Cultural Domain Analysis (CDA)

Steve Borgatti

Boston College

ABT Associates

8 January, 2002

Topics

- Overview of CDA
  - Theory
  - Data collection
  - Analysis
  - Applications
- Software Demonstration
  - Anthropac
  - UCINET/NetDraw

History

- Became popular in the 60s
  - In part because of availability of Bell Labs Fortran programs
- Linguistic anthropology → cognitive anthropology → marketing research
- Scientific, yet emic
  - From distinction between phonemic and phonetic
  - Describing & modeling the native's point of view
    - Models themselves remain in researcher's world
    - It is the objective that makes it emic, not the result
      - Informant ethnographies is yet another class of work

Underlying Notions

- Cognition organized around categories (domains)
  - Typically named, shared
  - Examples: illnesses, vegetables, countries
- Categories contain items
  - Some may be categories themselves
    - tree structure
- Items in semantic relations w/ each other
  - Part/whole, similar to, causes
- Items distinguished by attributes or features
  - What are the differences that make difference?

Componential analysis of horse terms

- Features
  - Stallion ← horse+male+adult
  - Mare ← horse+female+adult
  - Gelding ← horse+neuter+adult|adolescent
  - Filly ← horse+female+adolescent
  - Colt ← horse+male|female+child
  - Foal ← horse+male|female+baby
- Paradigm

Sex

Age

<table>
<tr><td>HORSE</td><td>male</td><td>female</td><td>neuter</td></tr>
<tr><td>adult</td><td>stallion</td><td>mare</td><td rowspan="2">gelding</td></tr>
<tr><td>adolescent</td><td></td><td>filly</td></tr>
<tr><td>child</td><td colspan="2">colt</td><td></td></tr>
<tr><td>baby</td><td colspan="2">foal</td><td></td></tr>
</table>

<table>
<tr><td>PIG</td><td>male</td><td>female</td><td>neuter</td></tr>
<tr><td>adult</td><td>boar</td><td>sow</td><td rowspan="2">barrow</td></tr>
<tr><td>adolescent</td><td></td><td>gilt</td></tr>
<tr><td>child</td><td colspan="2">shoat</td><td></td></tr>
<tr><td>baby</td><td colspan="2">piglet</td><td></td></tr>
</table>

Typical CDA Study

- Eliciting domain
- Eliciting items within a domain
- Analyzing structure of the domain
  - Semantic relations
  - Uncovering the meaningful attributes
- Analyzing structure of agreement among respondents
- Prediction
  - [People react similarly to similar things]

Elicitation & Measurement

- Domain membership
  - Free listing
- Measuring Similarities
  - Pile sorts, Triads, Direct rating, Map drawing
- Attributes
  - Eliciting:
    - Pile sort labeling
    - Interpreting MDS maps of similarities
  - Measurement:
    - Paired comparisons
    - Direct rating

Analysis Techniques

- Multidimensional scaling (MDS)
  - Of aggregate similarity data
- Cluster analysis
  - Of aggregate similarity data
- Property Fitting
  - Relating attributes to similarity data
- Consensus Analysis
  - Understanding variations in beliefs

Free Listing

- Basic idea:
  - Tell me all the <category name> you can think of
  - Typically loosely timed, no questions allowed
  - An example of Spradley's "grand tour" question
- Contrasts with survey open-ended question
  - Open-end is typically about the respondent:
    - what do you like about this product? what ice-cream flavors do you like? what illnesses have you had?
  - Free list is about the domain:
    - what ice-cream flavors are there? what illnesses exist?

Domain of Fruits

TABLE 2.1
Frequency of Mention of "Fruits" in Free List Task

Apple	37	Honeydew	9
Orange	35	*Avocado	8
Pear	34	Mango	8
Banana	33	Date	7
Grape	32	Fig	7
Peach	30	Prune	7
Tangerine	27	Gooseberry	6
Cherry	26	Raisin	5
Grapefruit	26	*Pumpkin	4
Pineapple	26	Casaba melon	3
Strawberry	22	Kumquat	3
Watermelon	21	Melon	3
Lemon	20	Breadfruit	2
*Tomato	19	Kiwi	2
Apricot	18	Passionfruit	2
Blueberry	18	Persimmon	2
Plum	18	Cranberry	1
Cantaloupe	17	Crenshaw melon	1
Lime	16	Currant	1
Nectarine	14	Elderberry	1
Papaya	14	Huckleberry	1
Raspberry	14	Loganberry	1
Blackberry	13	Manderine	1
Boisenberry	12	*Rhubarb	1
Tangello	11	Salmonberry	1
Guava	10	*Squash	1
Pomegranate	10	Taro	1
Coconut	9	Turnip	1

Weller & Romney. 1988.
Systematic Data Collection. Sage.

Domain of Vegetables

TABLE 2.2
Frequency Distribution of "Vegetables" Free Listing Task

Green beans	55	Chinese peas	6
Corn	50	Greens	6
Carrots	49	Okra	6
Peas	41	Summer squash	6
Lima beans	40	Blackeyed peas	5
Lettuce	38	Swiss chard	5
Broccoli	37	Wax beans	5
Califlower	36	Bamboo shoots	4
Brussels sprouts	35	Navy beans	4
*Tomatoes	32	Alfalfa sprouts	3
Onions	30	Chile peppers	3
Spinach	30	Endive	3
Asparagus	29	Kidney beans	3
*Squash	28	Leek	3
Cucumbers	26	Parsnips	3
Celery	25	*Pumpkin	3
Cabbage	24	Redleaf lettuce	3
Zucchini	24	*Rhubarb	3
*Turnips	23	Water chestnuts	3
Potatoes	20	Butterleaf lettuce	2
Artichokes	18	Green onions	2
Bell peppers	18	Kale	2
Radishes	18	Kolari	2
*Avocado	18	Red onions	2
Beets	13	Sauerkraut	2
Rutabaga	11	Butternut squash	1
Bean sprouts	10	Garlic	1
Eggplant	9	Hubbard squash	1
Mushrooms	8	Jicama	1
Parsley	8	Peapods	1
Pinto beans	8	Pickles	1
Yams	7	Soybeans	1

*Indicates items that appear on both "fruit" and "vegetable" lists.

Weller & Romney. 1988. *Systematic Data Collection.* Sage.

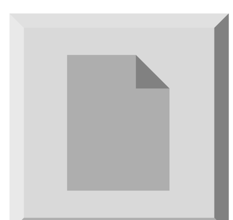
The “Bad Words” Domain

WARNING:
4-Letter words follow!

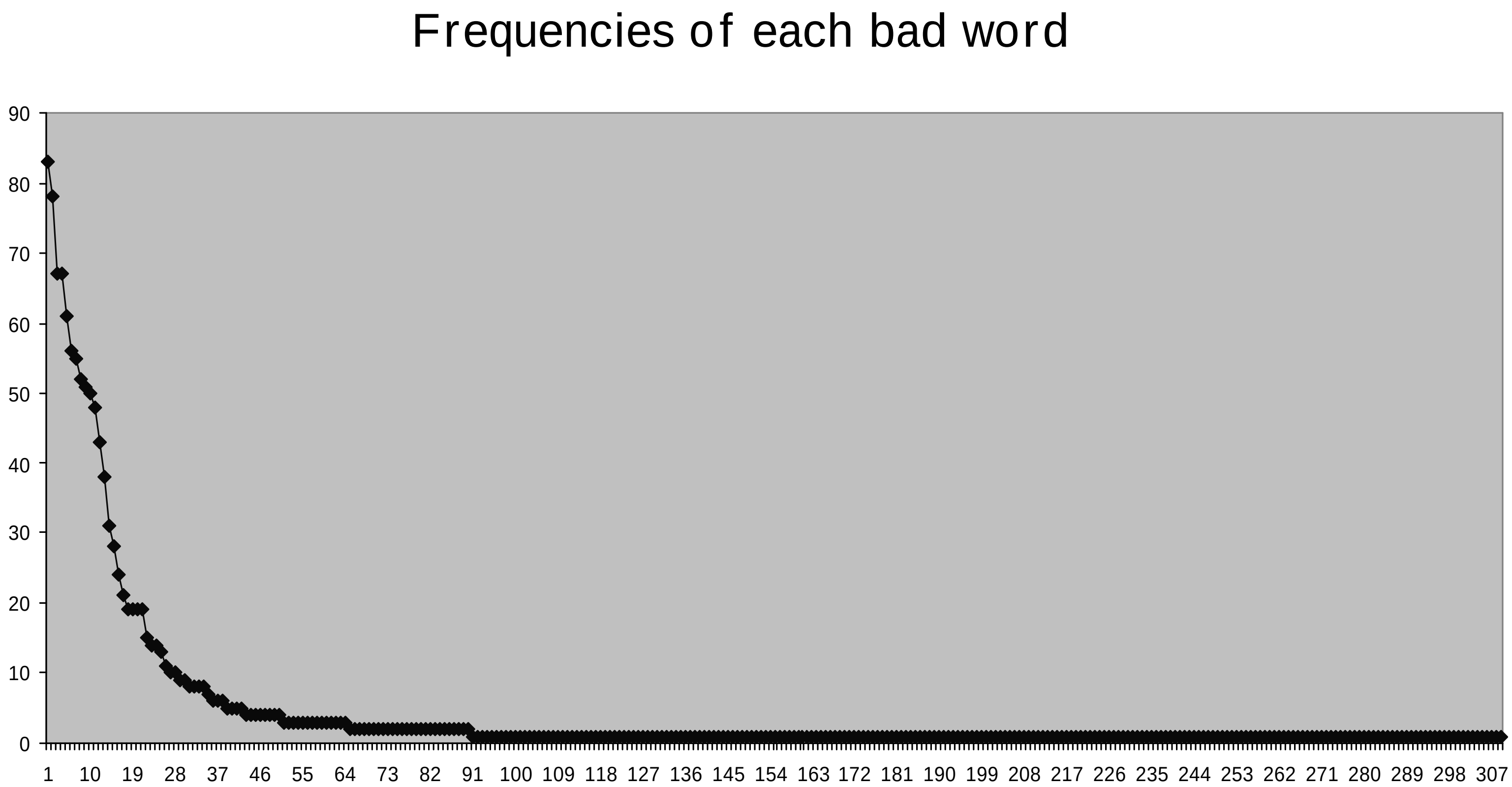
The squeamish and the moral should go back to work now!

Frequencies

- Sort in descending order
- Tally average position in lists
- Combine frequency and position to create salience measure
- May need editing to standardize spelling
- In some cases, want to collapse synonyms
  - Not in linguistics projects, though

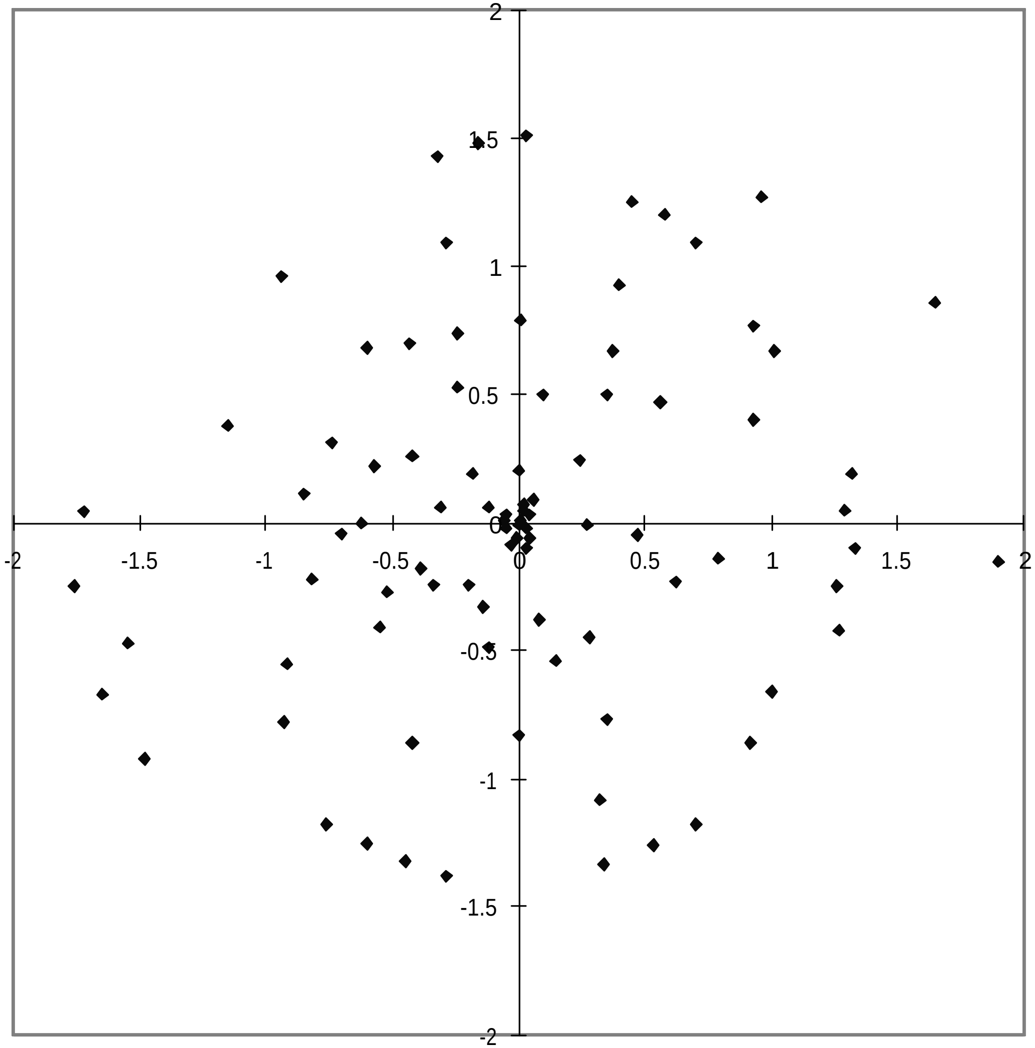


Domain borders are fuzzy



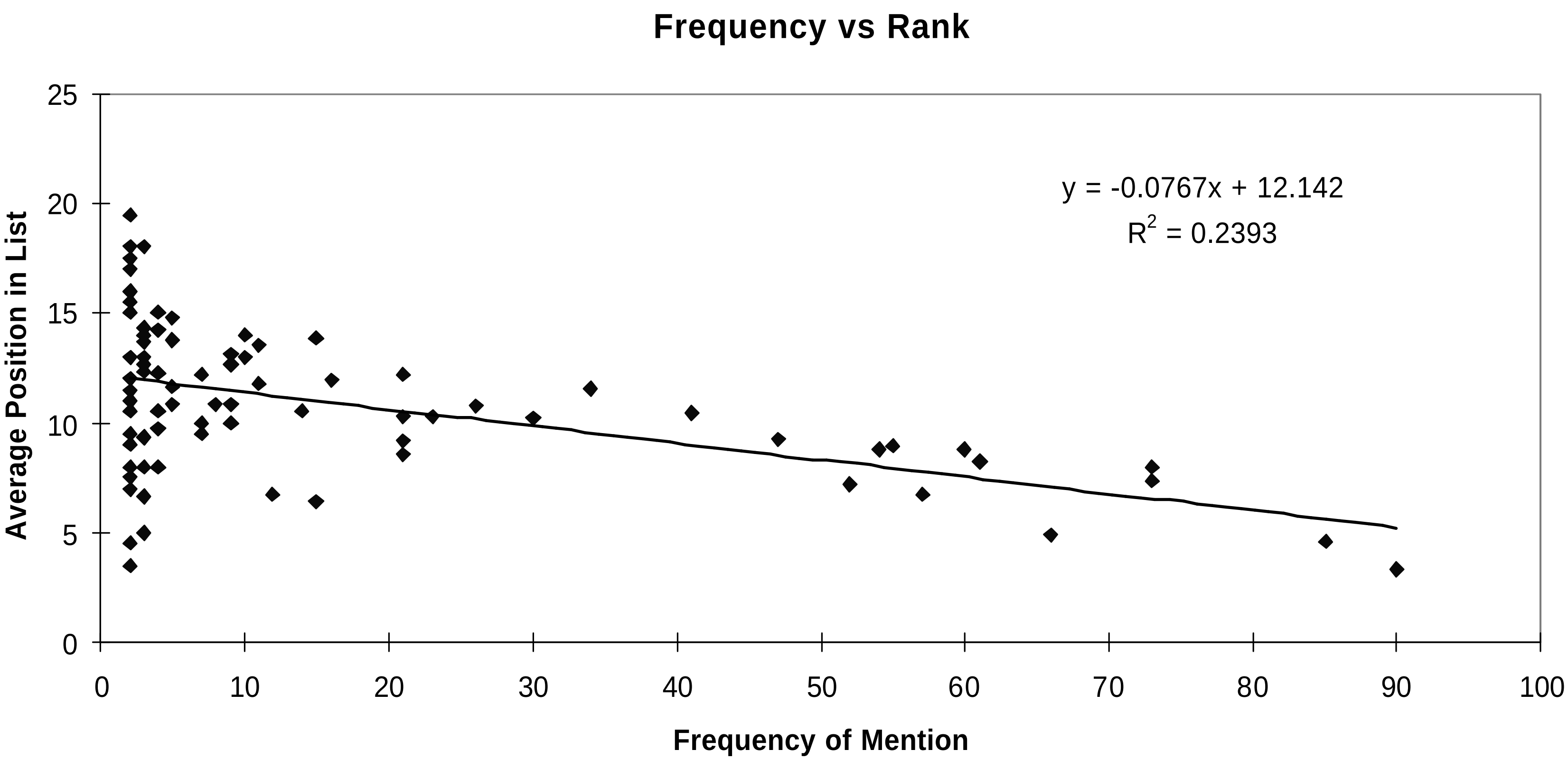
Domains have core/periphery structure

- MDS of item-item co-occurrences
- Each dot is a bad word
- Core items are in the center – in everybody's list – and co-occur with each other

