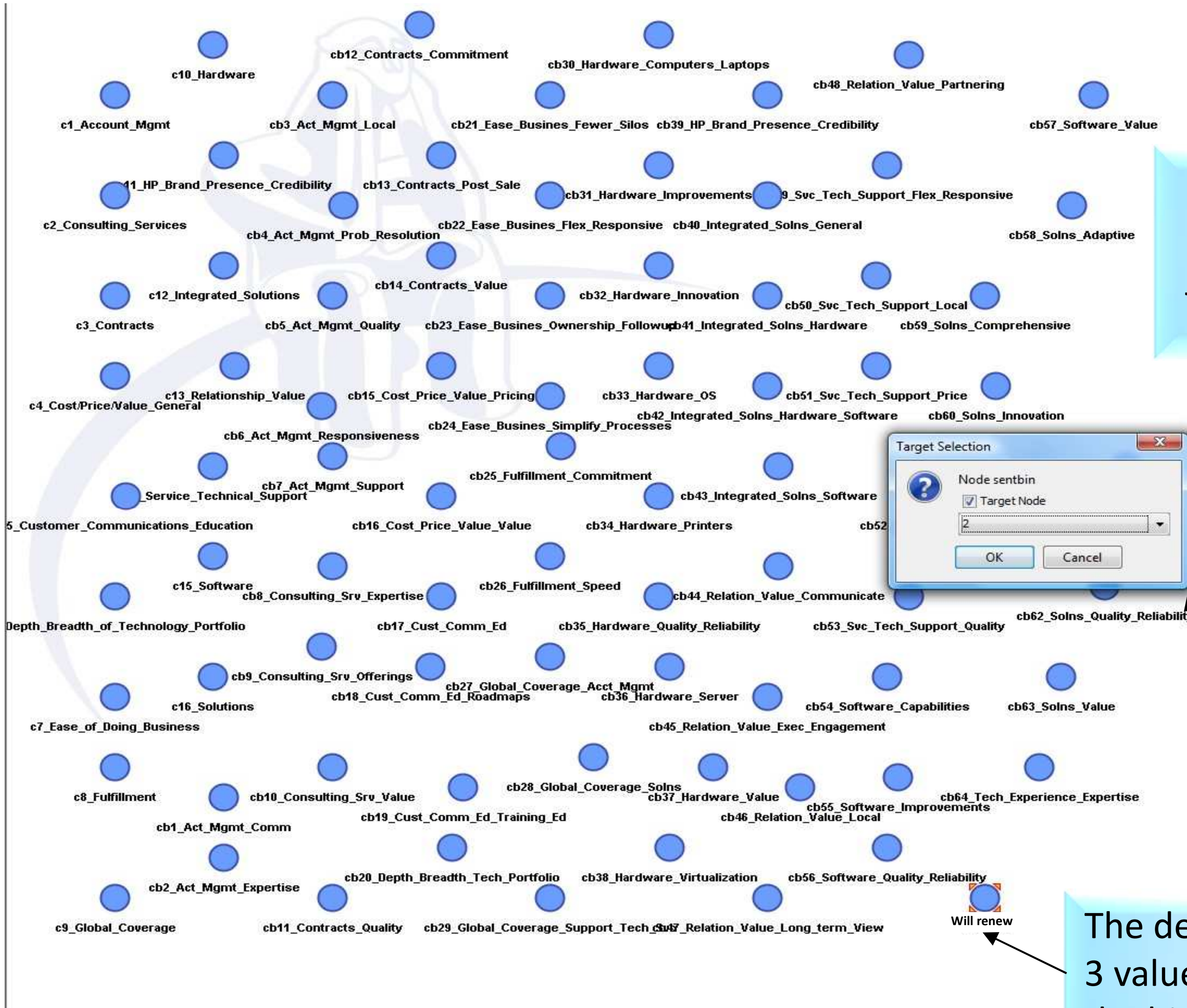
- Bayesian network are a remarkable new method discussed in more depth in another presentation[1] and an article[2]
- If you are familiar with structural equation models or PLS path models, these will look similar—variables and arrows
- However, they can largely be **self-constructed,** with data driving the patterns of connections
- Also, the rules are different
  - Even though arrows point in one direction between two variables, influence flows both ways
  - Any change in one part of the network **propagates** throughout the entire structure
  - The whole network is connected!
- Some terminology is required—
  - The variable at the start of an arrow is called a **parent**
  - The variable at the end is called a **child** of the parent
  - The parent node leads to (and can cause) the child node
    - Arrows can lead to or from a dependent variable
    - Children can have several parents and parents can have several children

The variables keep connected

[1] "Bayes Nets: Harnessing their power"

[1] "Bayes Nets: Understanding the best newest thing"

We start with 79 coded verbatim comments as independents and “will renew” as the dependent



Nothing connected yet: just the variables to be considered

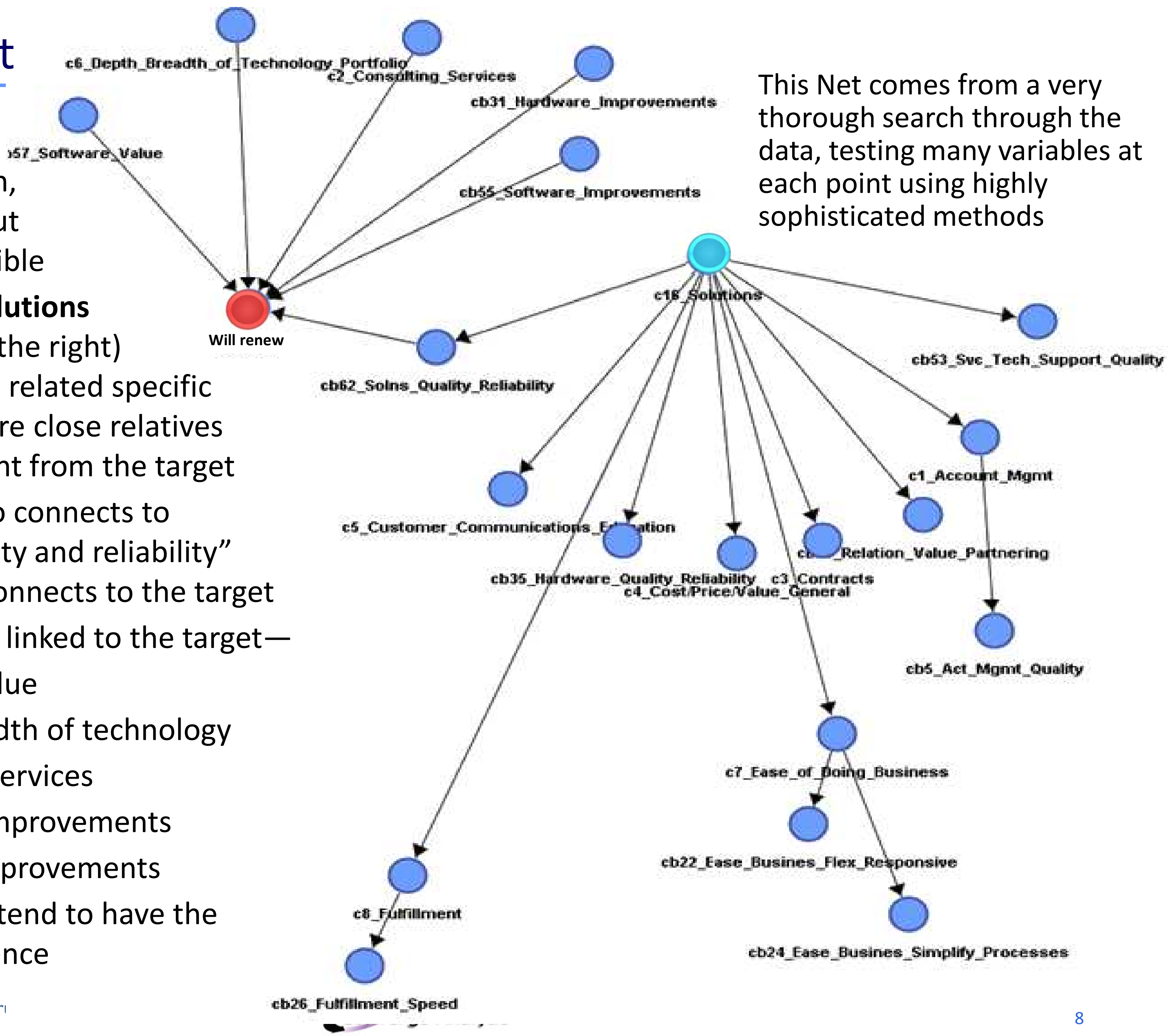
The dependent takes 3 values (0, 1, 2), with the highest the target

Isolating variables that belong (the Markov blanket)

- The **Markov blanket** comes from a search for strong connections—sifting through the variables many times, leaving just parents and children of the target—and co-parents of any child
- Down to 18 variables!
- Note that the arrows show us **unsupervised** directions chosen by the data
- Directions really do not matter unless we are seeking to find true **causation**—not possible with most data we will ever see
- Our dependent could be viewed as the parent rather than the child of most of these variables
  - That is, the independents **explain the target**, the way independent variables work in a regression

Best Bayes Net

- This data-driven, automatic layout looks very sensible
- The variable **solutions** (highlighted to the right) leads to several related specific variables that are close relatives and more distant from the target
- "Solutions" also connects to "solutions quality and reliability" which in turn connects to the target
- Other variables linked to the target—
  - Software value
  - Depth/breadth of technology
  - Consulting services
  - Hardware improvements
  - Software improvements
- Direct linkages tend to have the strongest influence



This Net comes from a very thorough search through the data, testing many variables at each point using highly sophisticated methods