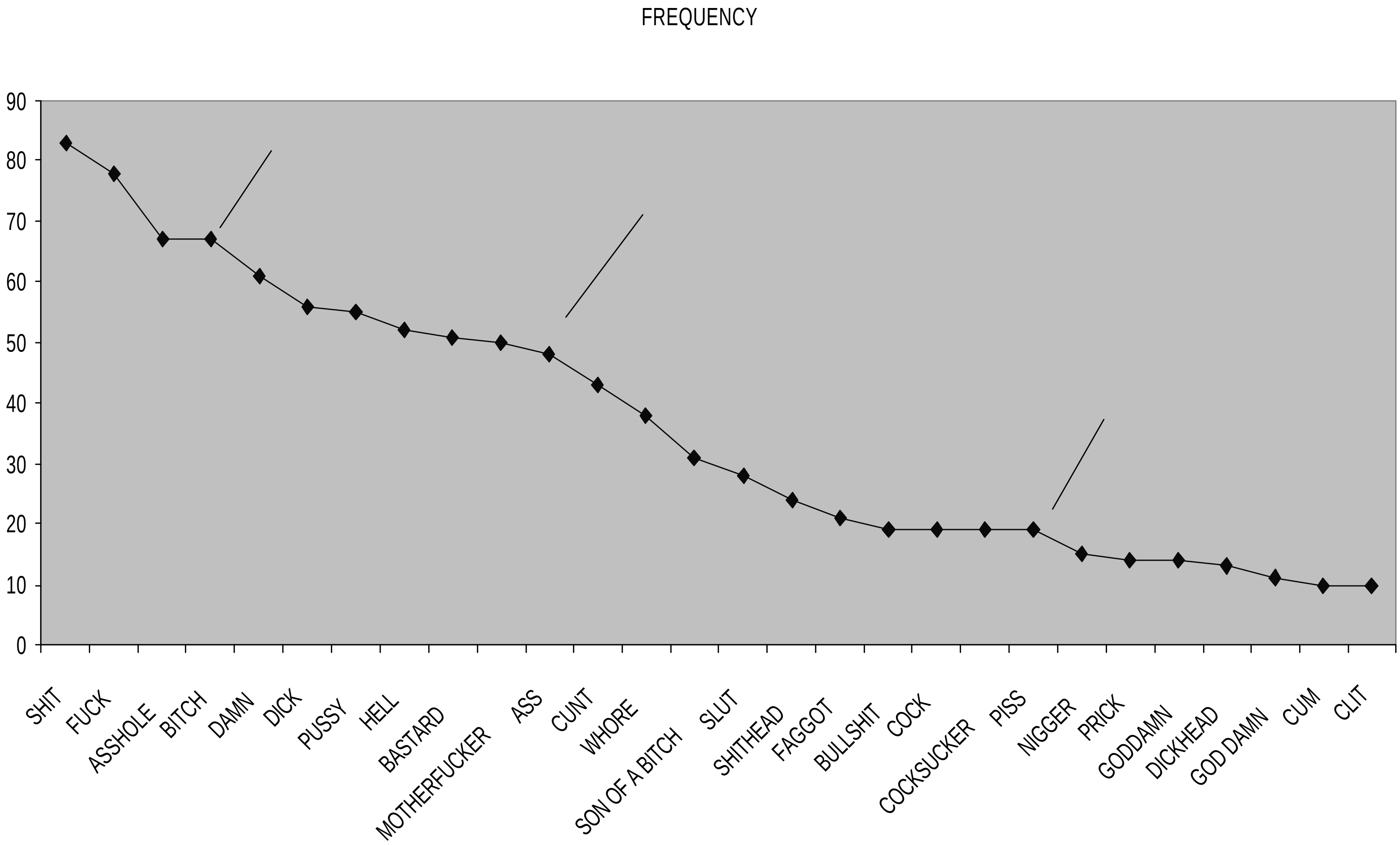


Core items typically mentioned first

Characteristic negative correlation between avg rank and frequency



Use scree plot to select core



Can analyze respondents as well

- Length of lists
- Conventionality of their lists (do they tend to list more popular items)
- Correlation between rank (position on list) and sample frequency
- Similarities (overlaps) in people's lists

Things to notice ...

- Boundaries of a domain are fuzzy
  - Not just artifact of aggregation
  - For additional data collection, need inclusion rules
- Simple, established cultural domains have
  - Core/periphery structure
  - Core items recalled first
  - Consensus among respondents:
    - Each list has core items + idiosyncratic
    - We don't see clusters
- Quantitative analysis of qualitative data

Animals Domain

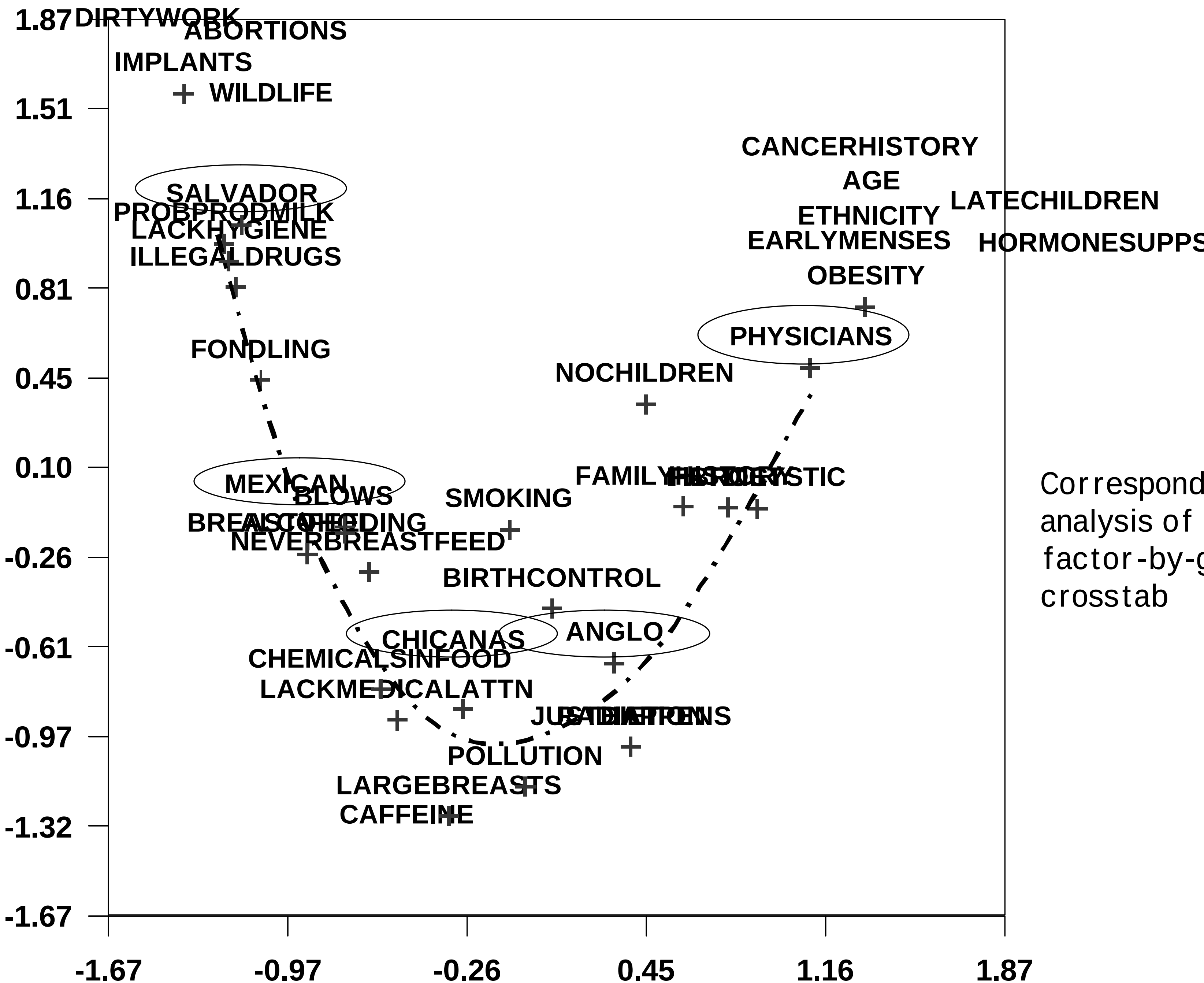
- Please grab a piece of paper and something to write with
- When I say ‘go’, please write down all the animals you can think of. You will have two minutes

Things to notice ...

- Ordering of items encodes ...
  - sub-category membership
  - Semantic relations such as similarity (lions & tigers) complementarity (forks & knives)
- Can reproduce map of domain from free lists

Causes of Breast Cancer

Salvadoran women ($N = 28$)	%[a]	Mexican women ($N = 39$)	%	Chicanas ($N = 27$)	%	Anglo women ($N = 27$)	%	Physicians ($N = 30$)	%
Blows, bruises	29	Blows, bruises	64	Chemicals in food	30	Family history	67	Family history	100
Problems producing milk	29	Never breast-feeding	33	Environmental pollution	26	Radiation	26	Obesity	37
Breast implants	21	Chemicals in food	28	Blows, bruises	26	Unhealthy diet	19	Hormone supplements	33
Disorderly, wild life	16	Excessive fondling	23	Lack of medical atten.	26	Smoking	19	First child after 30	30
Excessive fondling	14	Problem producing milk	23	Family history	26	Birth control pills	19	High fat diet	30
Smoking	14	Birth control pills	18	Never breast-feeding	22	Environmental pollution	19	Prior history of cancer	30
Never breast-feeding	14	Breast-feeding	15	Smoking	19	It just happens	15	Age	27
Lack of hygiene	14	Lack of medical atten.	15	High fat diet	11	Blows, bruises	15	No children	20
Family history	11	Smoking	13	Large breasts	11	Never breast feeding	11	Smoking	17
Abortions	11	Too much alcohol	13	Too much caffeine	11	Fibrocystic breasts	11	Fibrocystic breasts	13
Illegal drugs	11	No children	13	Birth control pills	11	High fat diet	11	Ethnicity	13
Dirty work environment	11	Lack of hygiene	8					Early menses	13
		Illegal drugs	8					Birth control pills	13
		Family history	8						



Correspondence analysis of factor-by-group crosstab

Things to notice ...

- <u>Comparative</u> analysis is particularly powerful
- Correspondence analysis
  - is clearly quantitative
    - Singular value decomposition of frequency matrix adjusted for row and column marginals
  - So we have quantitative analysis of qualitative data
  - On the other hand, the result is a picture – what can be more qualitative than that?

Uses of Free List

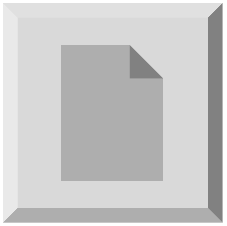
- First step in mapping the domain
  - i.e., getting a list of items to work with
- Analysis of the list itself
  - What makes something a fruit? A bad word?
  - Comparing salience of items for different groups
  - Examining similarities among respondents
    - Who lists the same items
  - Examining similarities among items
    - Which items tend to mentioned by the same respondents?
- Obtaining native terminology

Pile Sort Technique

- Basic idea:
  - On each of these cards is written the name of a thing. Please sort the cards into piles according to how similar they are. You can use as many or as few piles as you like.
- Outcome is quantitative measure of similarity among all pairs of items
  - For each pair of items, count the proportion of respondents who put them in the same pile
- Respondents only asked for non-quantitative judgments

Aggregate Proximity Matrix

- Item by item matrix gives the percent of respondents placing the two items in the same pile



- Typically visualize with MDS and cluster analysis

Triads

- Basic idea:
  - Present items to respondent 3 at a time, and ask which is most different

shark	seal	dog

- To elicit attributes
  - ask why they chose as they did, then try other triples
- To measure similarity
  - Systematically present all possible triples*
  - Each time an item is chosen most different it is a vote for the similarity of the other two
  - Arrange as an aggregate similarity matrix