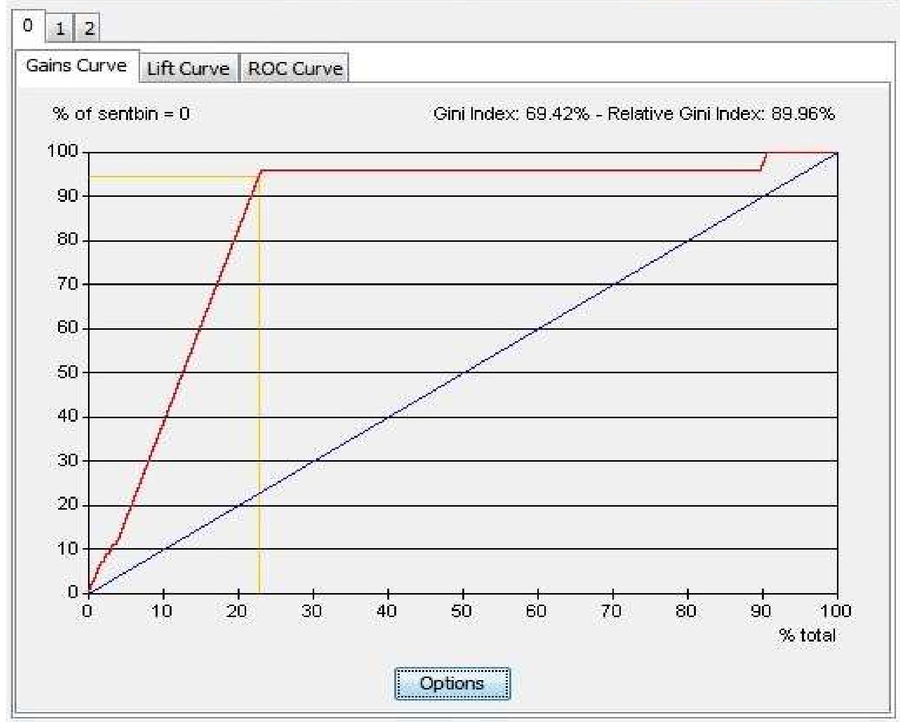
The network performed remarkably well

- Correct classification of the two extreme states (yes and no): 66% and 52%
- Nobody fell in the middle—nothing to predict there
- This level of prediction is remarkable for just open-ended responses
- The curve (right) compares how well the model got the positives right (true positives) vs. predicting a negative as a positive (false positives)
- The lower line is chance, so curve area above the line is improvement
  - This shows a strong level of improvement

Confusion Matrix

Occurrences | Reliability | Precision

Value	0 (4377)	1 (5142)	2 (9651)
0 (726)	66.25%	16.12%	17.63%
1 (0)	0%	0%	0%
2 (18444)	21.12%	27.24%	51.63%



The Net also calculates how much influence each variable has

Relative weights

Relationship Analysis

Parent	Child	Kullback-Leibler Divergence	Relative Weight	Global Contribution	Mutual information	G_{KL}-test	Degrees of Freedom	p-value	G-test (Data)	Degrees of Freedom (Data)	p-value (Data)
c7_Ease_of_Doing_Business	cb22_Ease_Busines_Flex_Responsive	0.340568	1.0000	35.77%	0.340568	9050.6803	1	0.00%	9050.6807	1	0.00%
c16_Solutions	cb62_Solns_Quality_Reliability	0.262175	0.7698	27.54%	0.262175	6967.3607	1	0.00%	6967.3613	1	0.00%
c8_Fulfillment	cb26_Fulfillment_Speed	0.099972	0.2935	10.50%	0.099972	2656.7809	1	0.00%	2656.7808	1	0.00%
c1_Account_Mgmt	cb5_Act_Mgmt_Quality	0.046354	0.1361	4.87%	0.046354	1231.8614	1	0.00%	1231.8662	1	0.00%
c7_Ease_of_Doing_Business	cb24_Ease_Busines_Simplify_Processes	0.029462	0.0865	3.09%	0.029462	782.9593	1	0.00%	782.9558	1	0.00%
c16_Solutions	c7_Ease_of_Doing_Business	0.028148	0.0827	2.96%	0.028148	748.0455	1	0.00%	748.0437	1	0.00%
c16_Solutions	c4_Cost/Price/Value_General	0.020191	0.0593	2.12%	0.020191	536.5712	1	0.00%	536.5706	1	0.00%
cb62_Solns_Quality_Reliability	sentbin	0.017928	0.0526	1.88%	0.014812	476.4454	64	0.00%	472.2646	2	0.00%
c16_Solutions	cb48_Relation_Value_Partnering	0.016835	0.0494	1.77%	0.016835	447.4034	1	0.00%	447.4045	1	0.00%
c16_Solutions	cb53_Svc_Tech_Support_Quality	0.015752	0.0463	1.65%	0.015752	418.6092	1	0.00%	418.6093	1	0.00%
c16_Solutions	c1_Account_Mgmt	0.012180	0.0358	1.28%	0.012180	323.6777	1	0.00%	323.6803	1	0.00%
cb31_Hardware_Improvements	sentbin	0.011515	0.0338	1.21%	0.010302	306.0188	64	0.00%	331.9024	2	0.00%
c16_Solutions	c8_Fulfillment	0.009495	0.0279	1.00%	0.009495	252.3322	1	0.00%	252.3319	1	0.00%
c16_Solutions	c5_Customer_Communications_Education	0.008832	0.0259	0.93%	0.008832	234.7143	1	0.00%	234.7149	1	0.00%
cb55_Software_Improvements	sentbin	0.008185	0.0240	0.86%	0.007184	217.5212	64	0.00%	236.4934	2	0.00%
c16_Solutions	cb35_Hardware_Quality_Reliability	0.007297	0.0214	0.77%	0.007297	193.9191	1	0.00%	193.9195	1	0.00%
c16_Solutions	c3_Contracts	0.005885	0.0173	0.62%	0.005885	156.4030	1	0.00%	156.4038	1	0.00%
c6_Depth_Breadth_of_Technology_Portfolio	sentbin	0.005083	0.0149	0.53%	0.002914	135.0694	64	0.00%	121.3369	2	0.00%
c2_Consulting_Services	sentbin	0.003226	0.0095	0.34%	0.002472	85.7265	64	3.63%	78.6077	2	0.00%

Close Save As... Print

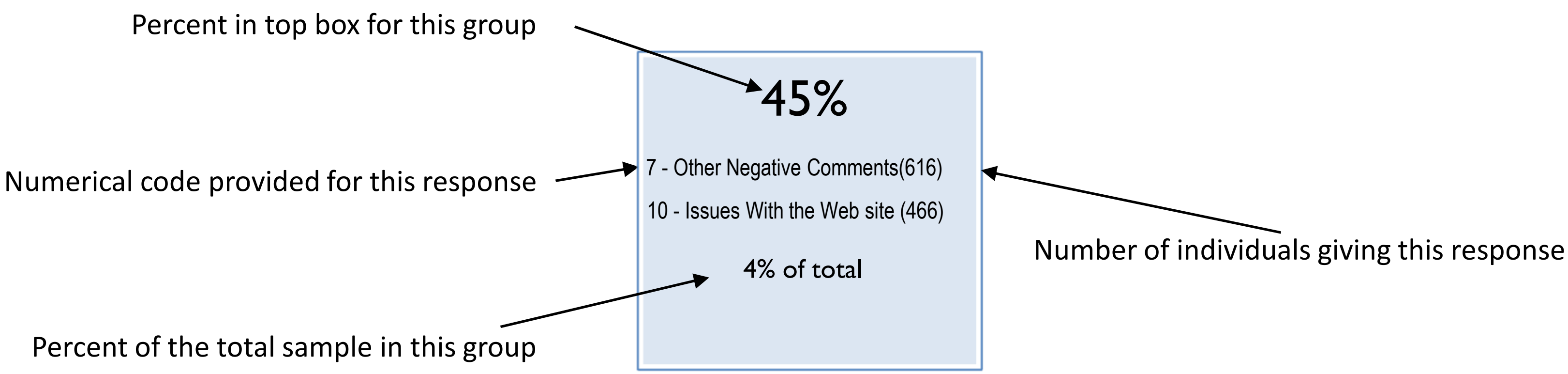
Example with classification trees (CHAID)

- CHAID provides a great deal of valuable information, but works differently from Bayes Nets
- While CHAID also predicts patterns in a dependent, it does so by grouping words or coded phrases
  - Each group will be associated with some level of a response
    - For instance, all the codes in one group will have on average 45% in the top box of ratings
- Most CHAID programs do not try to calculate variables' importances
  - This makes sense since variables are being grouped
- We will see how this works in the next slides . . .

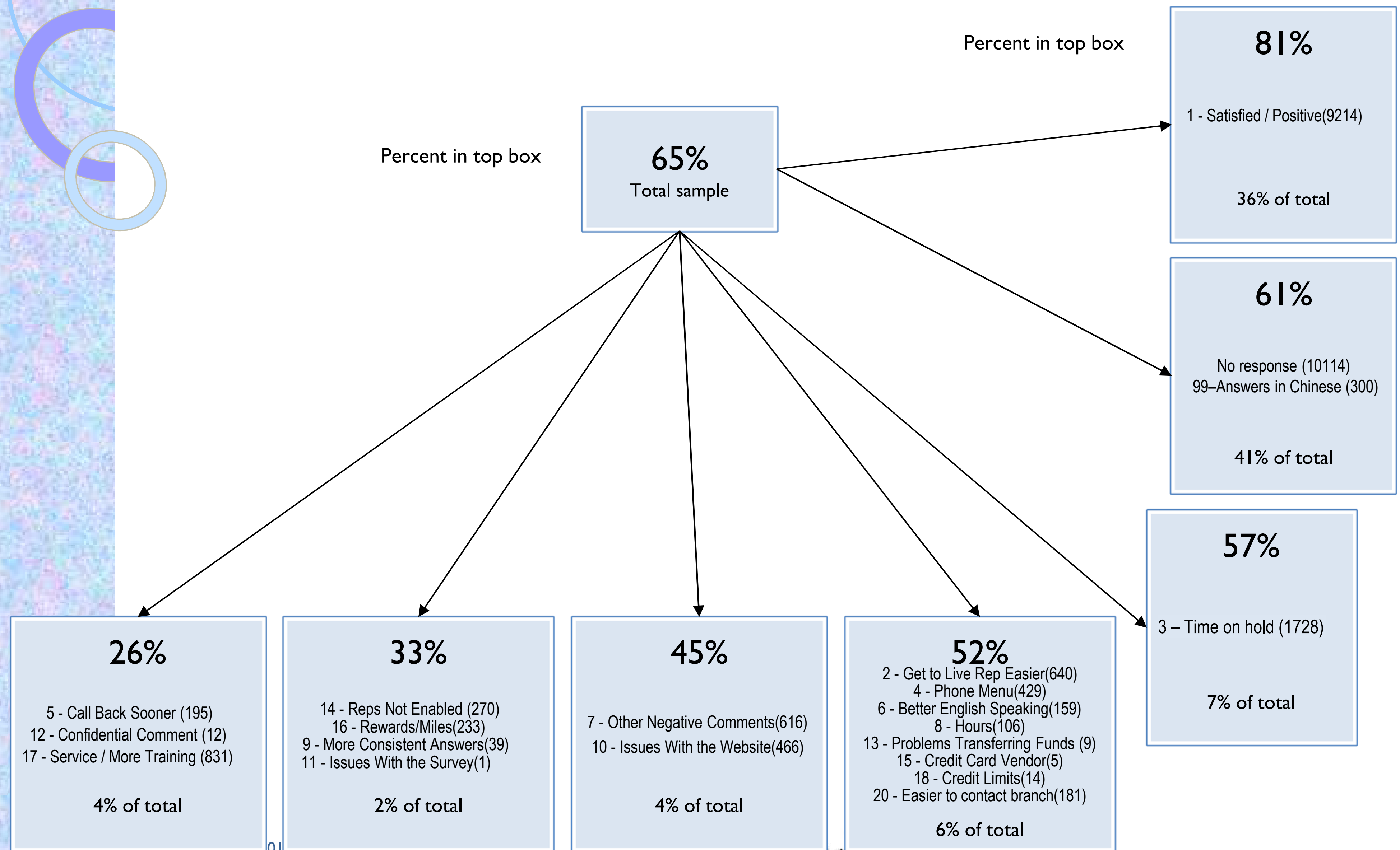
CHAID had nothing whatsoever to do with forming this group

How CHAID works: optimal recoding

- CHAID forms groups of open-ended responses that have statistically equal levels of positive responses
  - It examines every possible way of grouping and finds the way that produces the strongest statistical difference
    - With 30 codes this is **millions** of possible ways
- This powerful ability is called **optimal recoding**
- Here CHAID found seven groups of responses
  - In top box scores, these range from 81% down to 26%
    - A person in the least likely group (26% top box) is **less than 1/3 as likely** to be completely satisfied as one in the most likely group (81%)
- Even among the relatively small groups with lower levels of satisfaction, CHAID quickly uncovered the verbatim comments behind their ratings
- Here is the information shown for each group—



How verbatim codes align with percent top box in overall ratings



Descriptive text analytics

Types of descriptive text analysis

- The Bag–of–Words Model
  - Syntax is irrelevant
  - A collection of words is analyzed without regard to order or grammar
  - Analyses—
    - Distributions of words
    - Distributions compared to known distributions
    - Derived measures of importance from the most frequent words

Not our bag (of words)

- The Sequential Model
  - A search for words occurring near each other in the document
    - Analyses
      - A popular method is searching **nGrams**
        - Sequences of 3, 4, 5, etc. words
        - Longer nGrams are similar to phrases in ordinary writing
      - Finding the similarity of word pairs by counting how often each pair occurs in the same nGram
        - By running an **n–words window** through the document, stepping one word forward at a time, we compute the similarity of every possible pair of words

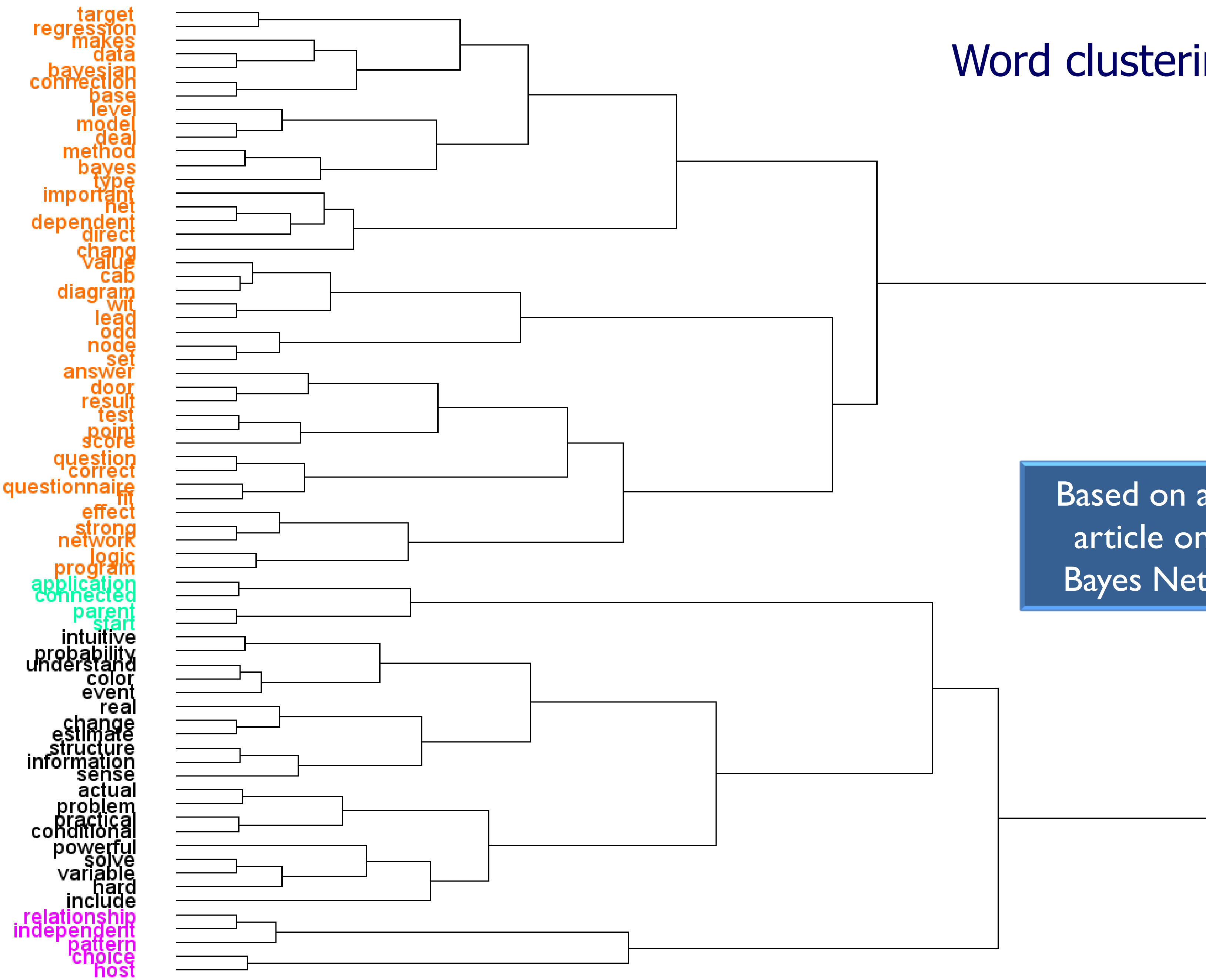
Descriptive analysis: Word clustering

- Creates a diagram where words that appear together most in the document are closest
- The lengths of lines in the diagram corresponds to how often the words were close to each other—shorter distance is more frequent
- This analysis clustered words via Ward's method[1]
- The distance measure is based on the square root of the number of times each pair of words appeared in a sliding **n-gram window**
  - Taking the square root normalizes the distribution of word frequencies, which has been shown to improve clustering
- Any leaves colored black do not belong to a cluster

[1] *For the statistically inclined, that's one of the more widely used hierarchical agglomerative methods*