* Or use clever balanced incomplete block design

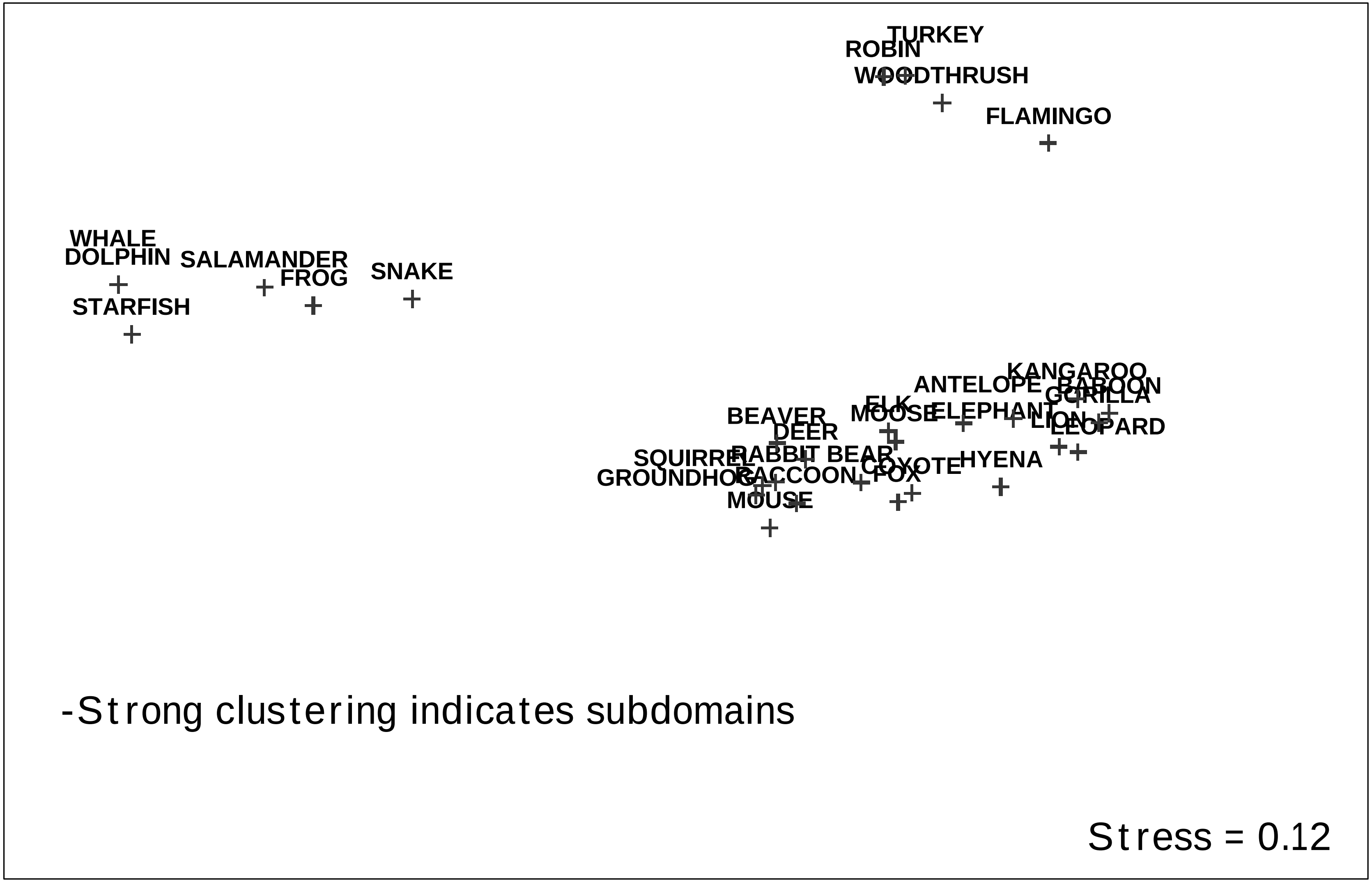
BIBDs

- Number of triples rises fast as items increase
  - n(n-1)(n-2)/6
  - For 30 items, have 4,060 triads to fill out …
- Each pair of items occurs n-2 times.
  - Let lambda stand for number of occurrences
- Balanced incomplete block design has each pair occurring same number of times, but lambda < n-2
  - Lambda-1 design: each pair occurs just once

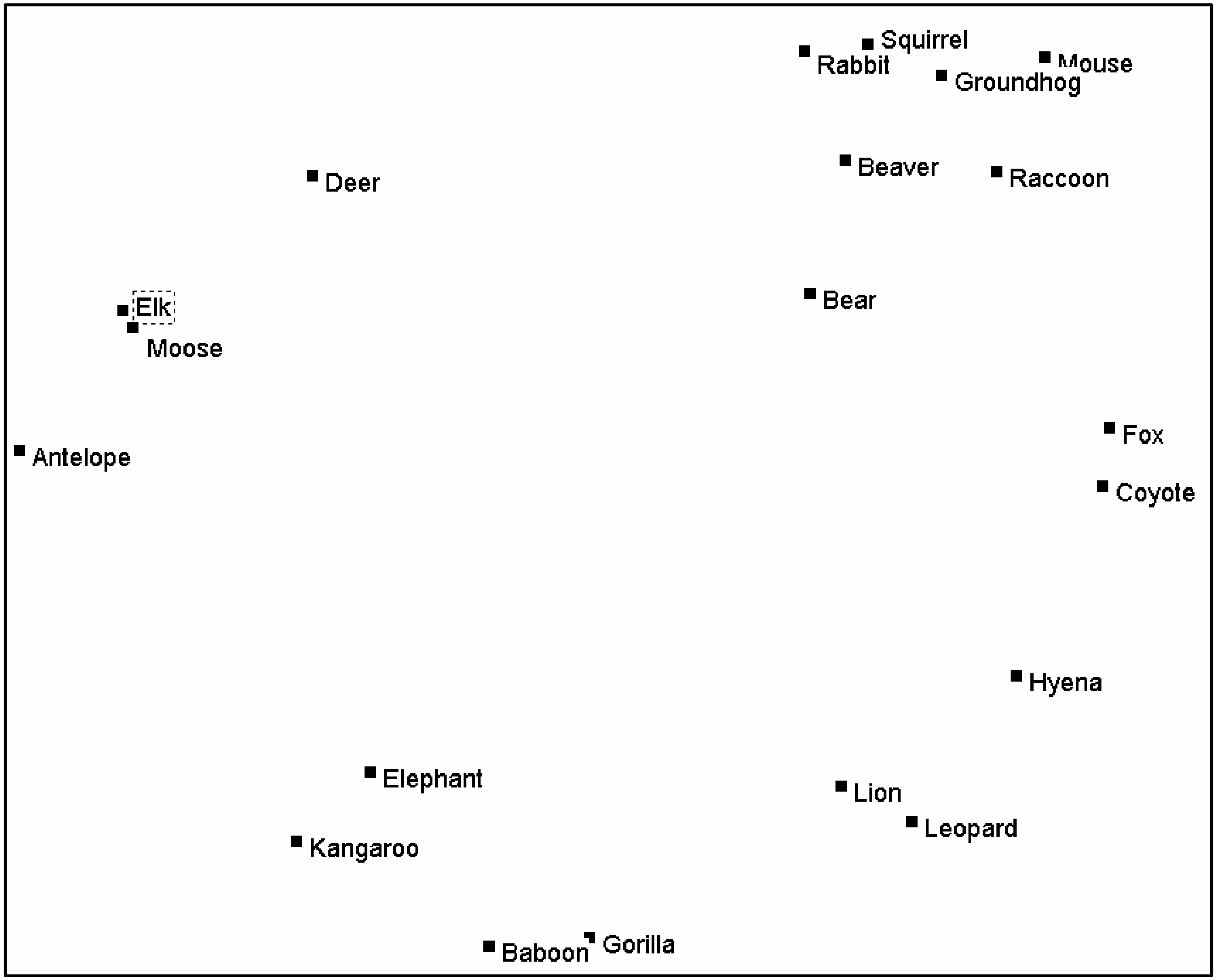
Representing Proximities

- Multidimensional scaling (MDS)
  - Maps items to points in Euclidean space such that points corresponding to more similar items are placed nearer to each other in the space
- Cluster analysis
- Network analysis techniques

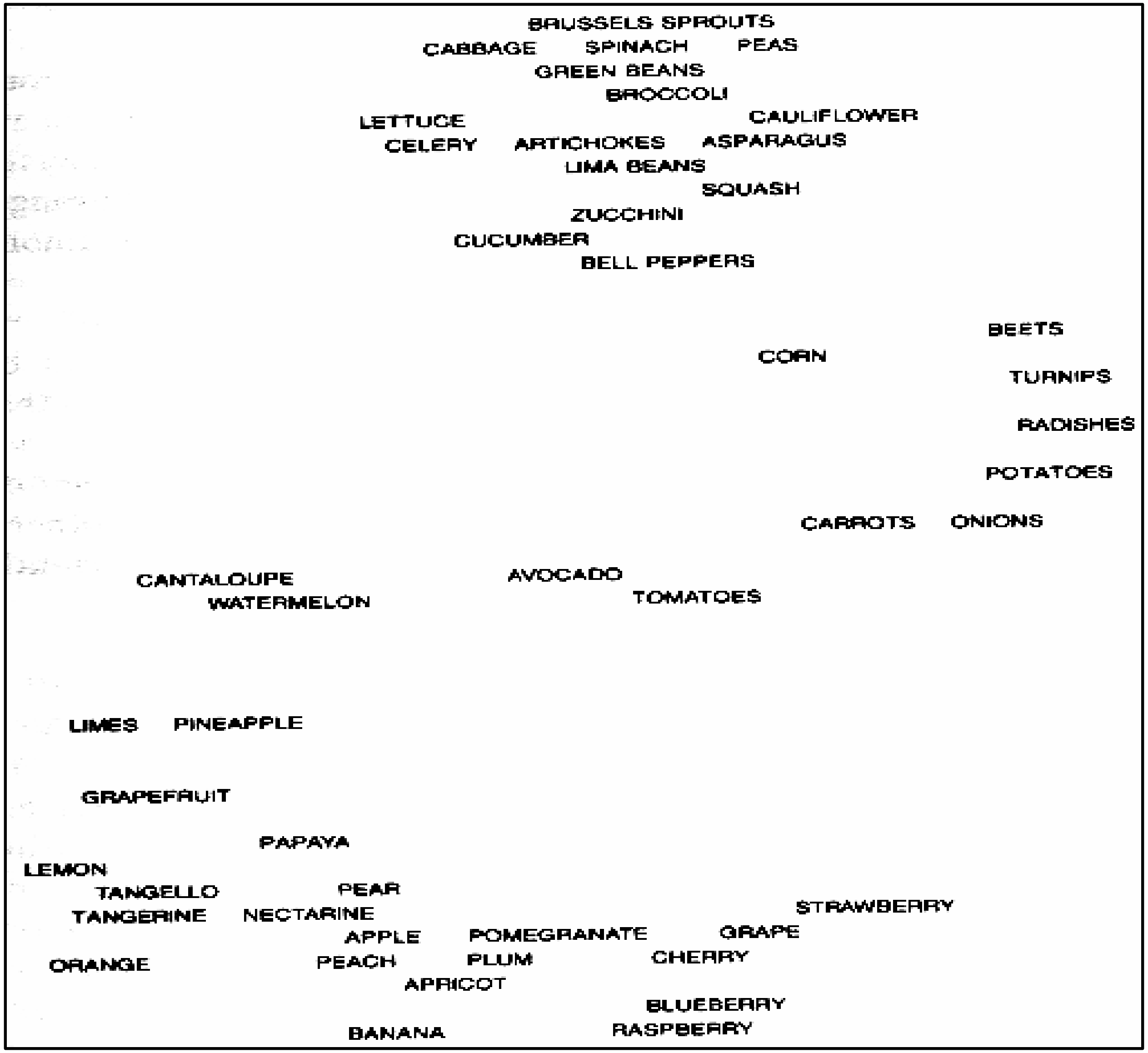
MDS of animals domain



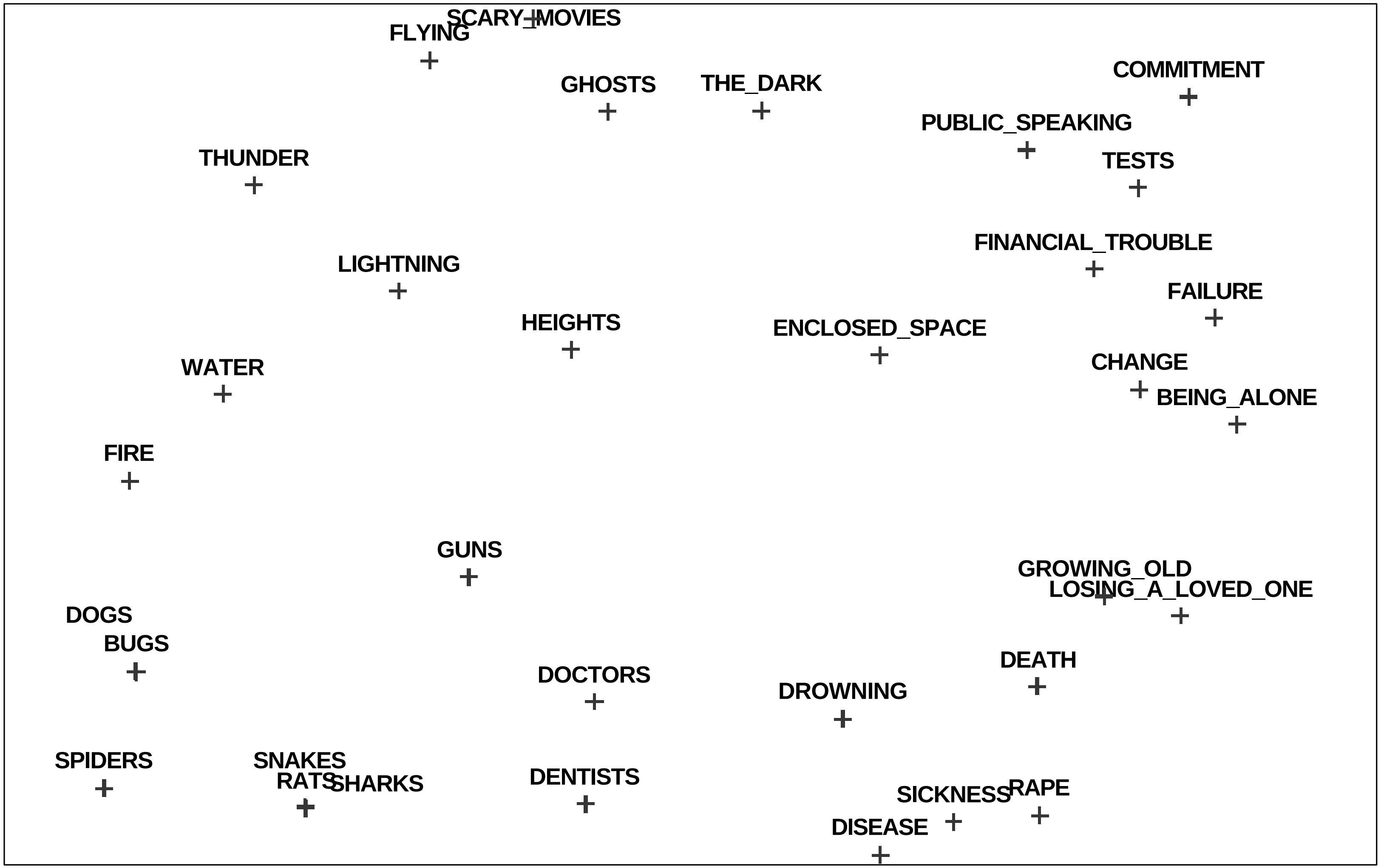
MDS of land animals only



Fruits & Vegetables



Things people are scared of



Things to notice ...