Pretty enough but apparently these leaves are not members

Word clustering



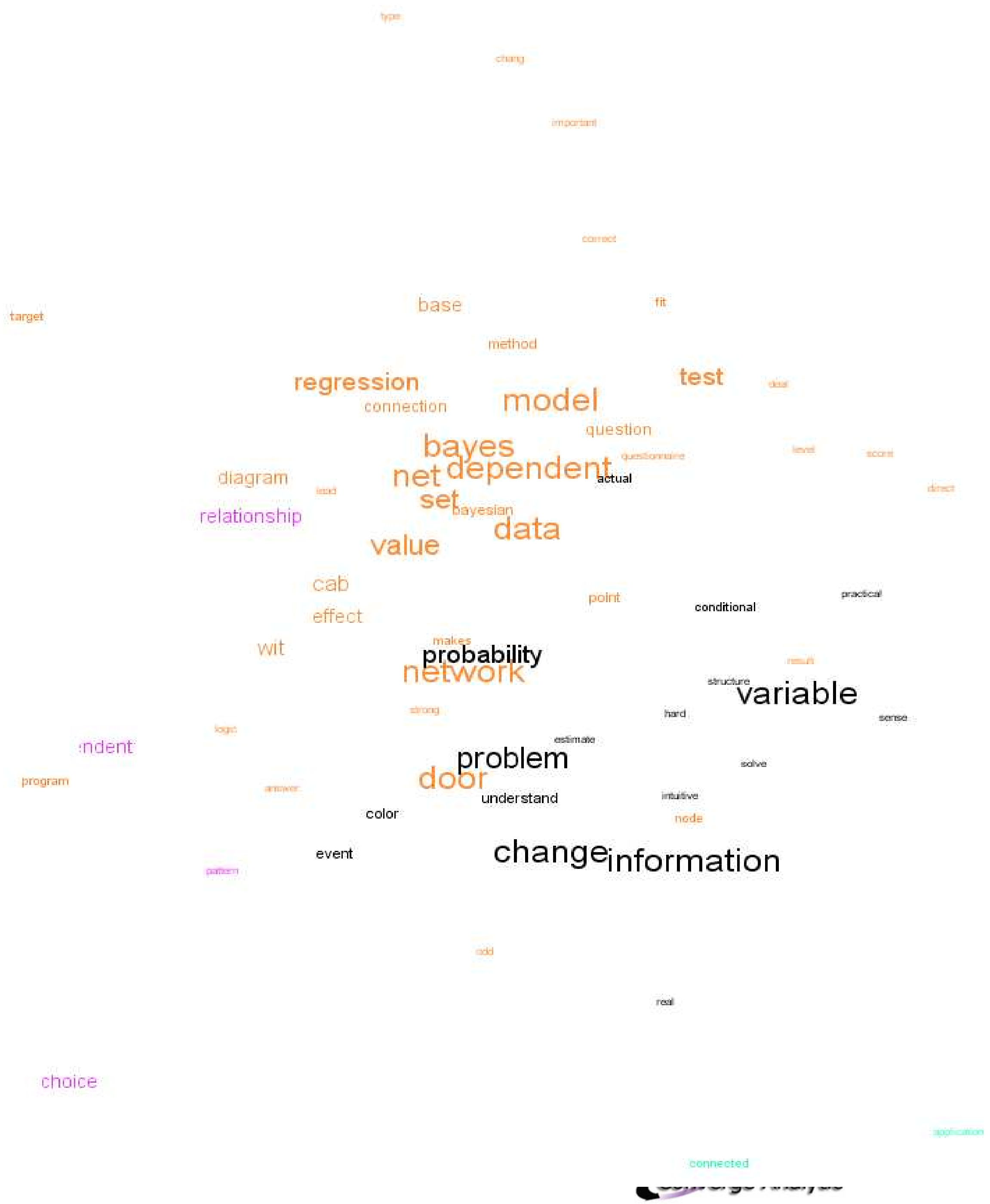
Based on an article on Bayes Nets

Making a word cloud

- A word cloud is a spatial diagram showing how often words occur near each other
  - Multidimensional scaling (MDS) creates a graph layout of the co–occurrences of words within a sliding n-gram window
  - The words also are sized according to the square root of their frequency of occurrence in the document
    - Once again, the square root transformation is used to normalize the distribution of frequencies, making the plot more coherent

It seems the Babylonians knew about square roots. With some luck, this presentation is more intelligible than the inscription