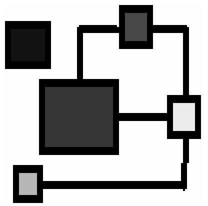
- Can use MDS with any proximity matrix
  - Aggregate similarities, Direct ratings, Confusion matrices, Correlation matrices, etc.
- Typically use 1-3 dimensions (mostly 2)
- Measure of fit (stress)
- Simplifies complex data
- Interpretation centers on
  - Looking for dimensions (quantitative item attributes)
  - Looking for clusters (qualitative item attributes)

Holidays

- Demo of Visual Anthropac pre-release version

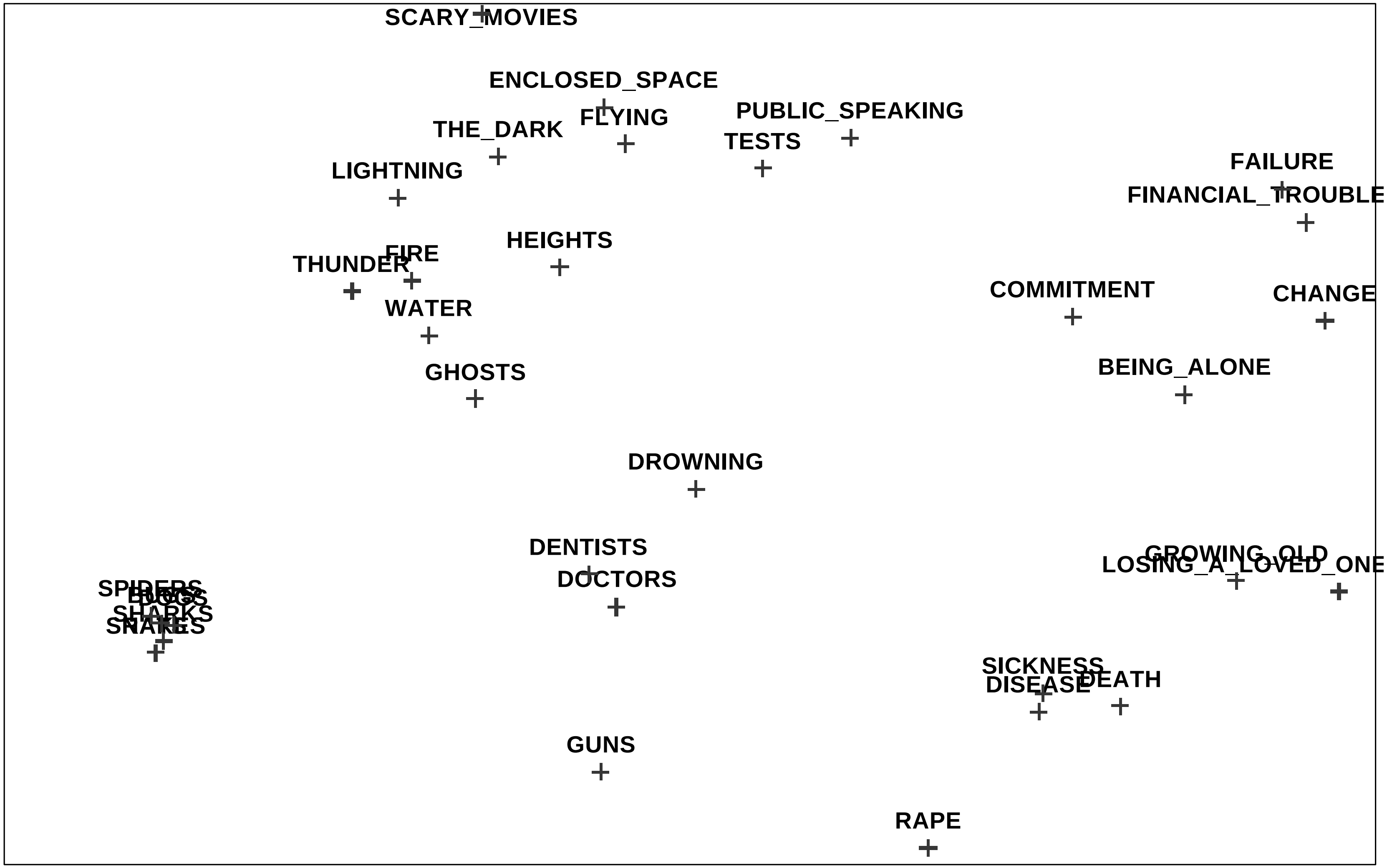
Network analysis

- Crimes dataset
- Animals
- Holidays



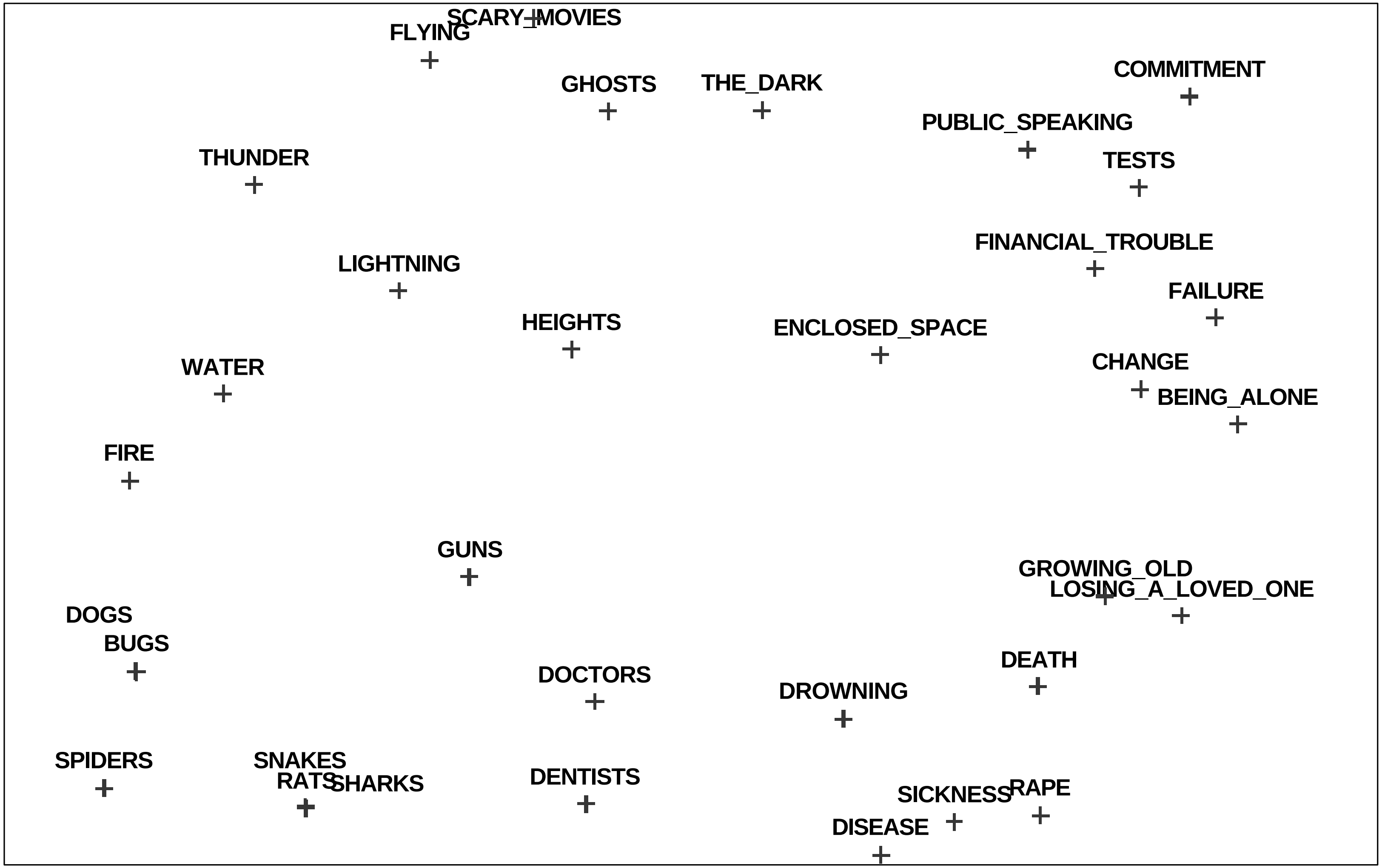
Things people are scared of

Female respondents

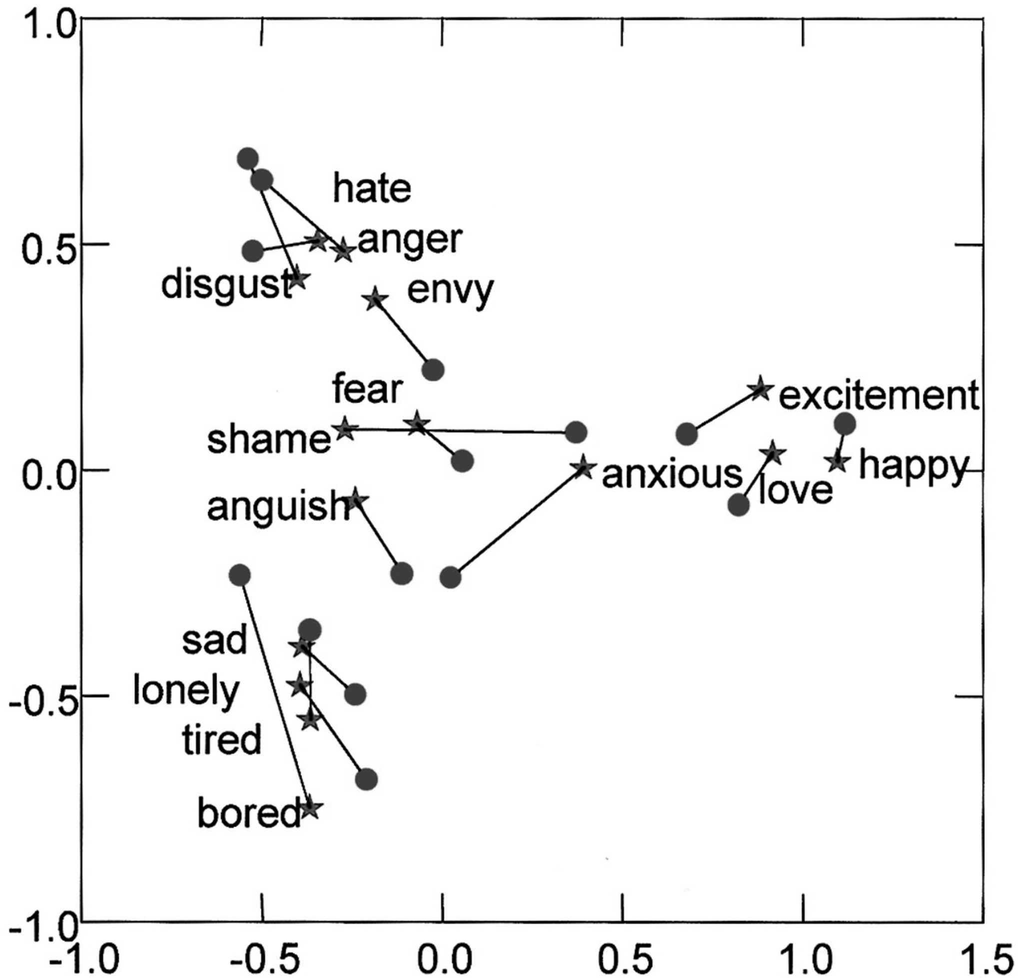


Things people are scared of

Male respondents



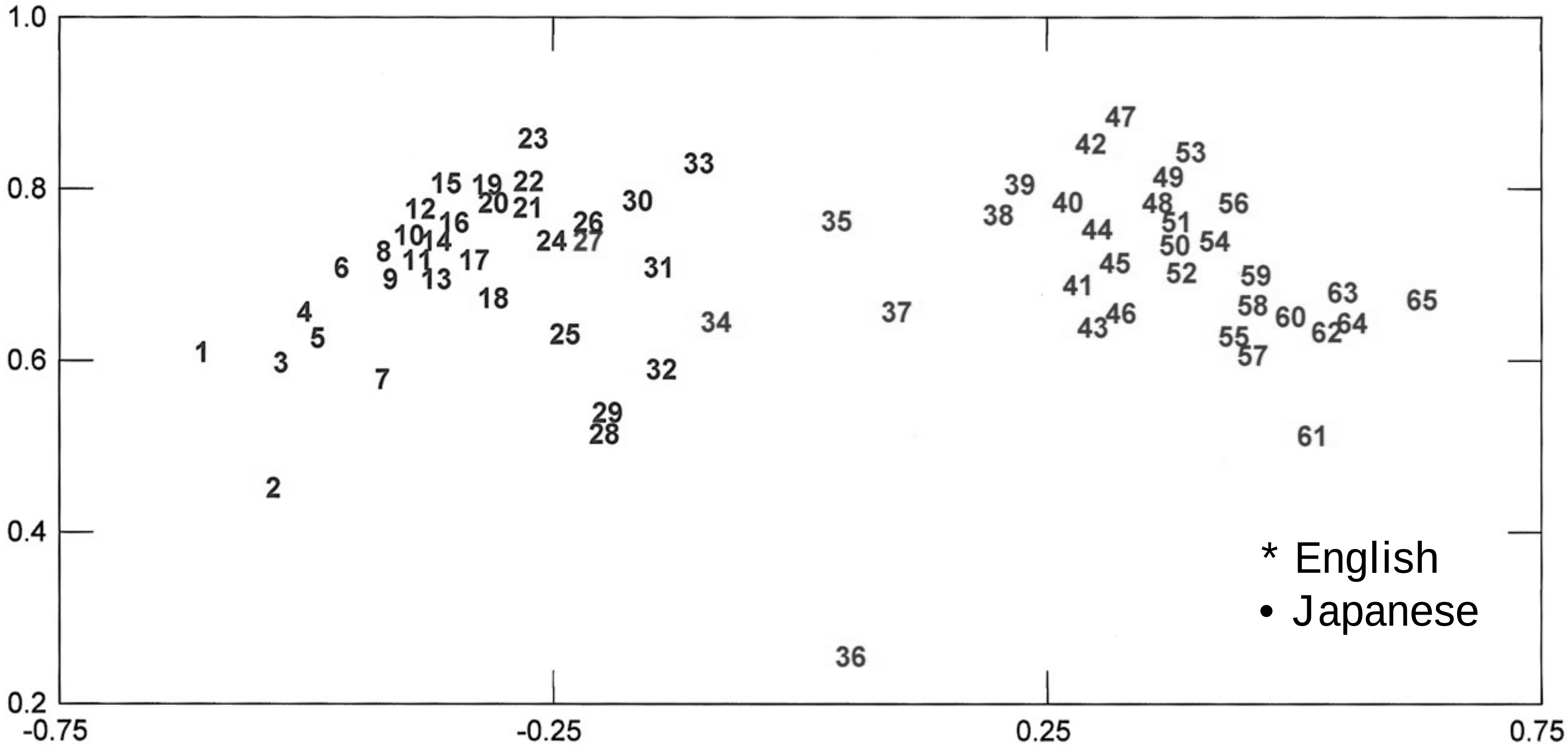
Discrepancy Analysis



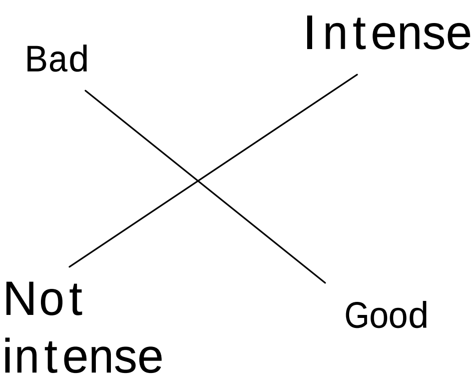
* English
• Japanese

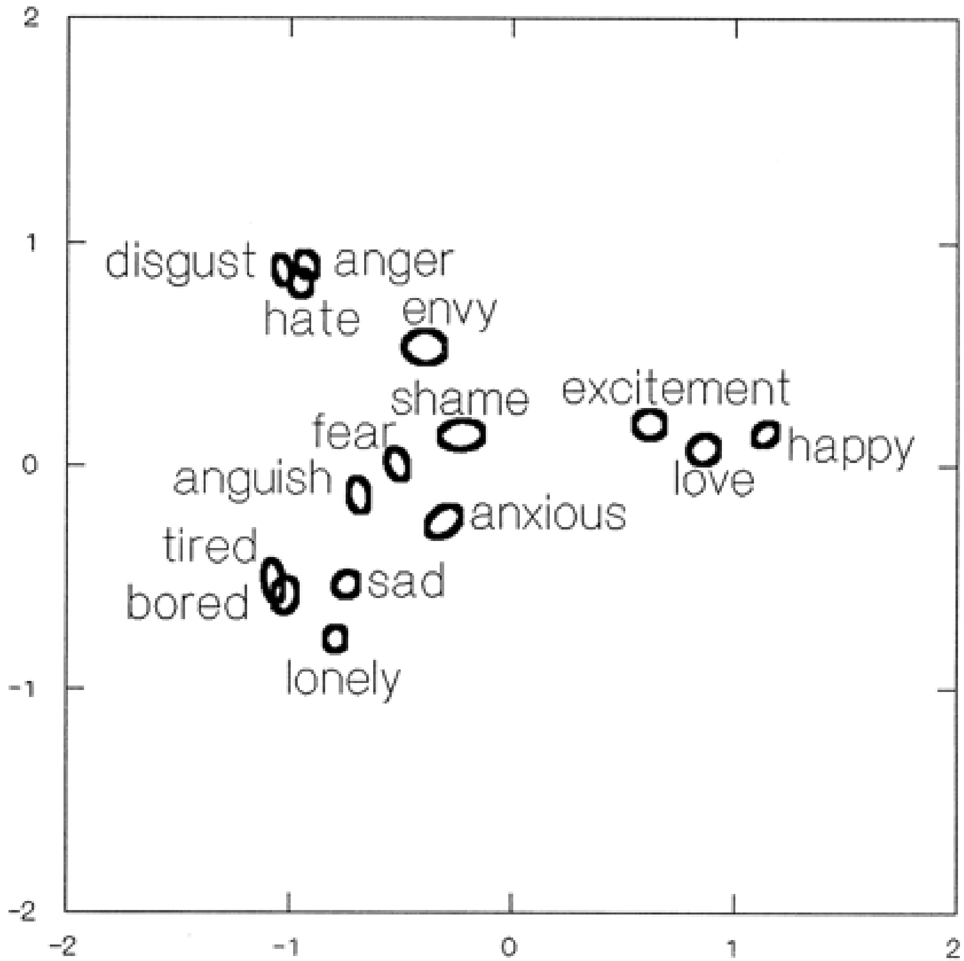
Romney, Moore, Batchelder and Hsia. 2002. Statistical methods ... PNAS 97(1): 518-523

MDS of similarities in respondents' sorts

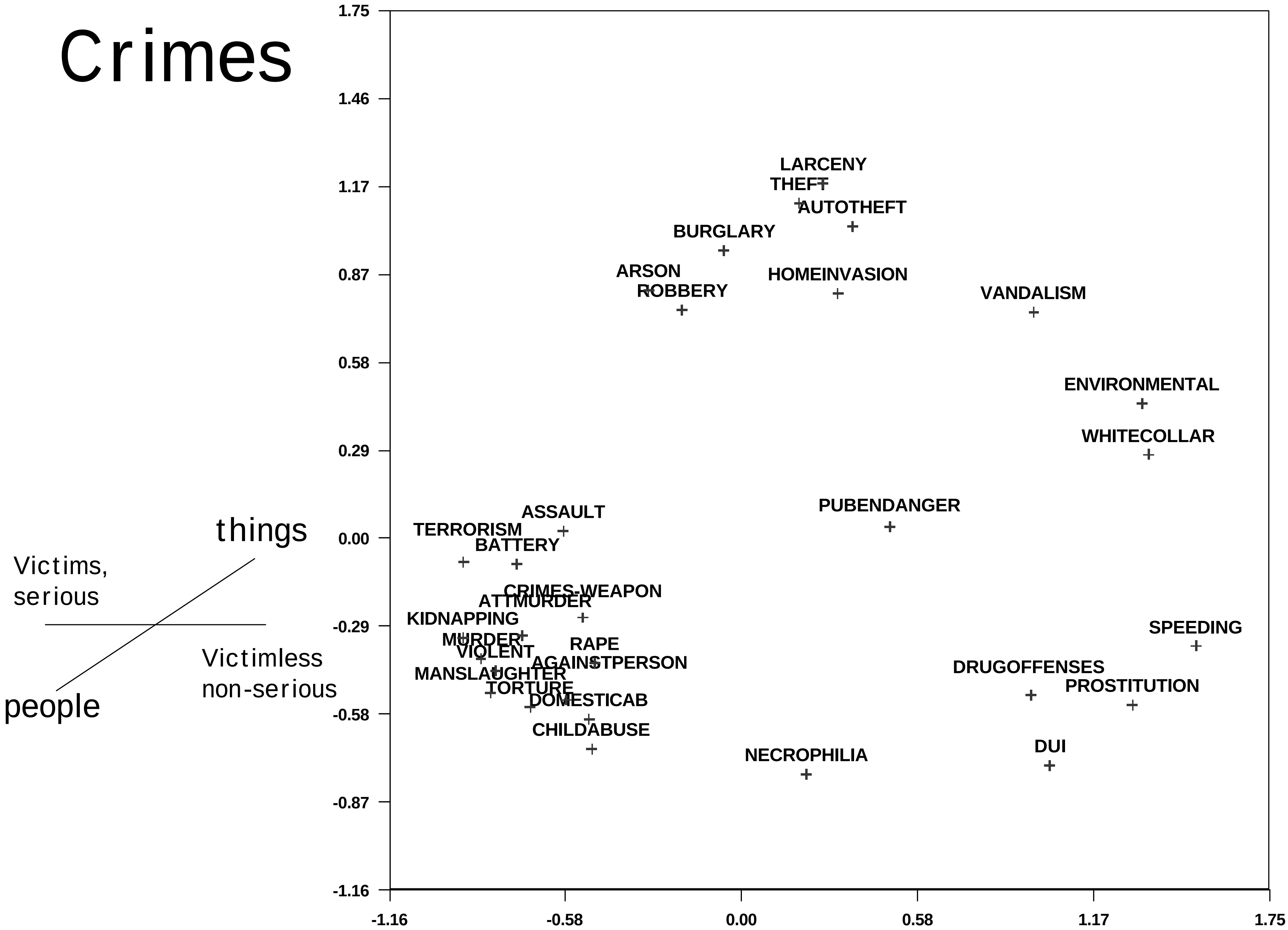


Emotion Terms



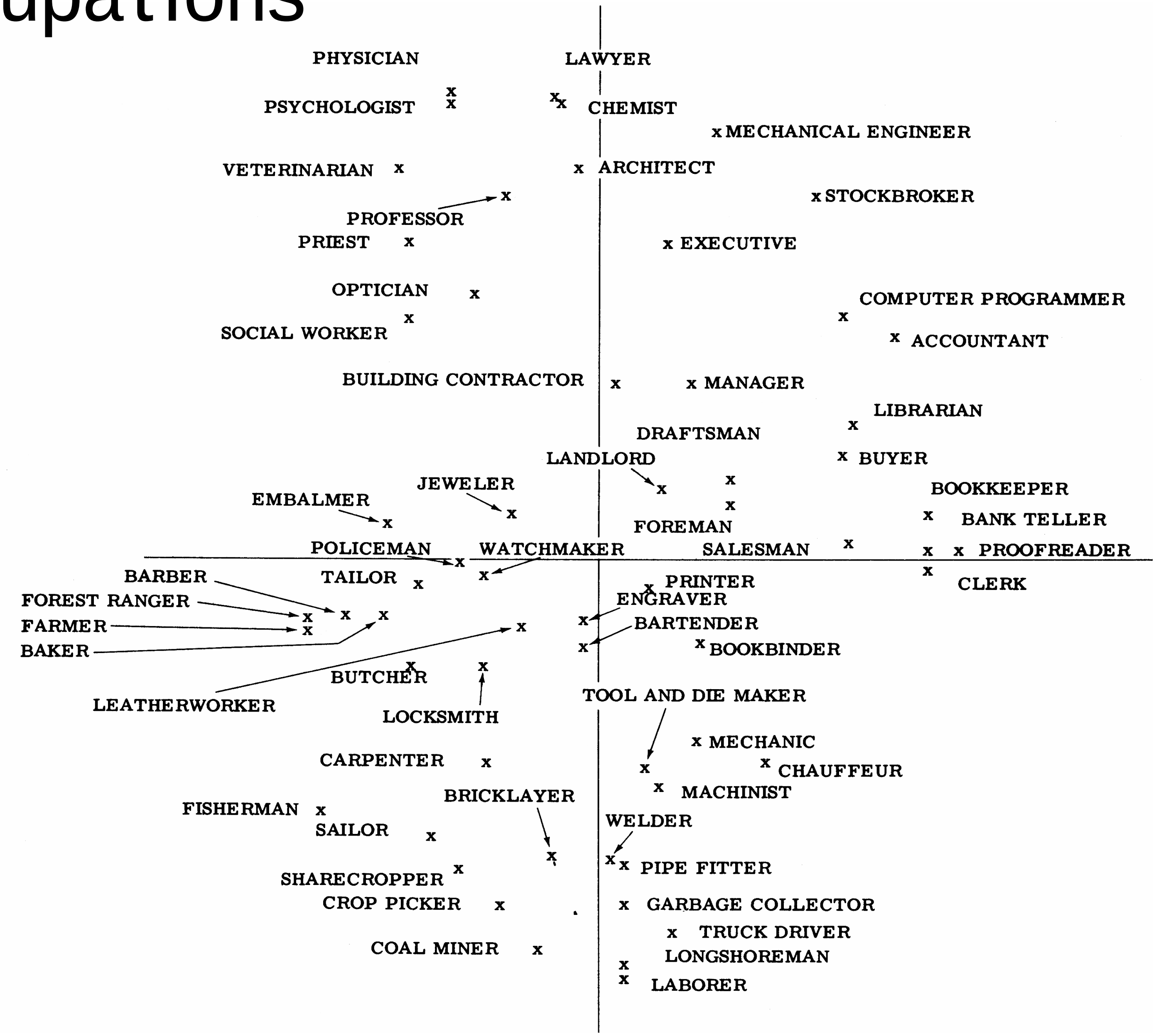


Crimes



LARCENY
THEFT
AUTOTHEFT
BURGLARY
ARSON
ROBBERY
HOMEINVASION
VANDALISM
ENVIRONMENTAL
WHITECOLLAR
PUBENDANGER
ASSAULT
TERRORISM
BATTERY
CRIMES-WEAPON
ATTMURDER
KIDNAPPING
MURDER
RAPE
VIOLENT
AGAINSTPERSON
MANSLAUGHTER
TORTURE
DOMESTICAB
CHILDABUSE
NECROPHILIA
SPEEDING
DRUGOFFENSES
PROSTITUTION
DUI
1.75
1.46
1.17
0.87
0.58
0.29
0.00
-0.29
-0.58
-0.87
-1.16
-1.16
-0.58
0.00
0.58
1.17
1.75
things
Victims,
serious
Victimless
non-serious
people

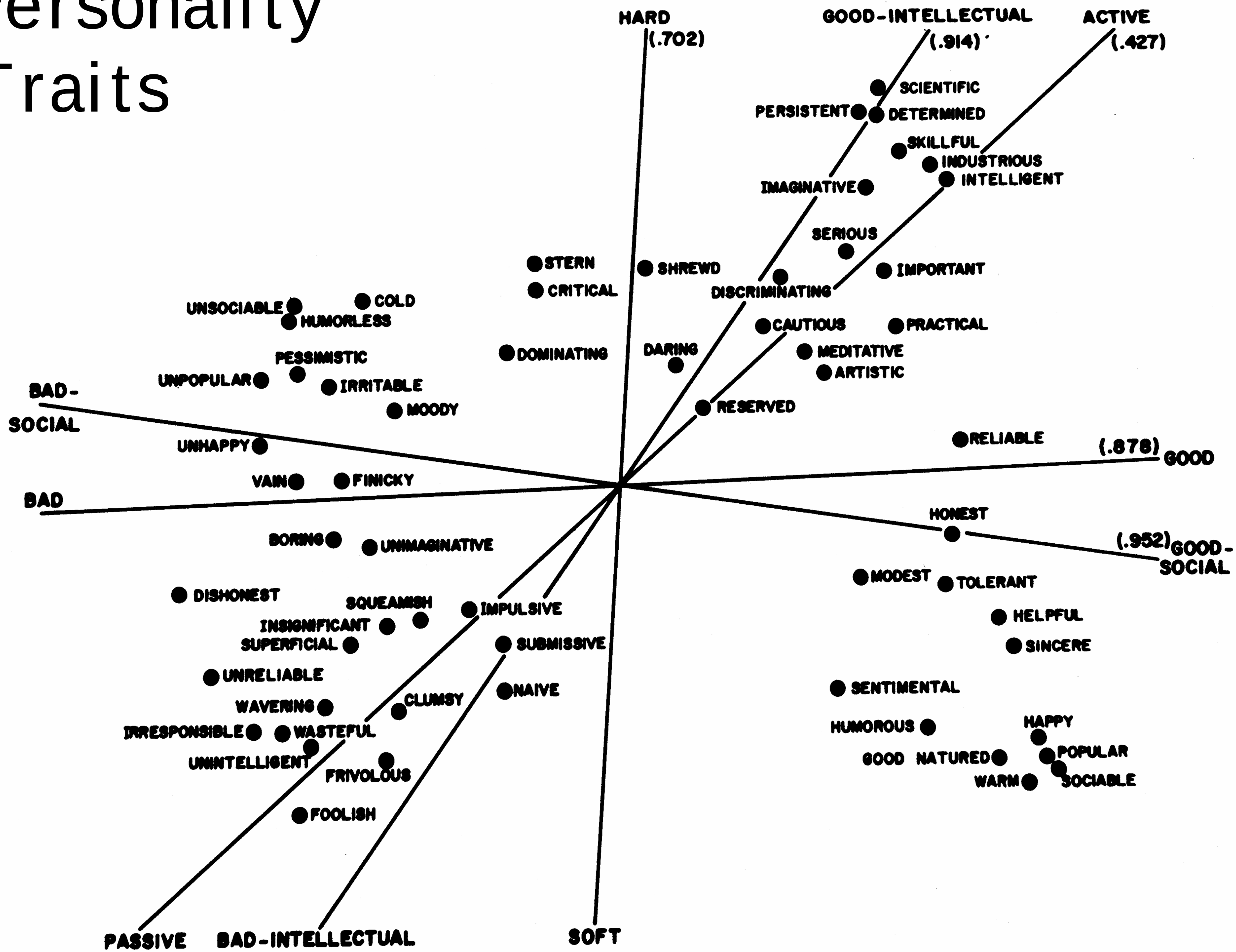
Occupations



Property Fitting (PROFIT)

- Testing hypotheses about dimensions in mds maps
  - Were respondents influenced by this dimension when they did the pile sorts or triads?
- Ask sample of respondents to rate each item on this dimension
- Aggregate across all respondents
- Regress average score on map coordinates
  - Prestige = b1*X_coordinate + b2*Y_coordinate
- Calculate vector angles from regression coefs