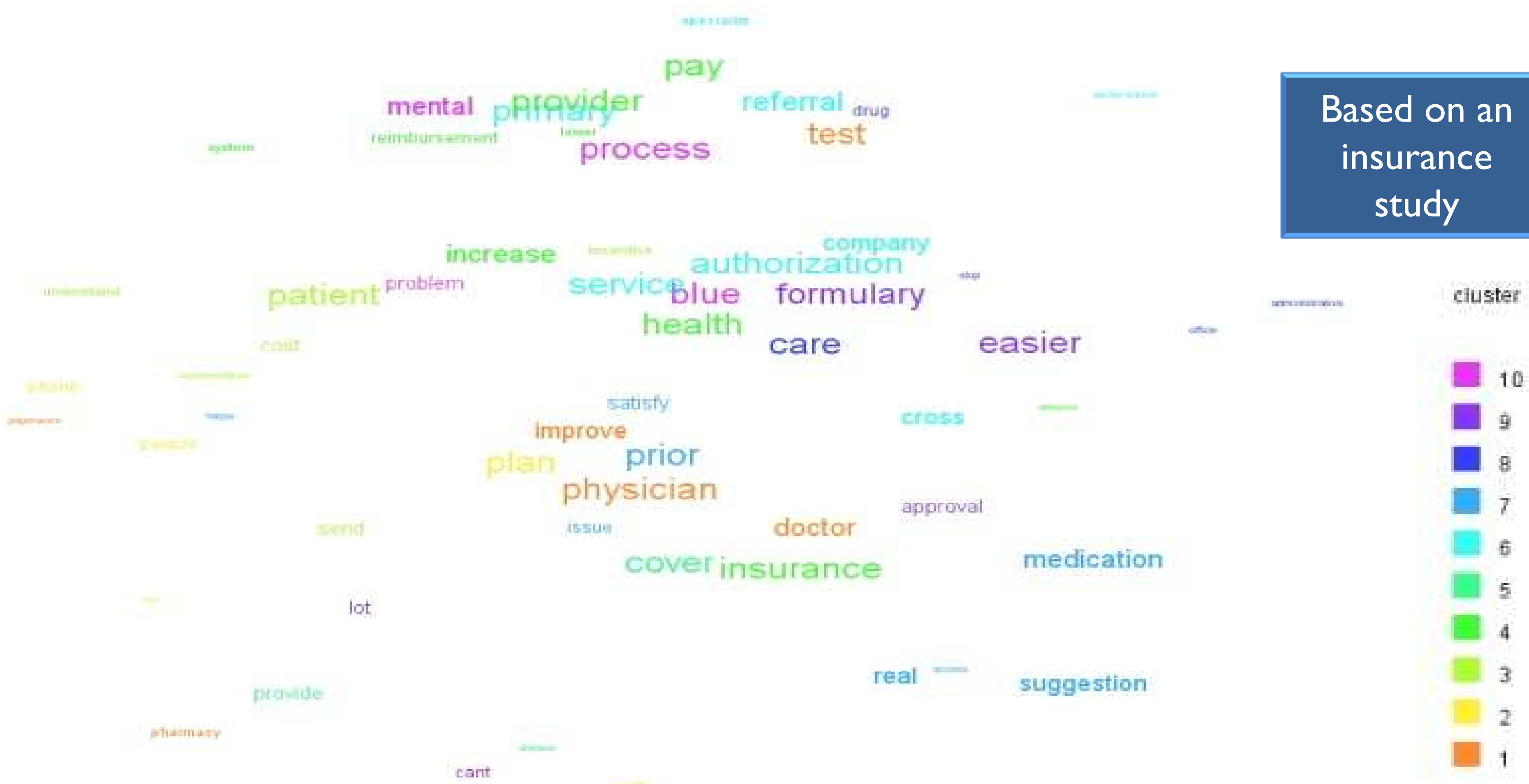
Word cloud



Based on an article on Bayes Nets

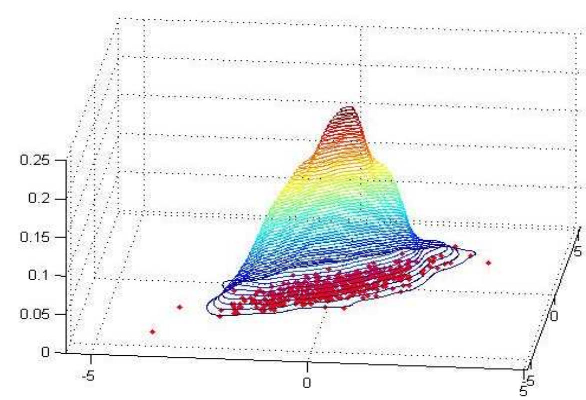
Word Cloud

Based on an insurance study



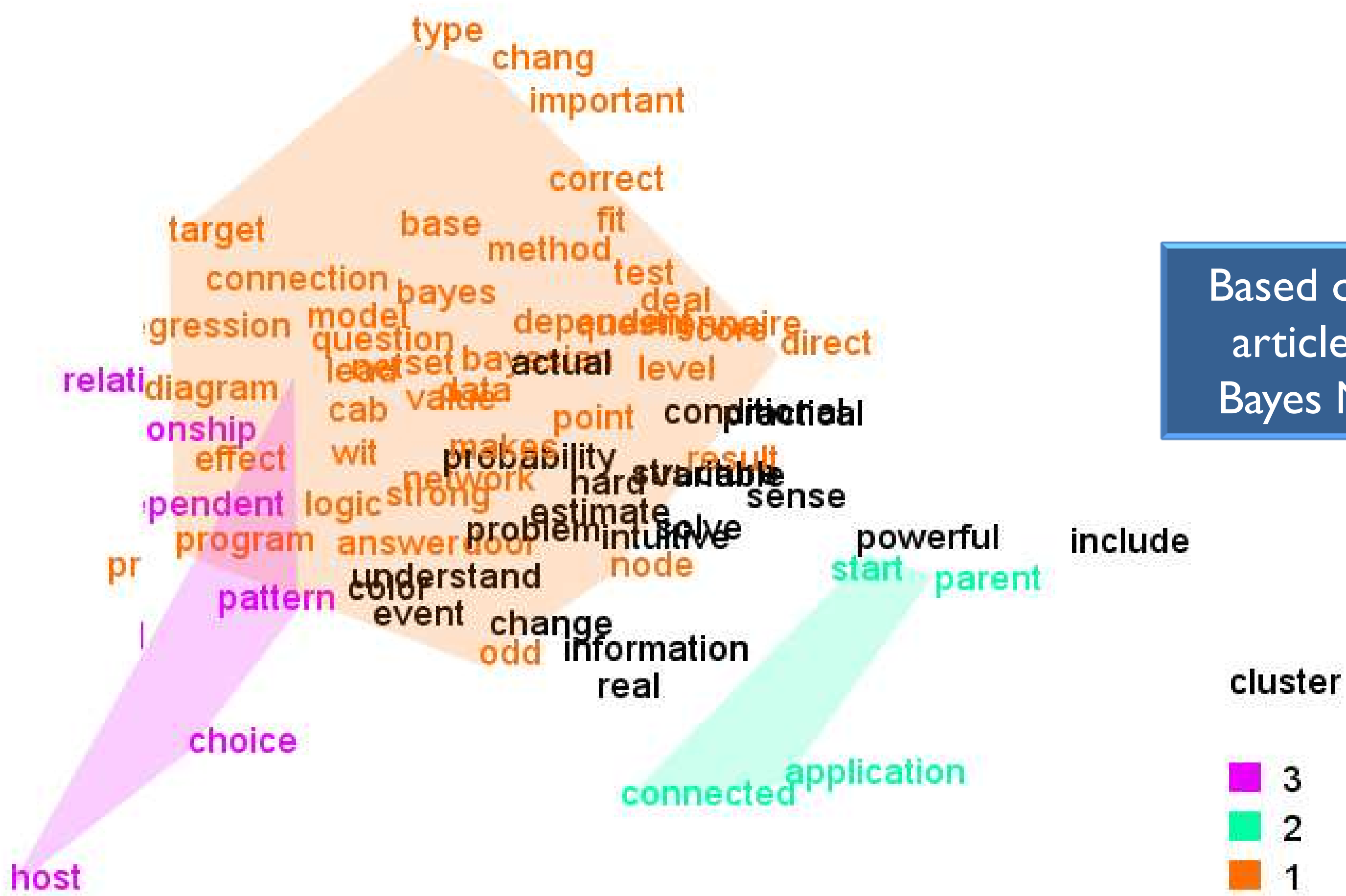
Word clouds with covers or boundaries

- We also can draw covers or boundaries around the word clouds, which may give a better idea of where the closest associations begin and end
- There are several types of covers, which can give different results
- There is no default or best type of cover
  - One type of cover is called a **convex hull**
  - An **alpha hull** may make a tighter cover on the points
    - In our example, these two types came out the same
- Another type, the **kernel density estimate,** may generate somewhat looser boundaries
- Examples of clouds with covers come from the article



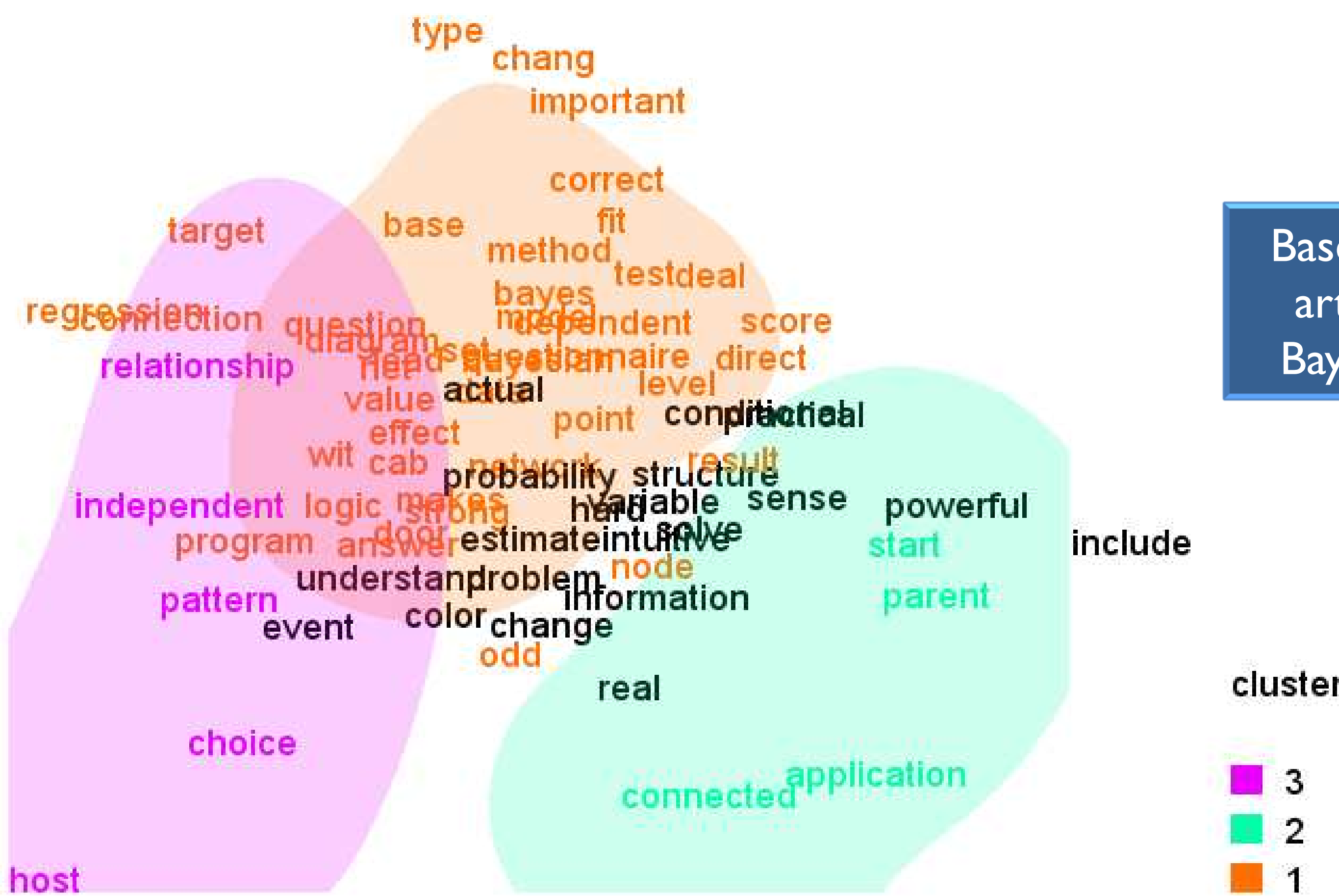
Kernel density: This should clear up everything