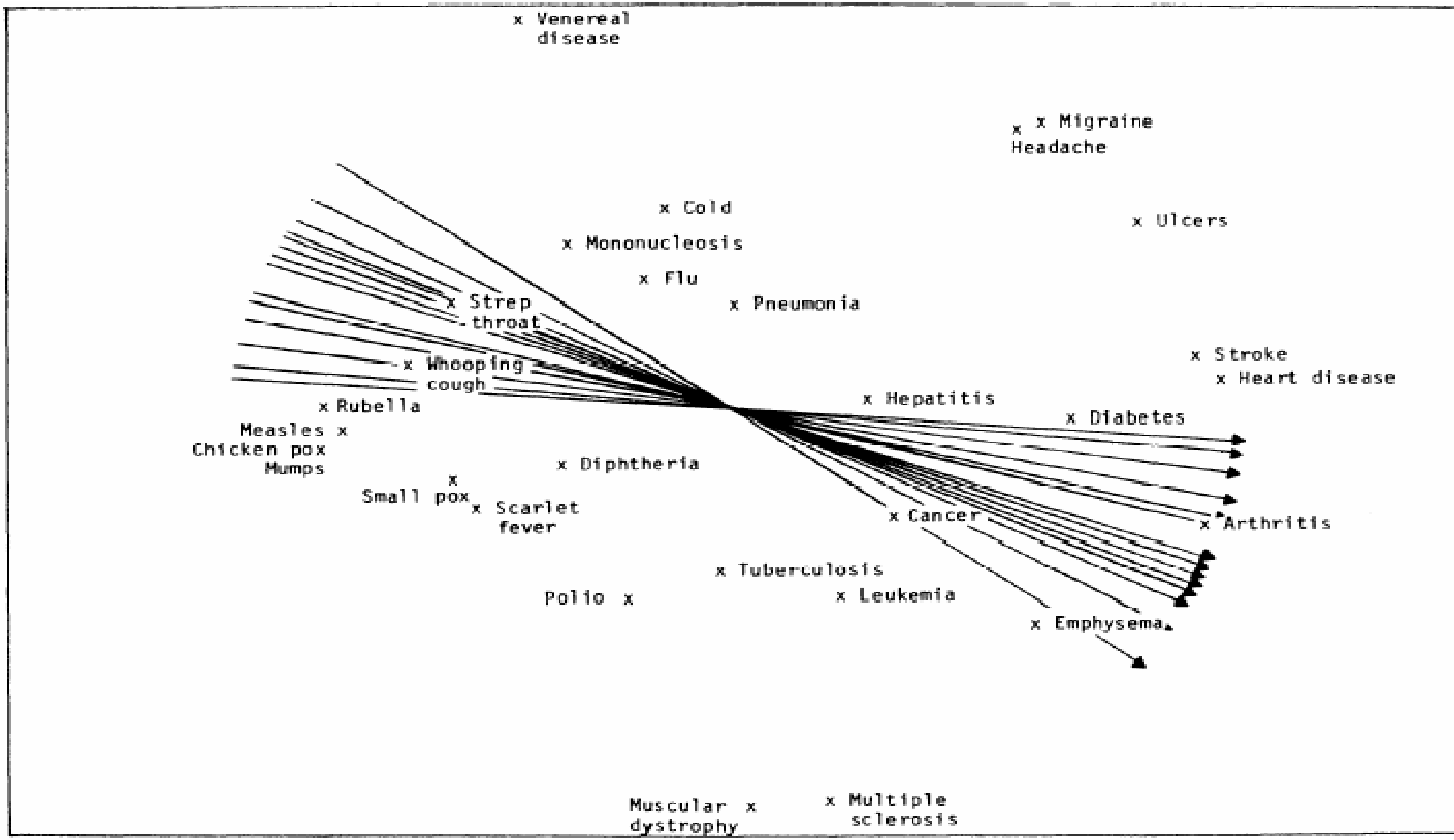
Personality Traits



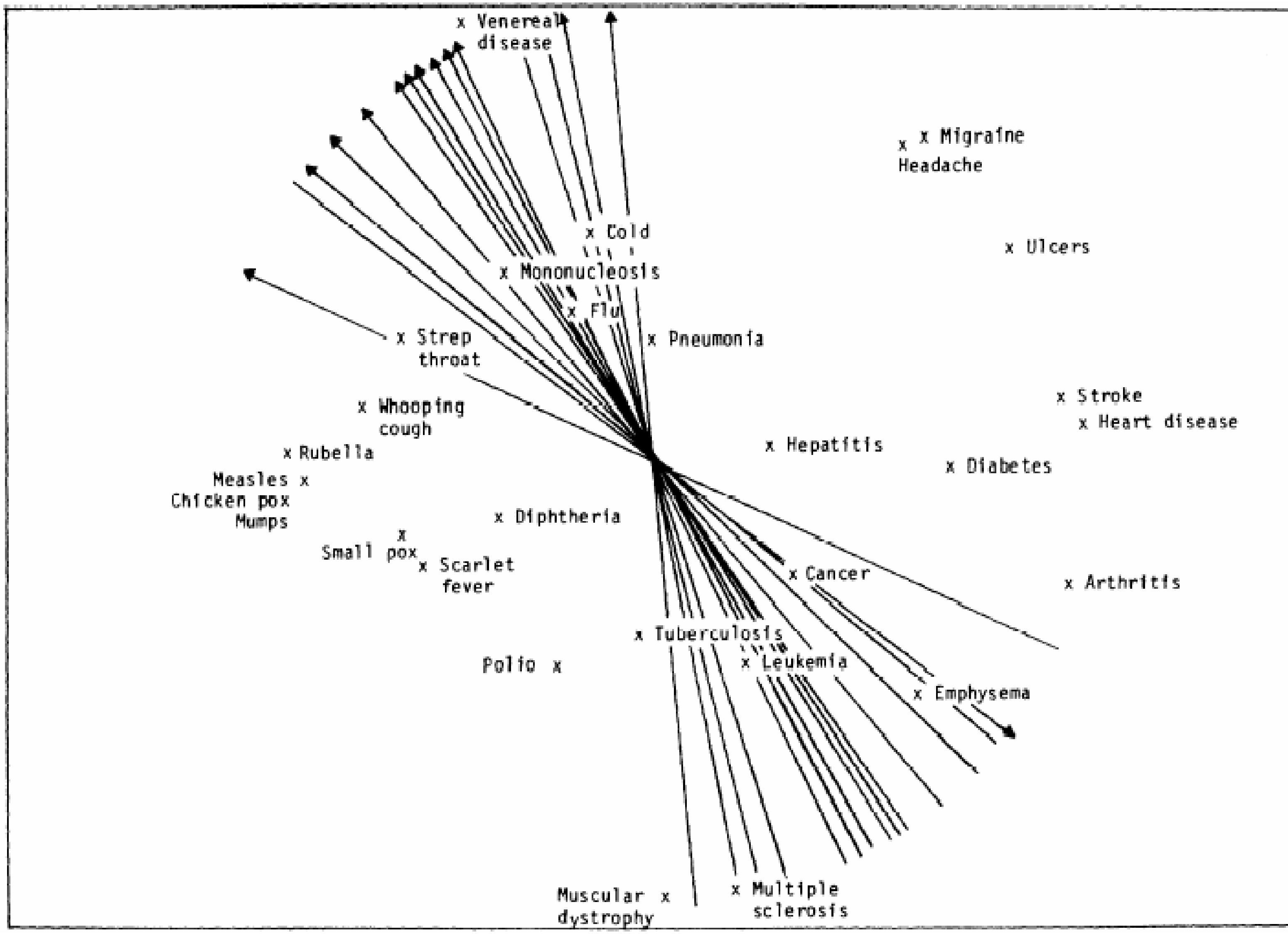
PROFIT

- The cases in the regression are items
- The dependent variable is the average rating of each item on the hypothesized attribute
- Look for significant r-square > 0.80
- If r-square is low, then we can discredit an attribute as being a factor in people's judgments
- If r-square is high, then they <u>may</u> have been using this attribute (or a highly correlated one) in their thinking
- Can also use un-averaged ratings: a different rating vector for each respondent

Contagiousness (US)

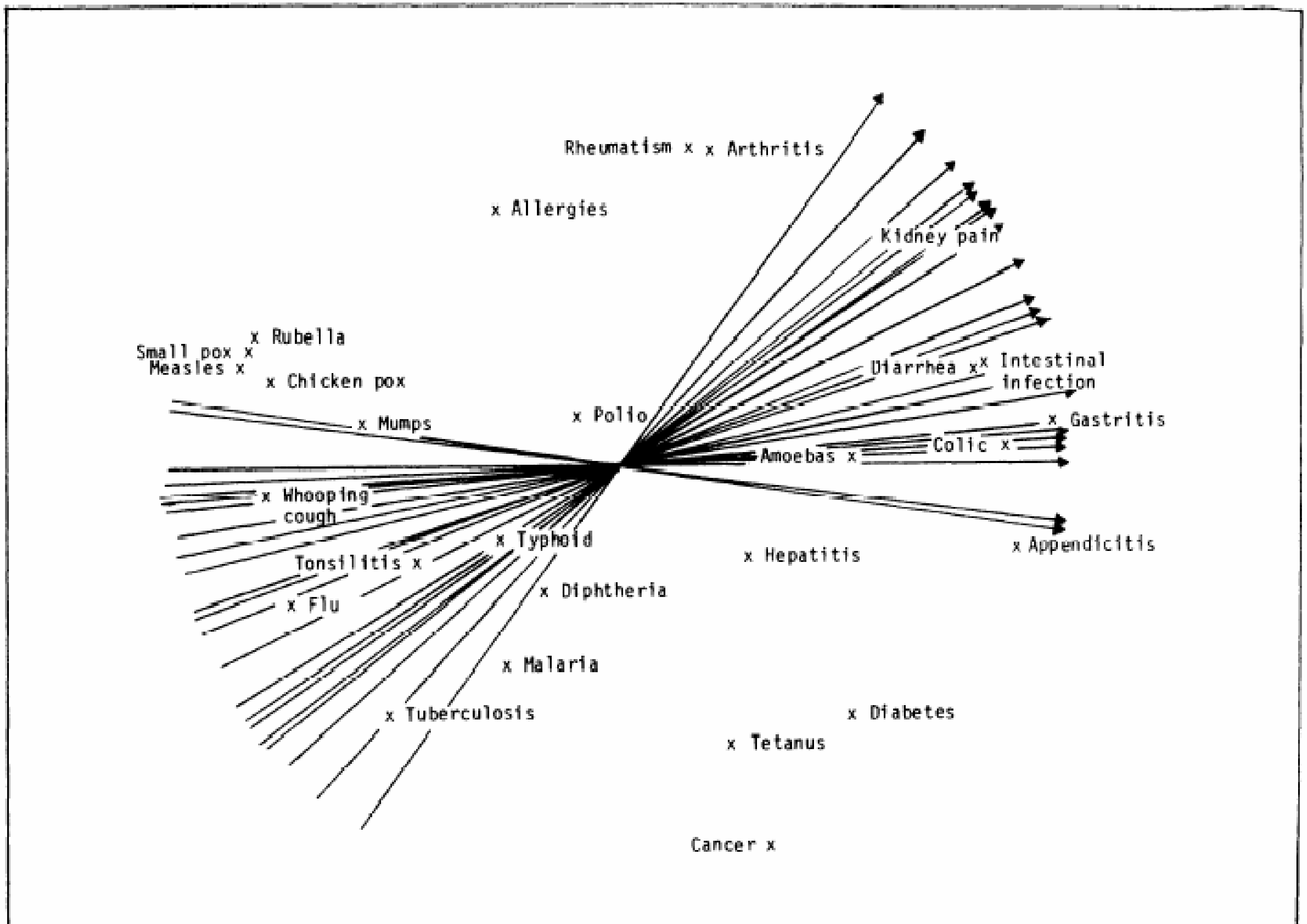


Severity (US)

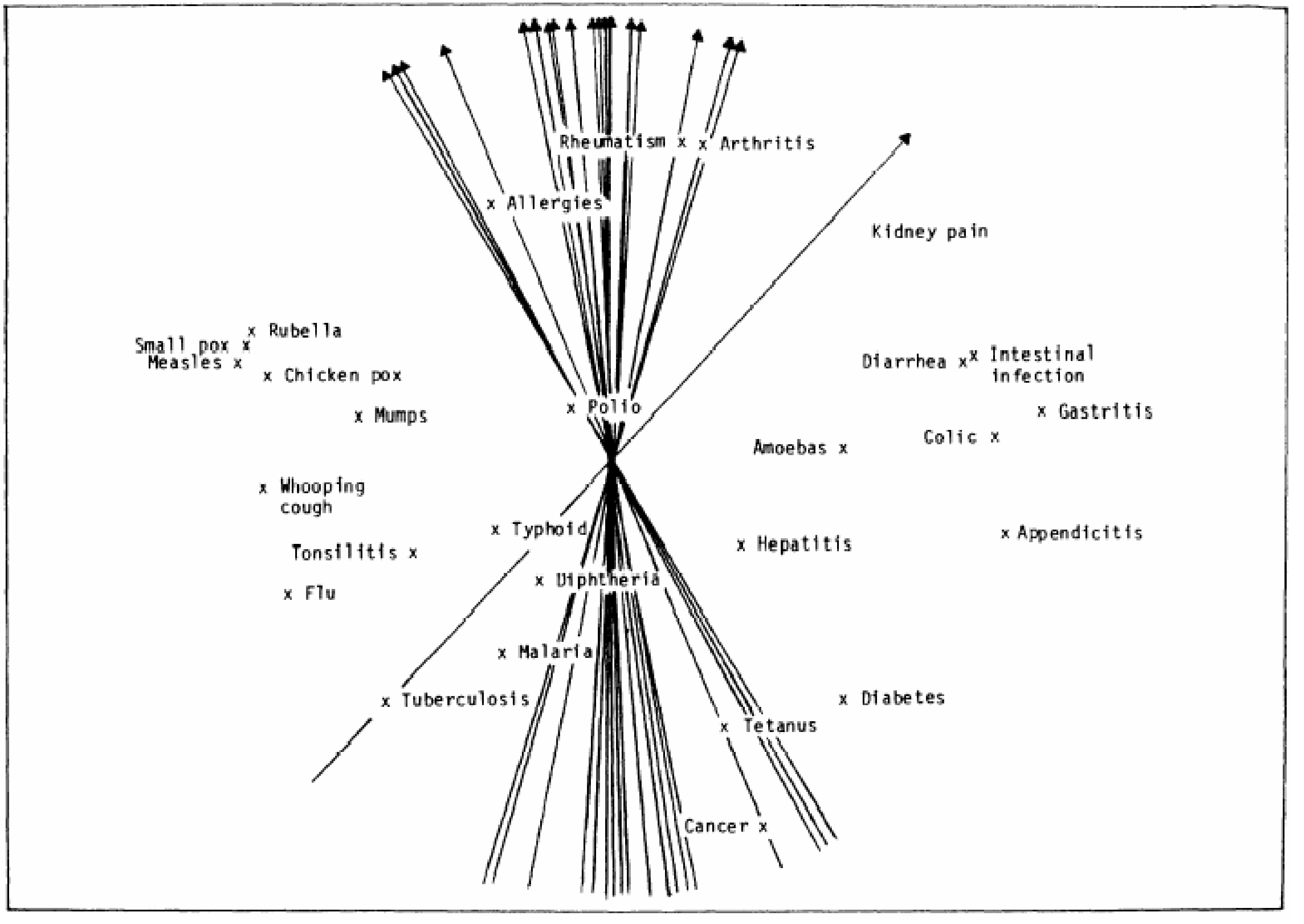


x Venereal disease
x Migraine
x Headache
x Cold
x Ulcers
x Mononucleosis
x Flu
x Strep throat
x Pneumonia
x Stroke
x Heart disease
x Whooping cough
x Hepatitis
x Rubella
x Diabetes
Measles x
Chicken pox
Mumps
x Diphtheria
x Small pox
x Scarlet fever
x Cancer
x Arthritis
x Tuberculosis
Polio x
x Leukemia
x Emphysema
Muscular x dystrophy
x Multiple sclerosis