Word cloud with convex hull or Alpha hull cover



Based on an article on Bayes Nets

Word cloud with density kernels hull cover



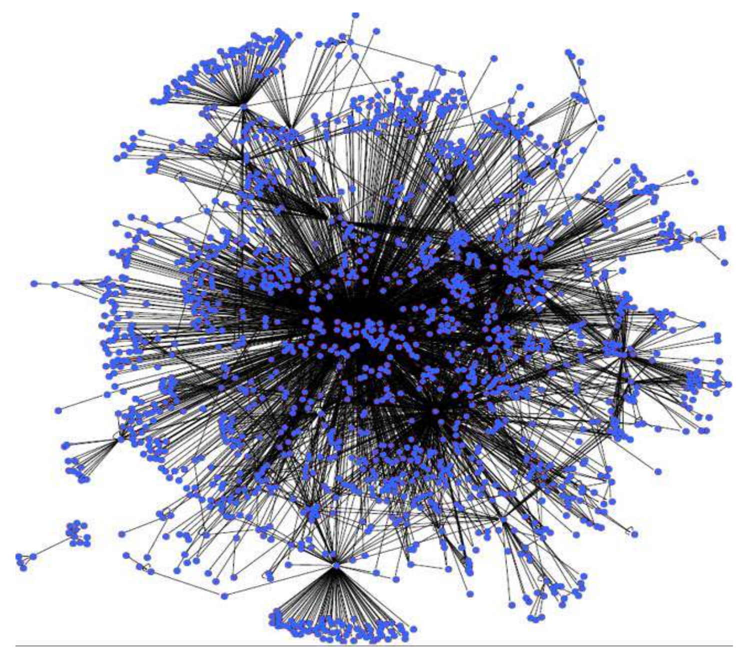
Based on an article on Bayes Nets

Converge Analytic

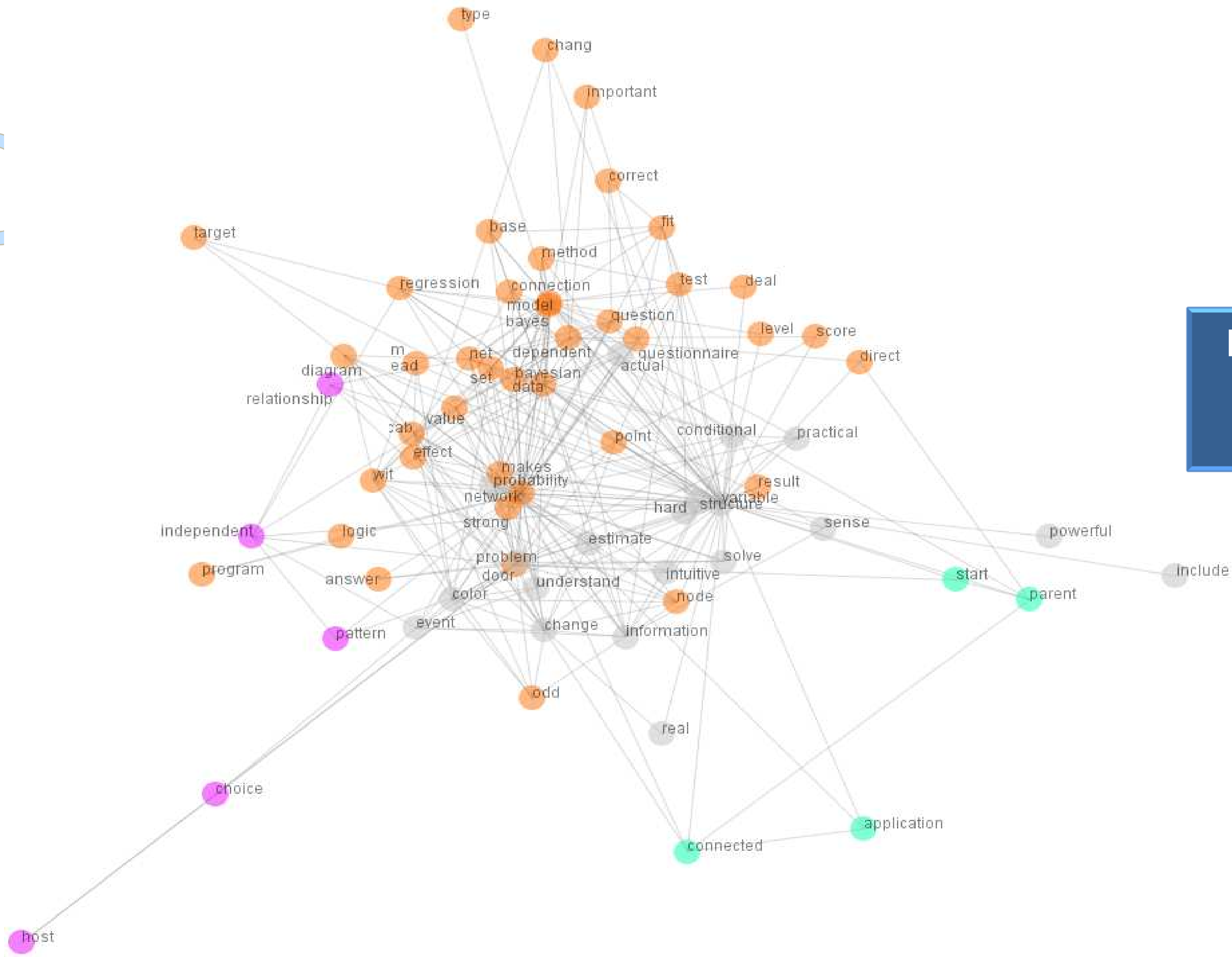
Graph layout of words: Clouds with something extra

- Information in this diagram is similar to that in the word cloud
- This looks somewhat different with words having **edges**, or connections to other words
- Words with a lot of edges have a high **degree**, meaning that they show up in connection with many other words in the document
- This has the possible advantage of looking like a network
  - It conveys some of the complexity of relationships among words
  - That also could be a disadvantage, as the tangle of lines may obscure some relationships

Not a graph layout, yet amazingly similar—the extreme complexity of E. coli's transcriptional regulatory network

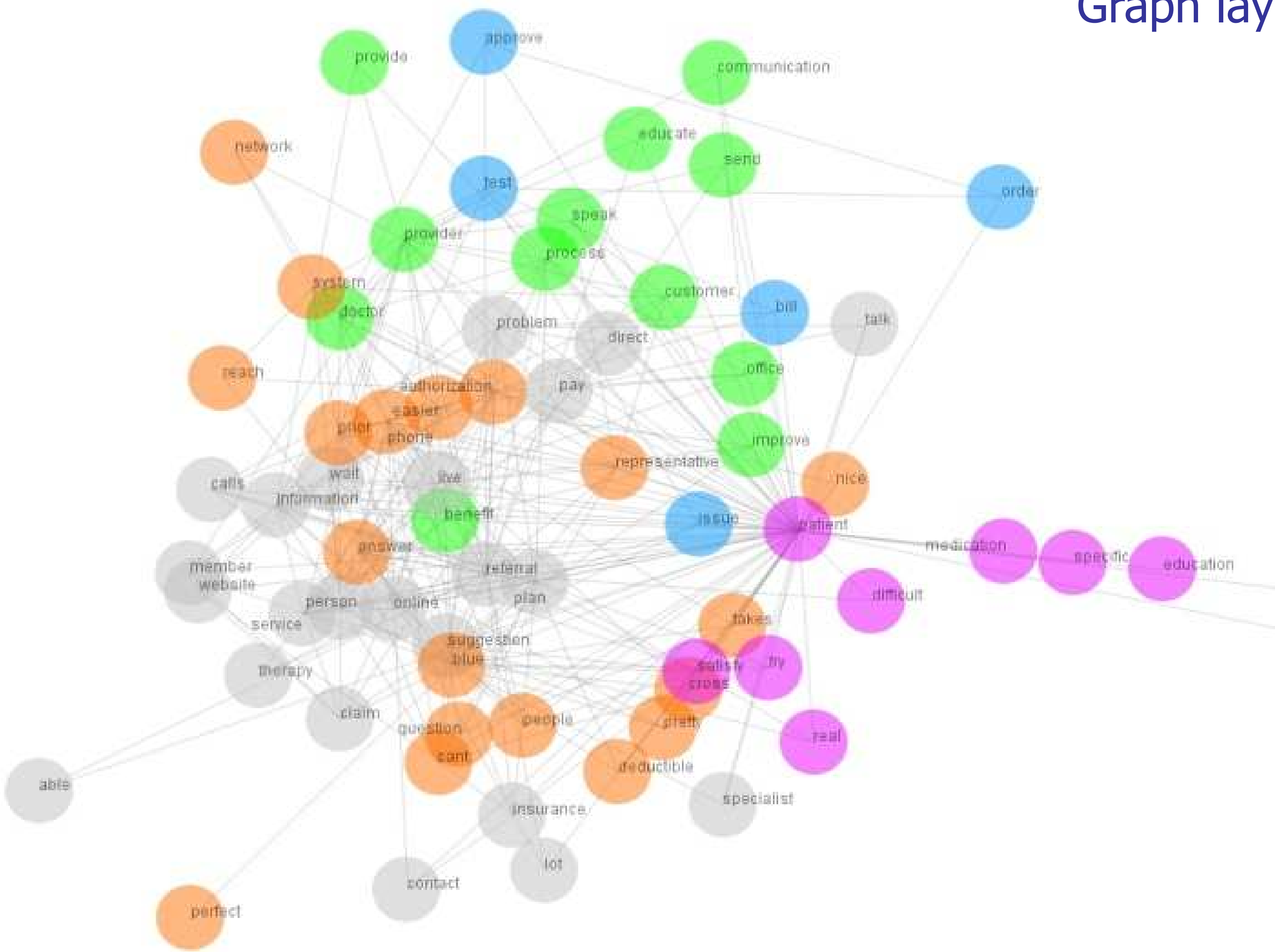


Graph layout of words



Based on an article on Bayes Nets

Graph layout of words

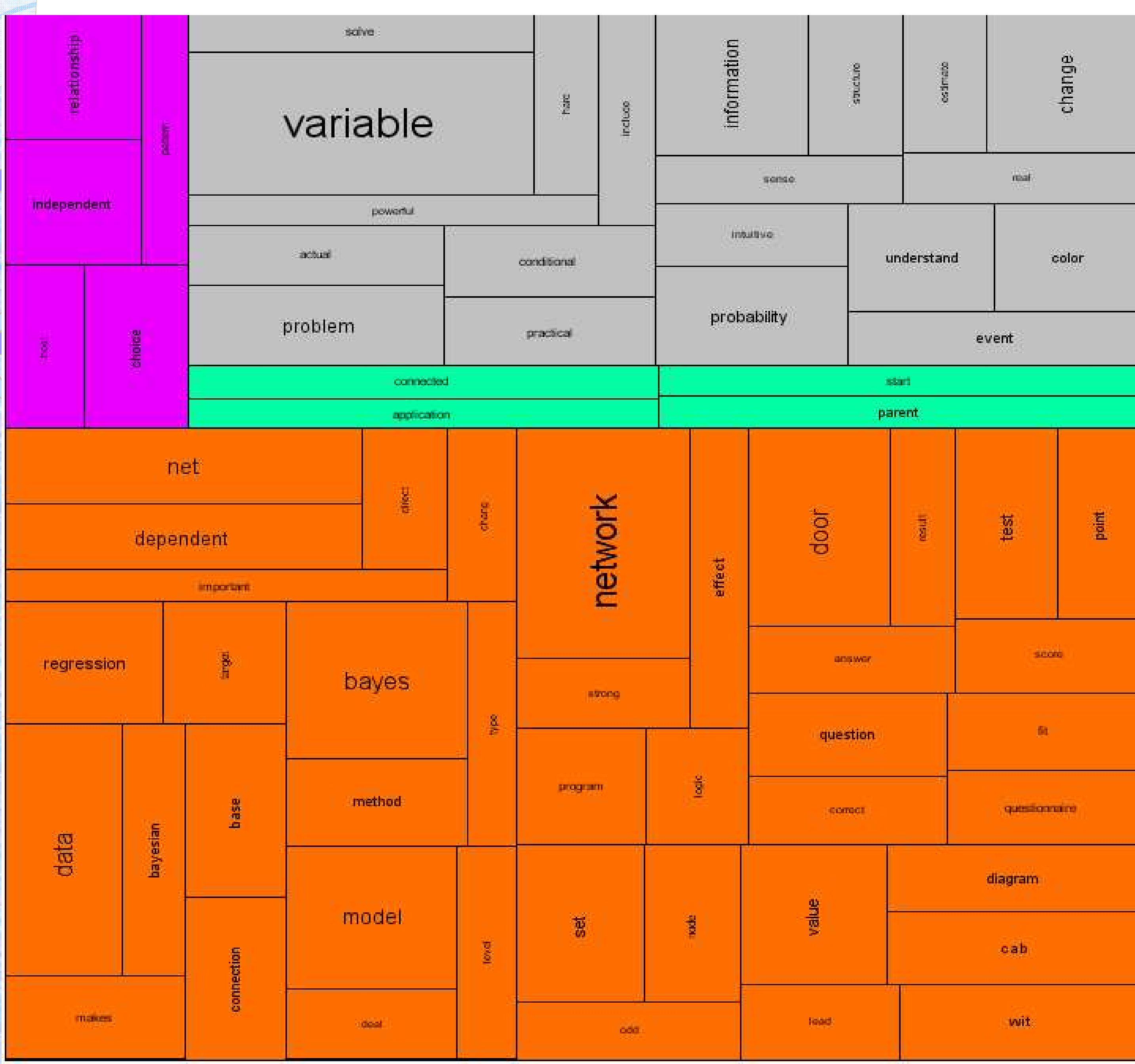


Based on an insurance study

Treemap of words

- Displays something like this have appeared in press stories, and for some may exemplify text analytics
- Words appear in a block, with words most associated with each other closer and words less associated farther apart
- Not all treemaps are alike
  - A conventional treemap does not order the rectangles statistically
  - A **Wordle** packs words as closely as possible and sizes them according to frequency
    - Adjacent words may not be related or located near each other
- In this map, the sizes of the words and their surrounding rectangles are proportional to the square root of word frequency in the document
  - This is the same square root transformation that used elsewhere
  - The rectangles are colored based on the hierarchical cluster analysis
    - If regions of rectangles are all one color, that increases our confidence that the cluster analysis was not due to chance
    - However, if colors are scattered throughout the rectangles, then the clusters likely were not coherent
    - We then would need to read them cautiously

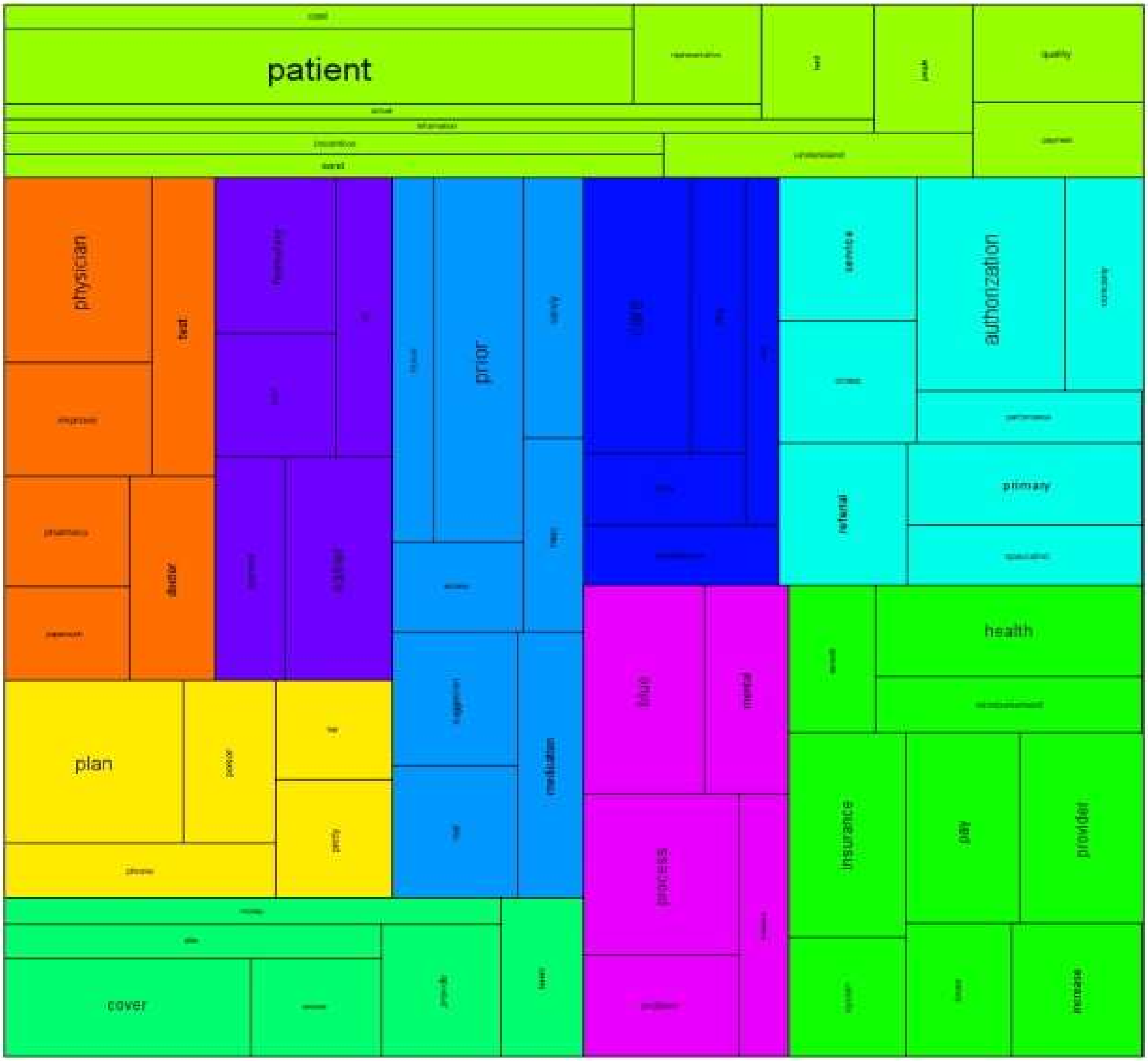
Not our tree map



Tree Map

Based on an article on Bayes Nets

Converge Analytic



Tree Map

Based on an insurance study

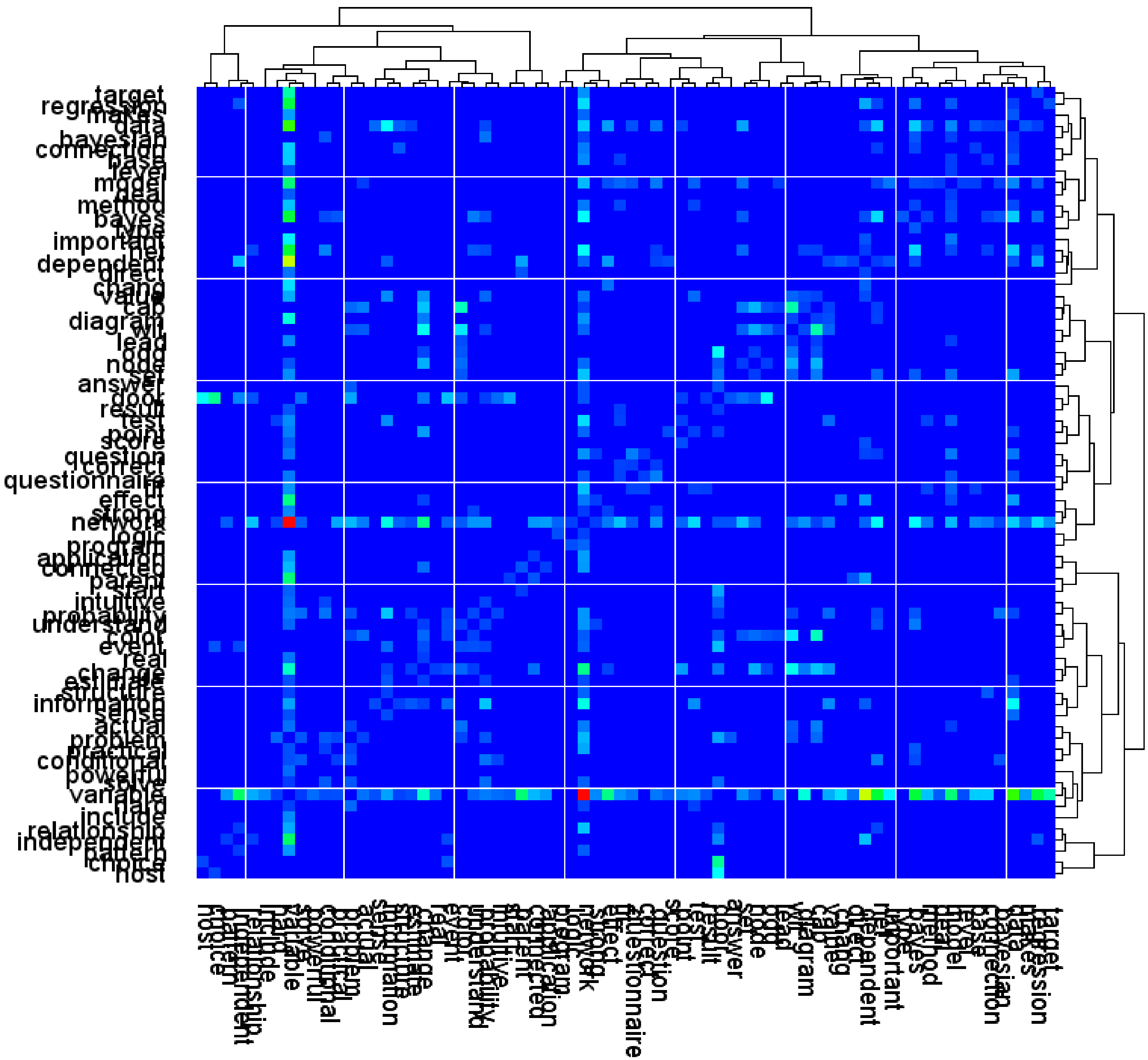
Converge Analytic

Heatmap of word associations

- Heatmaps have appeared as high science in the press
- This heatmap is based on the hierarchical clustering shown earlier
- Color represents the strength of the associations of pairs of words taken from the most frequent words in the document
  - These were computed using the sliding **n-gram window** run through the document
  - Black pixels show little or no association between words
- The clustering scheme appears along the edges of the map
- The relative popularity of heatmaps is somewhat puzzling
  - Research by Cleveland (1984) shows that people have the most difficulty using color hue, saturation and density as comparative measures