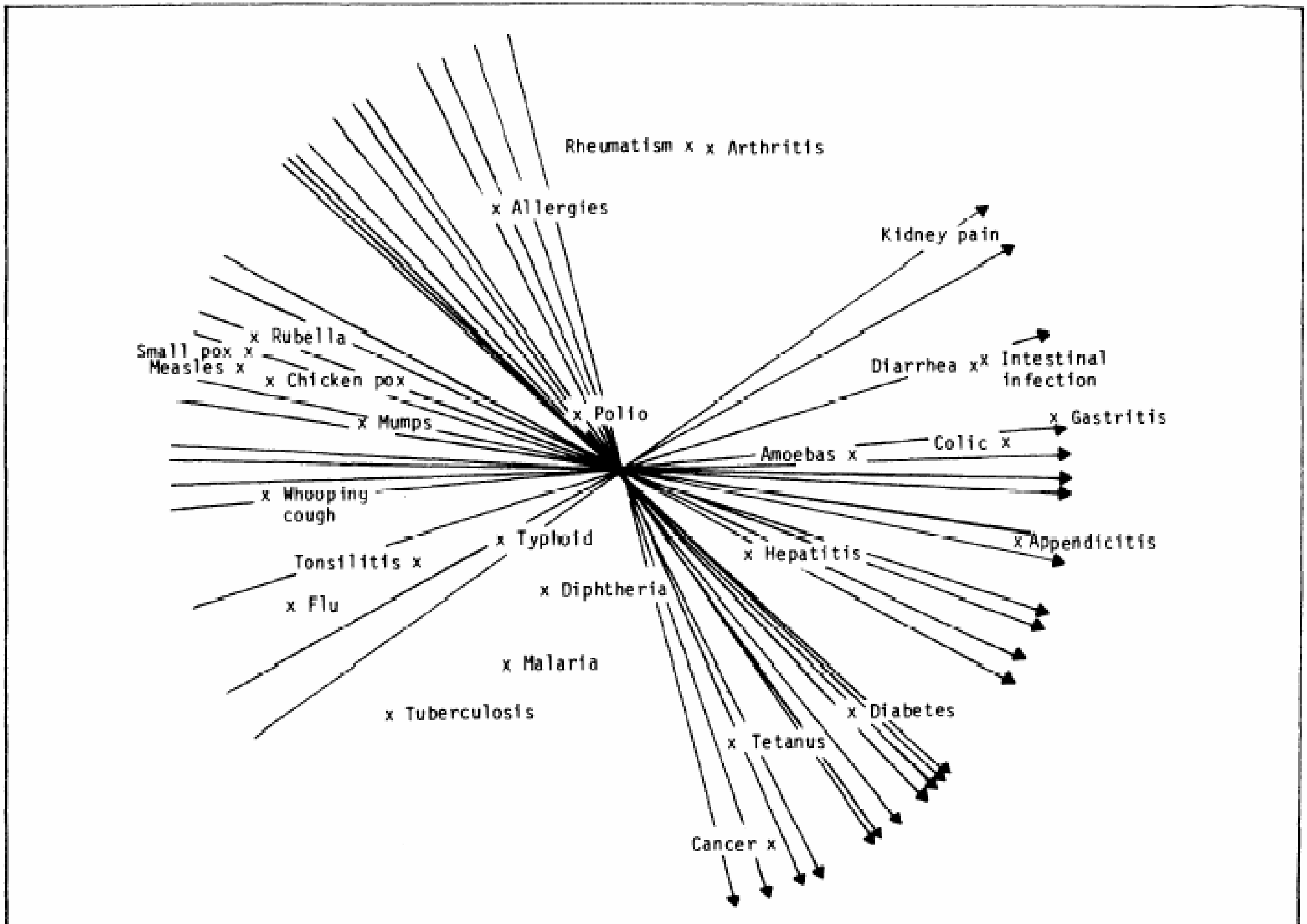
Contagion (Guatemala)



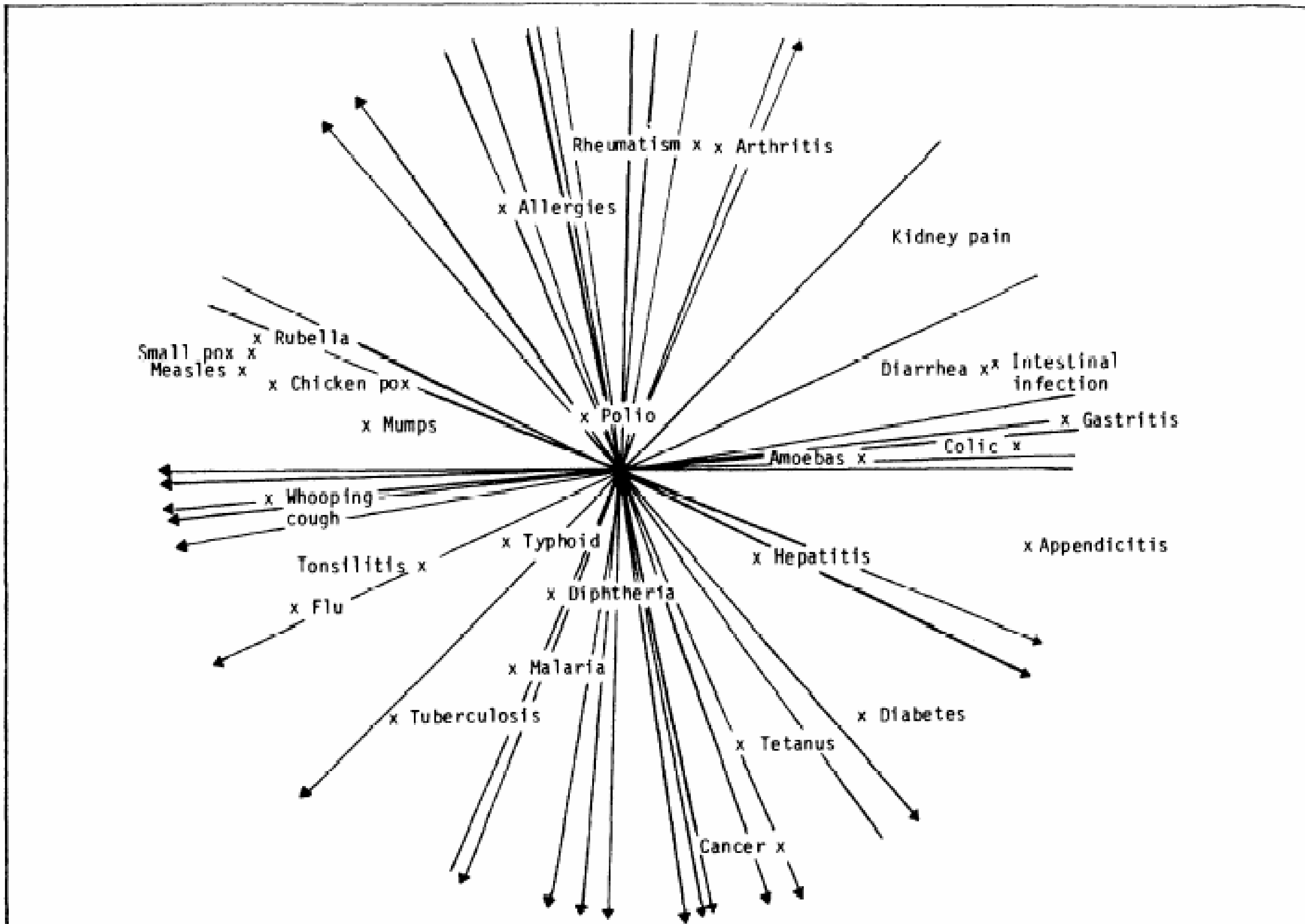
Severity (Guatemala)



Age of the Infirm (Guatemala)



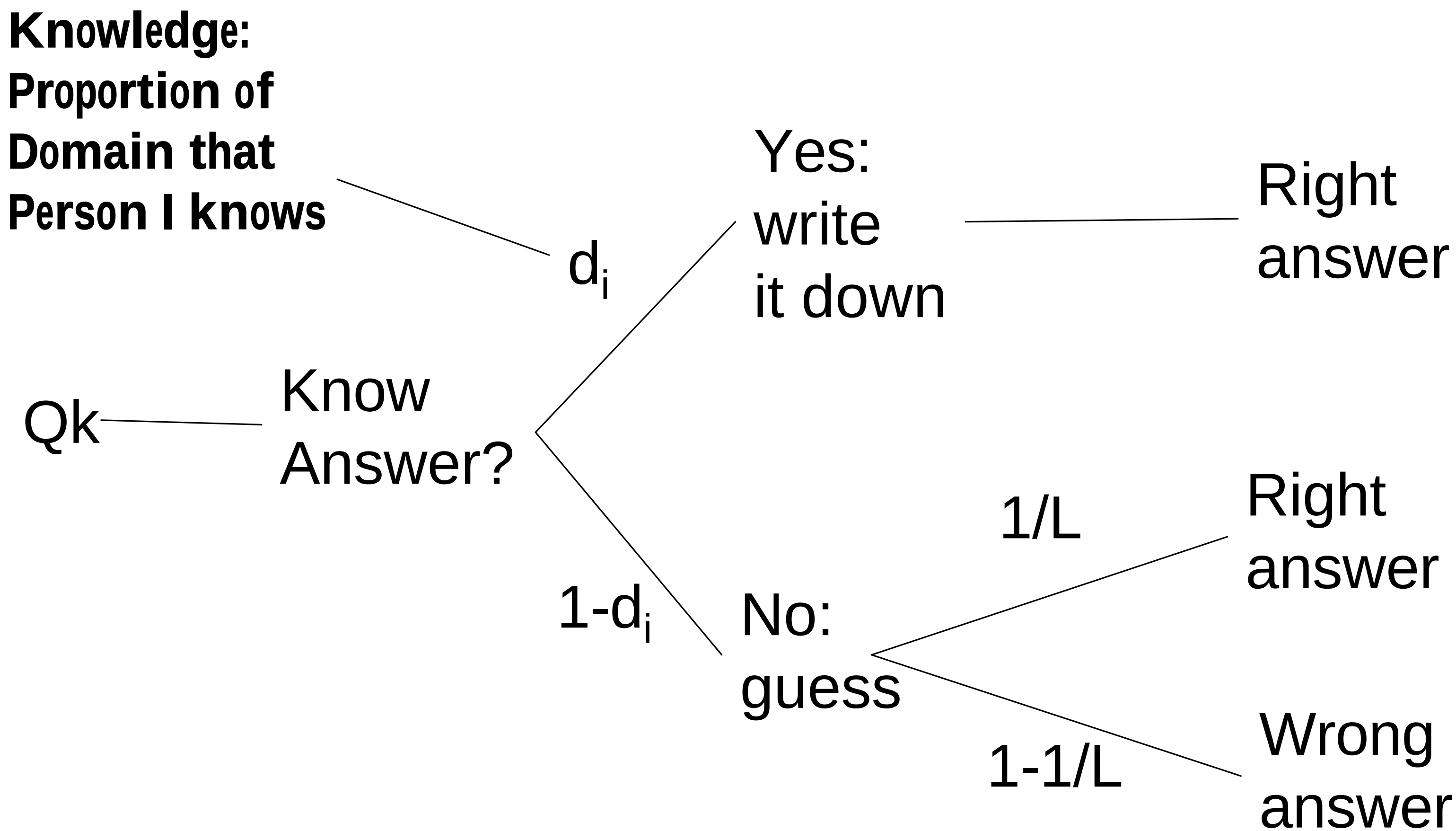
Hot-Cold (Guatemala)



Consensus Analysis

- Is it ok to aggregate across respondents?
  - Only if they belong to same culture – averaging systematically different sets of answers just gets mush
  - Similar to interpreting average of a bi-modal univariate distribution
- Can we tell which respondents know what they are talking about (or have conventional views) and which don’t (are out in left field)?
- Consensus theory of Romney, Weller & Batchelder can help

Response model



$$Prob(correct) = m_i = d_i + \frac{(1-d_i)}{L}$$

L = # of choices In multiple choice question.

Prob of agreement, m_{ij}

(between respondents I and J)

Case	Probability
1. Both know answer	$d_i d_j$
2. I knows and J guesses right	$d_i(1-d_j)/L$
3. J knows and I guesses right	$d_j(1-d_i)/L$
4. Neither knows, both guess the same	$(1-d_i)(1-d_j)/L$

Neither Knows, Guess Same

Person J

Person I

	1	2	...	L	
1	$(1/L)^2$				1/L
2		$(1/L)^2$			1/L
...			$(1/L)^2$		1/L
L				$(1/L)^2$	1/L
	1/L	1/L	1/L	1/L	1

$$(1/L)^2 + (1/L)^2 + \ldots = L(1/L)^2 = 1/L$$

Pairwise agreement m_{ij}

- Agreement m_{ij} is sum of four cases:

$$m_{ij} = d_i d_j + d_i(1-d_j)/L + d_j(1-d_i)/L + (1-d_i)(1-d_j)/L$$
$$m_{ij} = d_i d_j + (1-d_i d_j)/L$$

- Or rearrange terms:

$$(Lm_{ij}-1)/(L-1) = d_i d_j$$

- Agreement between respondents is a multiplicative function of knowledge level of each

Factor Analysis

observed

unknown

- Left side of $(Lm_{ij}-1)/(L-1) = d_i d_j$ is just obs agreement adjusted by constants. If we let $m^*_{ij} = (Lm_{ij}-1)/(L-1)$ then we can write more simply: $m^*_{ij} = d_i d_j$
- We solve for d's by factor analyzing M*
  - Spearman's fundamental equation of factor analysis $r_{ij} = f_i f_j$
    - Corr between two variables is a function of the extent each is correlated with the latent factor

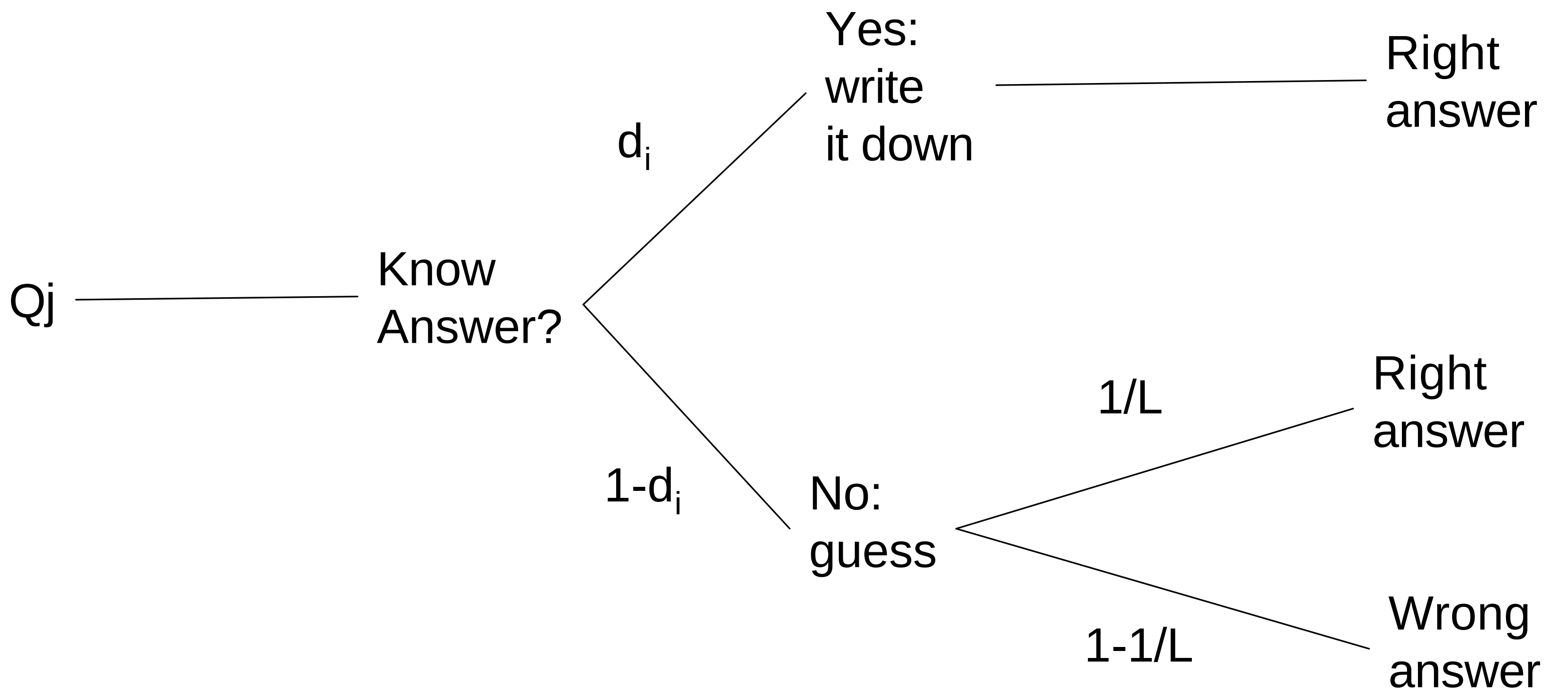
We can figure out how much people know without having an answer key !!!!!!!!!!!!