Not our Cleveland—
but 1984

Heat map



Based on an article on Bayes Nets

Word counts and betweenness

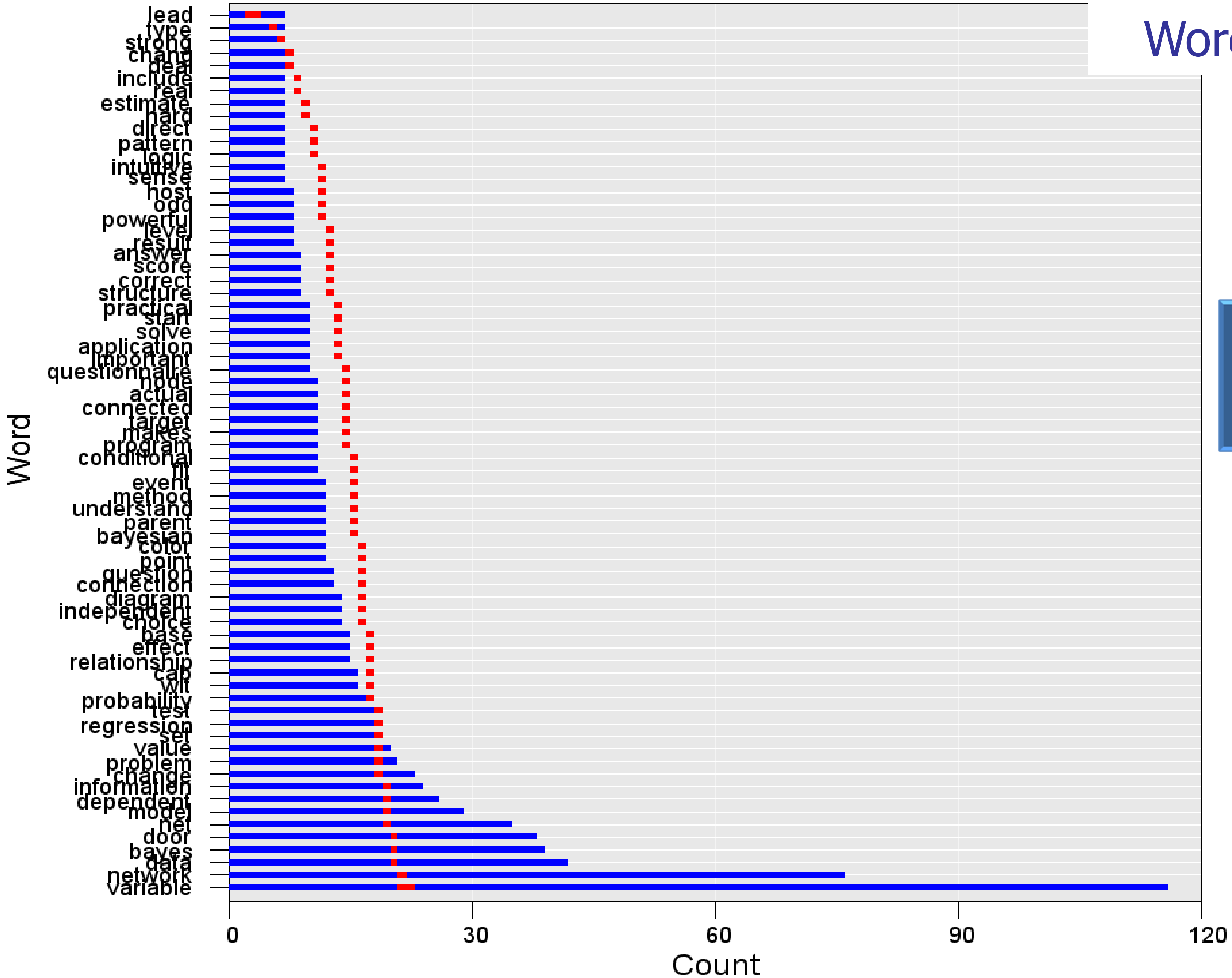
- This is very basic information about the document yet easily overlooked
- The frequencies of words are compared with what we would expect, assuming that all words are equally probable
  - The red dots show the 95% acceptance intervals for a completely random set of words
  - With frequencies of words random and equally probable, all the bars would fall inside the red dots
- The second bar graph shows the words' **betweenness centrality**
  - **Betweenness centrality** usually identifies people in a social network who are connected by many relationships
  - In text analysis, words with high **betweenness** appear near many other words that represent different concepts
    - These words can be viewed as having multiple meanings or nuances, or as key connectors of themes

We need to be alert for multiple meanings

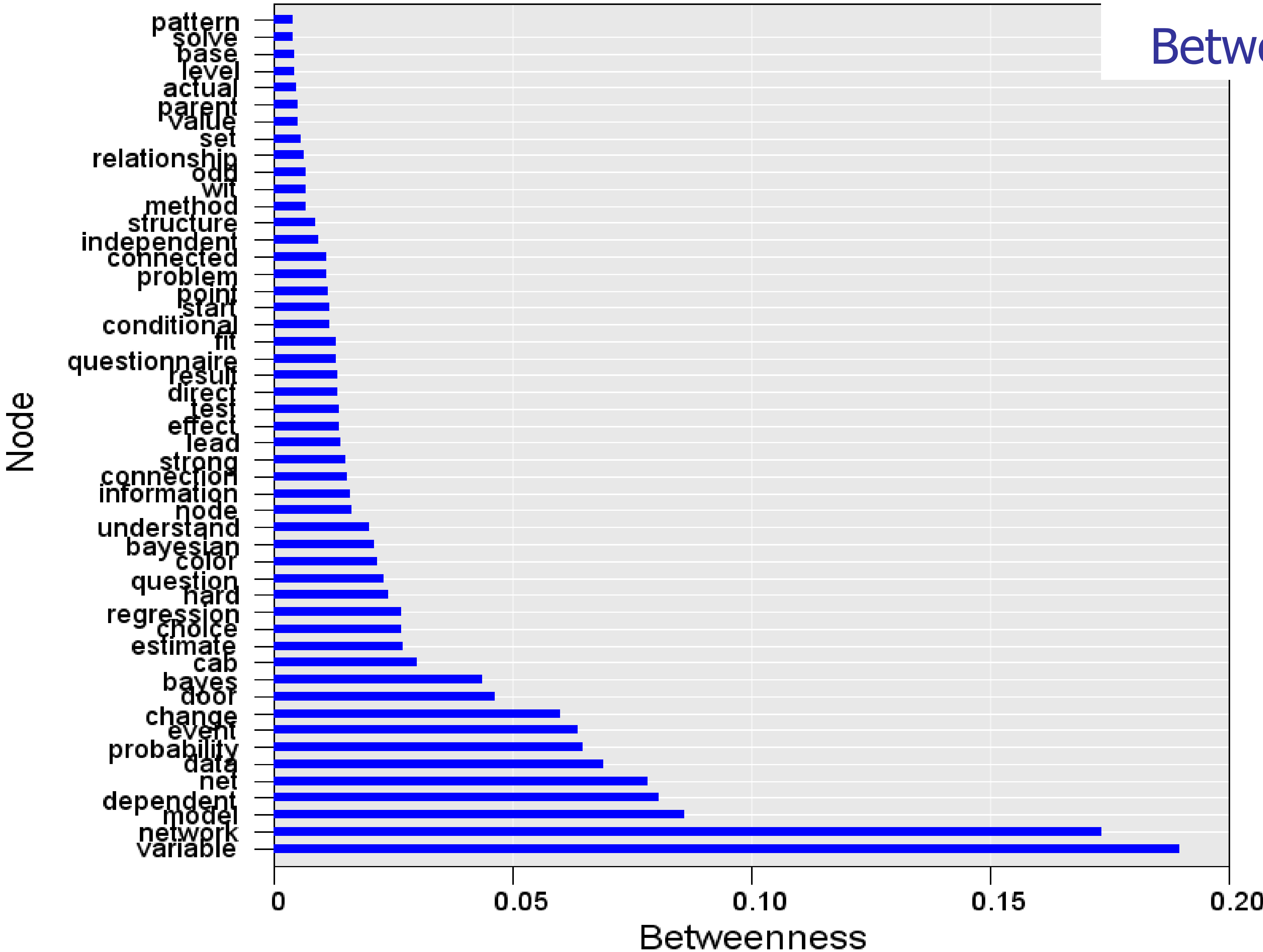
Converge Analytic

Word counts

Based on an article on Bayes Nets



Betweenness



Based on an article on Bayes Nets

Associations with flagged words

- Another kind of counting exercise, this shows which words occur along with each of four selected words
- Like betweenness, this gives a view of how many ideas link to each word—but with detail on the specific linkages
- This example comes from an insurance study

Word		Associated Words
1	authorization	answer authorization calling card deal decrease diagnose direct easier education efficient eligibility extended guidance hardest hour implement information insurance looking medication medicine MRI nice patient people period prior problem read referral request requirement resource revisit simplify sooner suggestion test trained urgent wed
2	educate	able allow backdate company frequent online panel patient process real refer referral satisfy system
3	education	care delineation difference frequent improve information member patient process question refer referral responsibility specialist
4	educational	network requirement service

Converge Analytic

Prediction and description in text analytics

- These two approaches give different views of what happens in a block of text, whether it is, e.g., a set of verbatim responses or a complete document
- Predictive approaches seek to find the words or combinations of words that predict patterns in a dependent variable, like share or overall rating
- Descriptive methods give more of an overall feeling or a "lay of the land"
  - While qualitative in nature, this can enhance understanding of themes and ideas in the text
  - Most text analysis appears to fall under this heading
  - Is this supplemental or sufficient? We need to decide

There is much to learn from the broad outlines— but is it what we need?

Converge Analytic

Questions? Comments? Need more information?

Dr. Steven Struhl

smstruhl@convergeanalytic.com

smstruhl@gmail.com

☎ (847) 624-2268