Inferring knowledge

- Factoring the observed agreement matrix M* solves for the unknown values d_i
  - The d values given by the factor loadings
- The d values are the amount of knowledge each person has
  - Literally, the correlation of the person's responses with the unknown answer key
- So factoring the agreement matrix gets us exact estimates of the amount of knowledge each person has
  - And no answer key is needed!!!
  - Exactly what we were looking for

What's the catch??

- The response model must be right



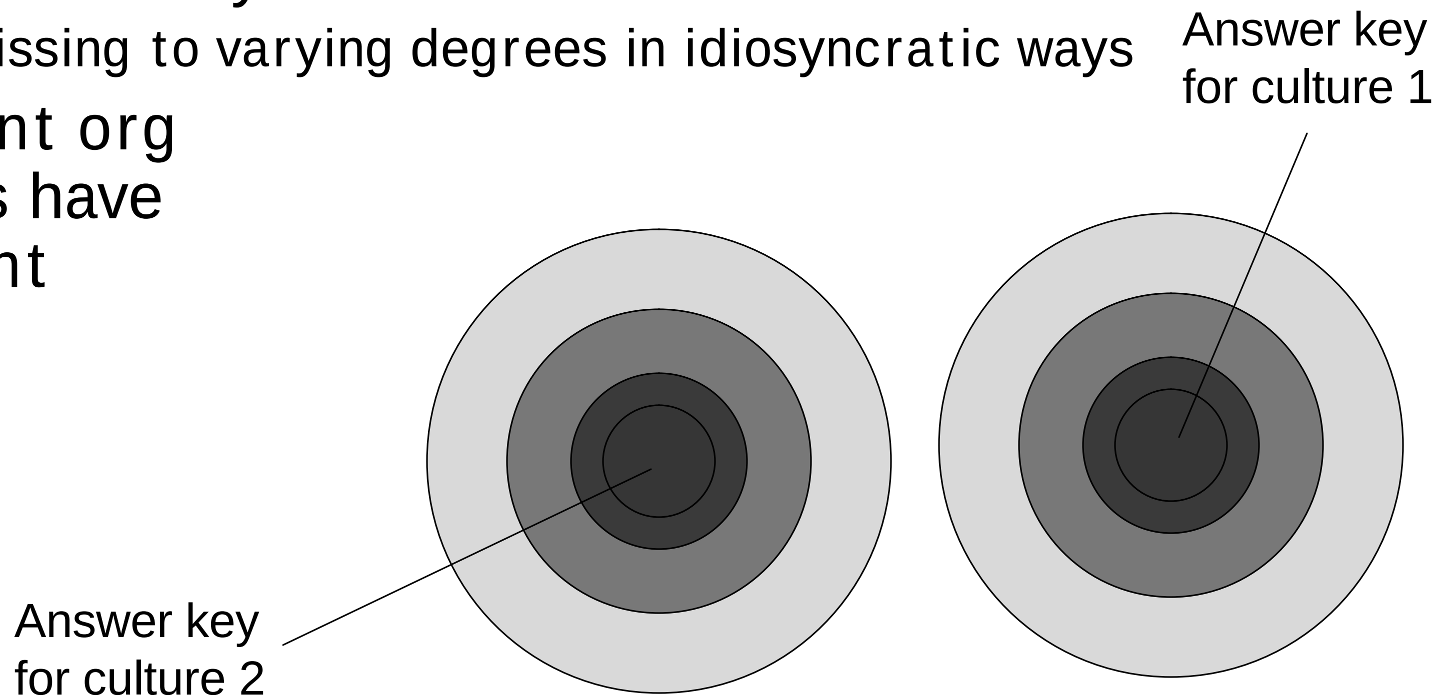
- Can characterize this model as follows

Three conditions

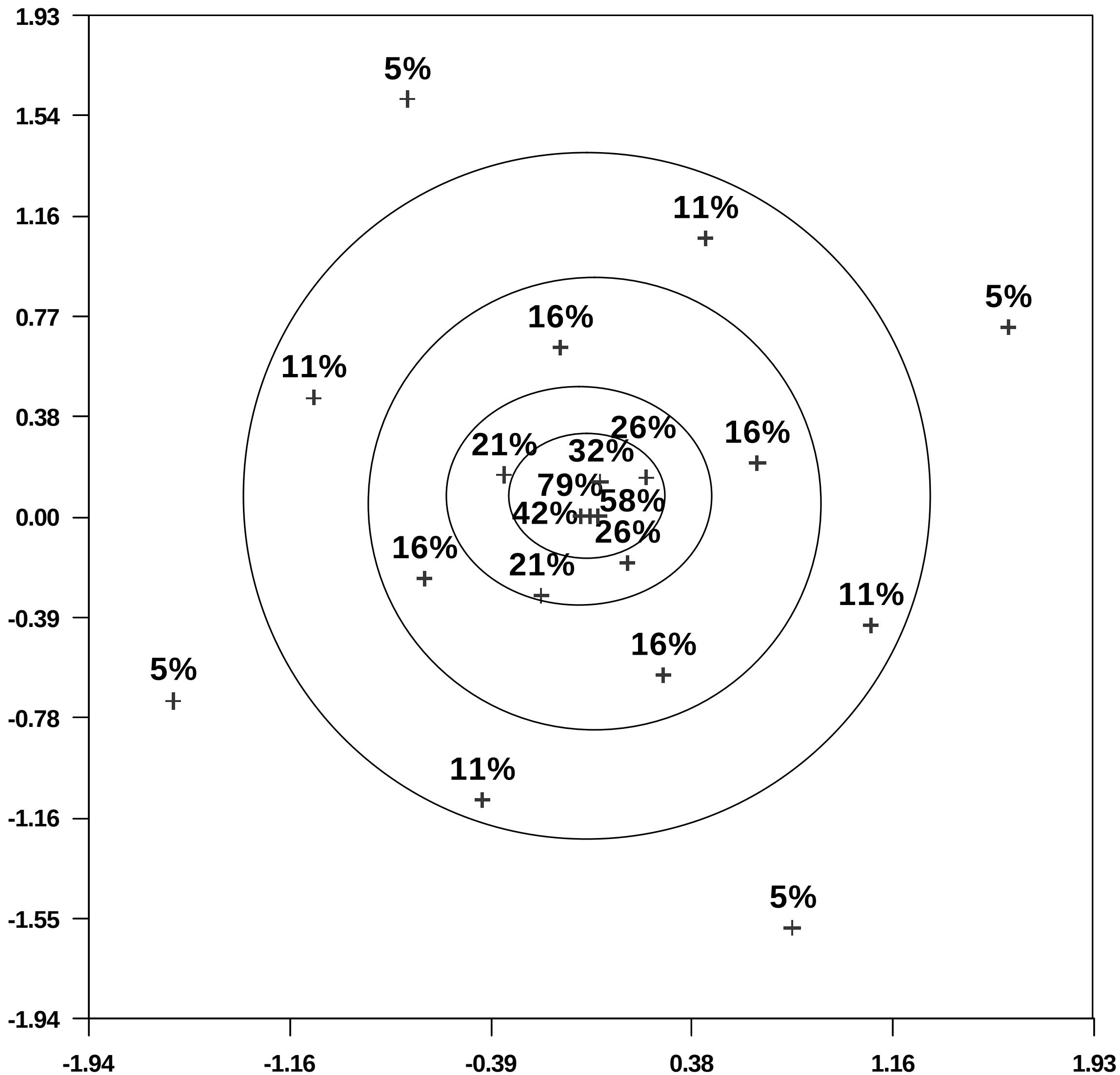
- Common Truth
  - each question has exactly one right answer, applicable to entire sample of respondents
    - Sample drawn from one pop w/ same answer key
- Local Independence
  - resp-item response variables x_{ij} are independent, conditional on the truth
- One Domain
  - All questions drawn from same domain, i.e.:
    - can model knowledge w/ one parameter, d_i

Bullseye Model

- Two people agree to the extent that each is correlated with the truth
  - Truth is culturally correct answer key
- Each member of culture is aiming at same answer key
  - but missing to varying degrees in idiosyncratic ways
- Different org cultures have different targets



Expected Agreement Pattern



Partitioning variability

- Model identifies two sources of variability in responses (beliefs)
  - Cultural: multiple answer keys
  - Individual: variation in knowledge
- Within each culture, we still expect (and can measure), variability due to differential access to information, ability, etc.

Test of consensus model

- Undergraduate class with 92 students
- Multiple choice final exam with 50 questions
- Instructor's answer key provides gold standard to compare against
- Each student asked to guess test score of all acquaintances, including self

Measures

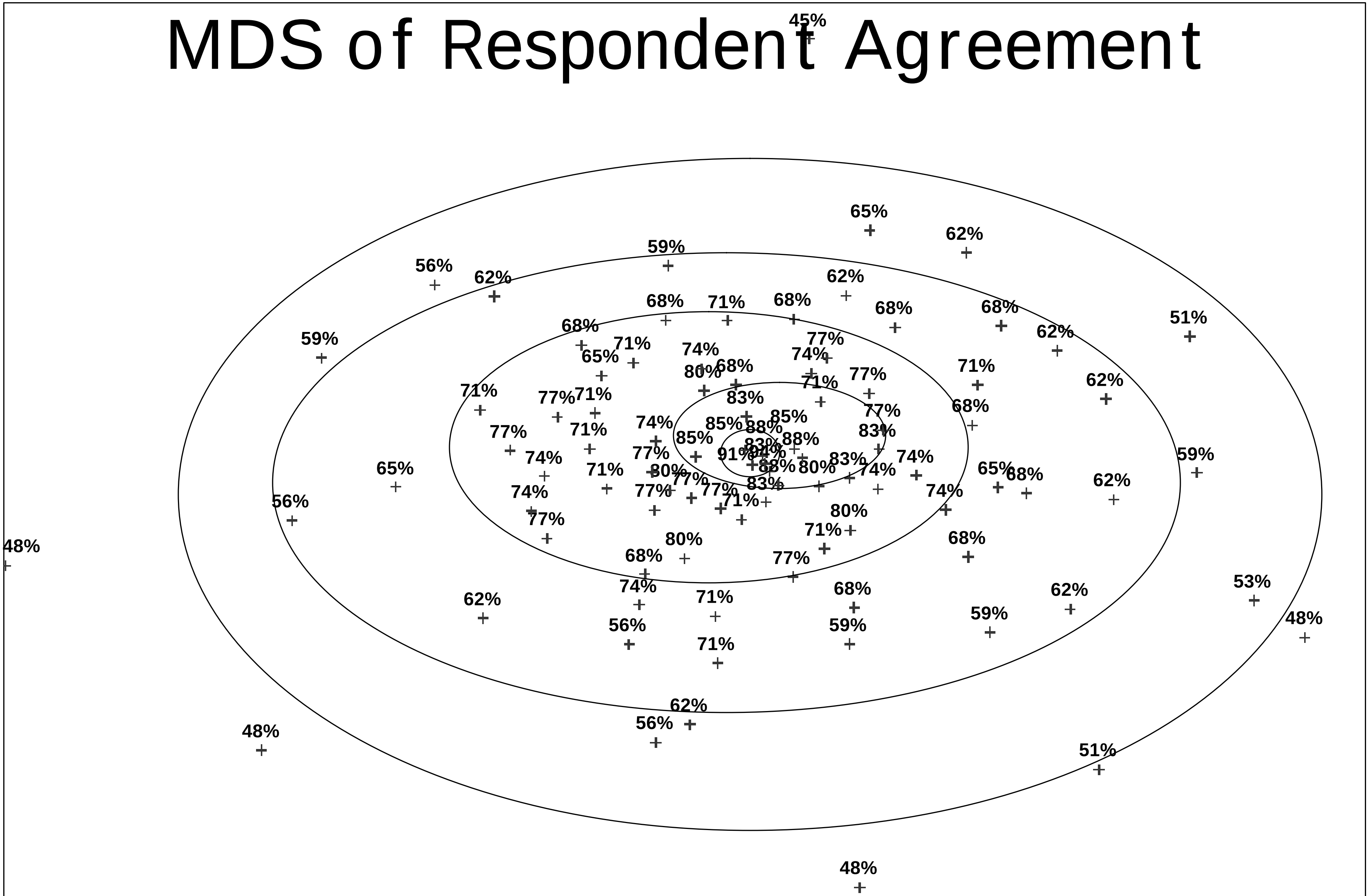
- Self-report model
  - Each person's estimate of their own score
- Network model
  - for each person, use average estimate of their scores (persons with fewer than 5 acquaintances were excluded)
    - All acquaintances
    - Only friends
- Consensus model
  - Factor loadings of minimum residual factor analysis of student-by-student agreement matrix
- Gold standard
  - % correct based on instructor's answer key

Factor Analysis of Agreements

Factor	Eigenval	Percent	Cum %	Ratio
1	51.323	93.6	93.6	28.308
2	1.813	3.3	96.9	1.065
3	1.702	3.1	100	

- Results consistent w/ single answer key
  - therefore we can use loadings to estimate knowledge

MDS of Respondent Agreement

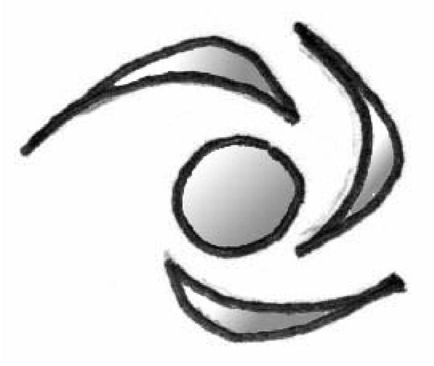


Correlations

	Gold	Self	Acquaint	Friends	Consen
Gold	1.000				
Self	0.479	1.000			
Acquaint	0.334	0.564	1.000		
Friends	0.398	0.556	0.891	1.000	
Consen	0.947	0.471	0.342	0.400	1.000

- Consensus estimates virtually identical to gold standard (r = 0.947)
- Self-report better than network model

Running Consensus



Summary

- CDA is about mapping structure of emic domains
- Data collection relies on text statements or simple categorical judgments
  - Listing terms
  - Piling, choosing most different, choosing greater of two items
- Analysis uses sophisticated computational techniques but mostly delivers pictures