Finding patterns in data

Data

Git Repository: **FindingPatterns** folder

https://github.com/shh-dlce/qmss-2017/tree/master/FindingPatterns

http://tinyurl.com/qmssData2

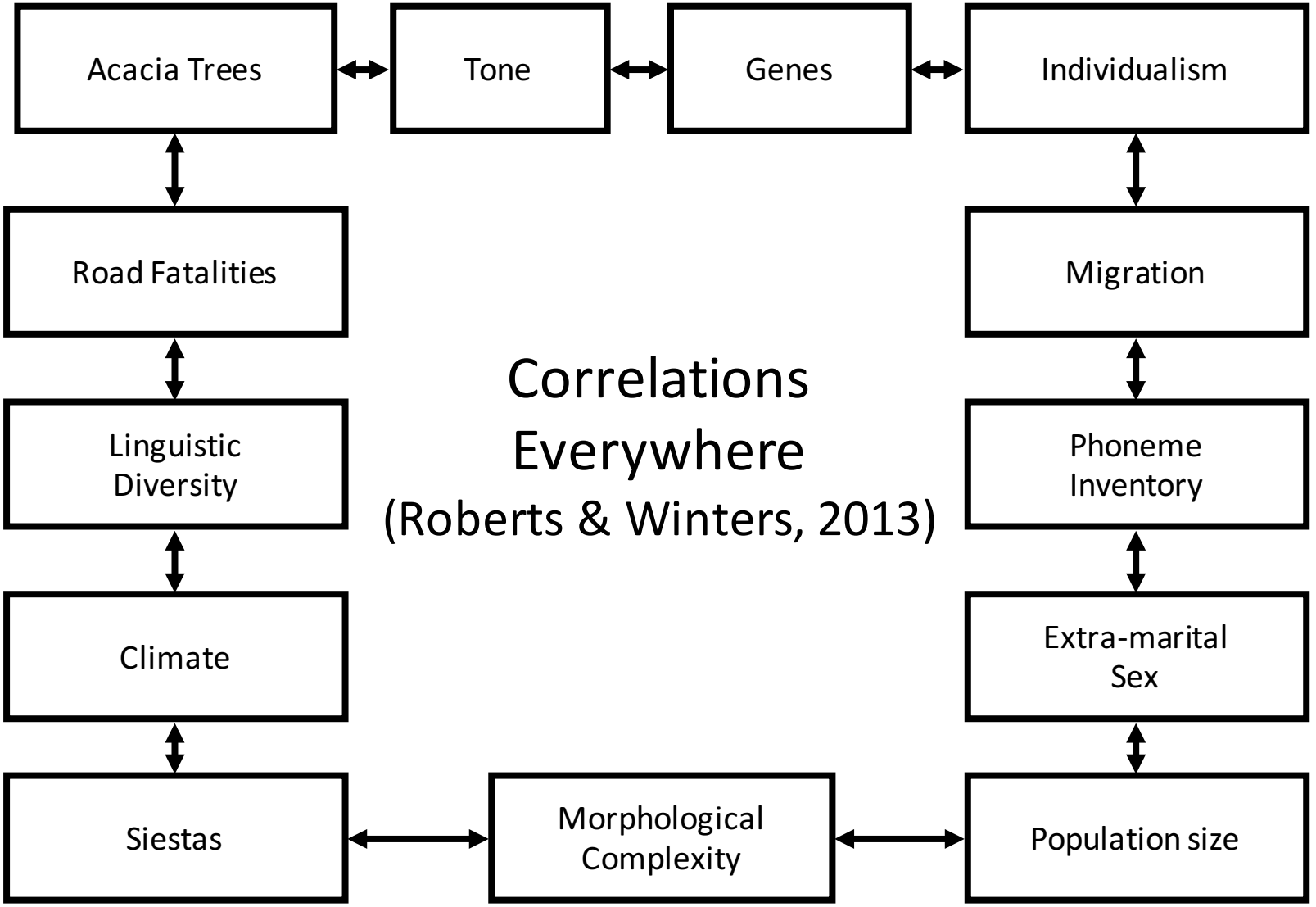
Finding patterns

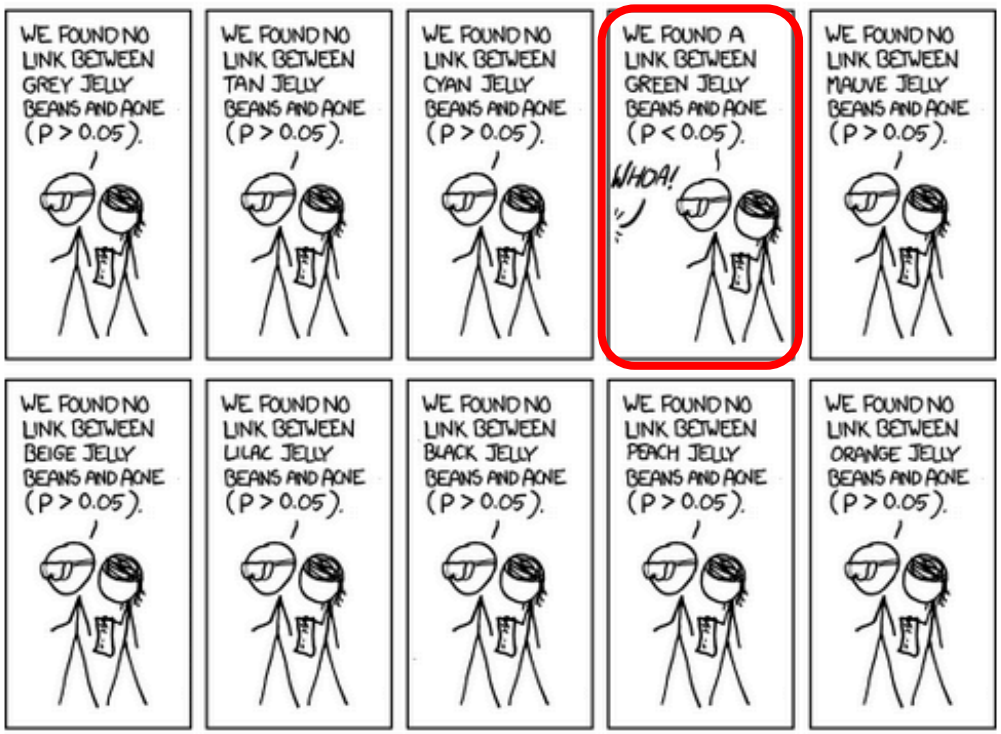
Visualising relationships

- General patterns

Finding patterns:

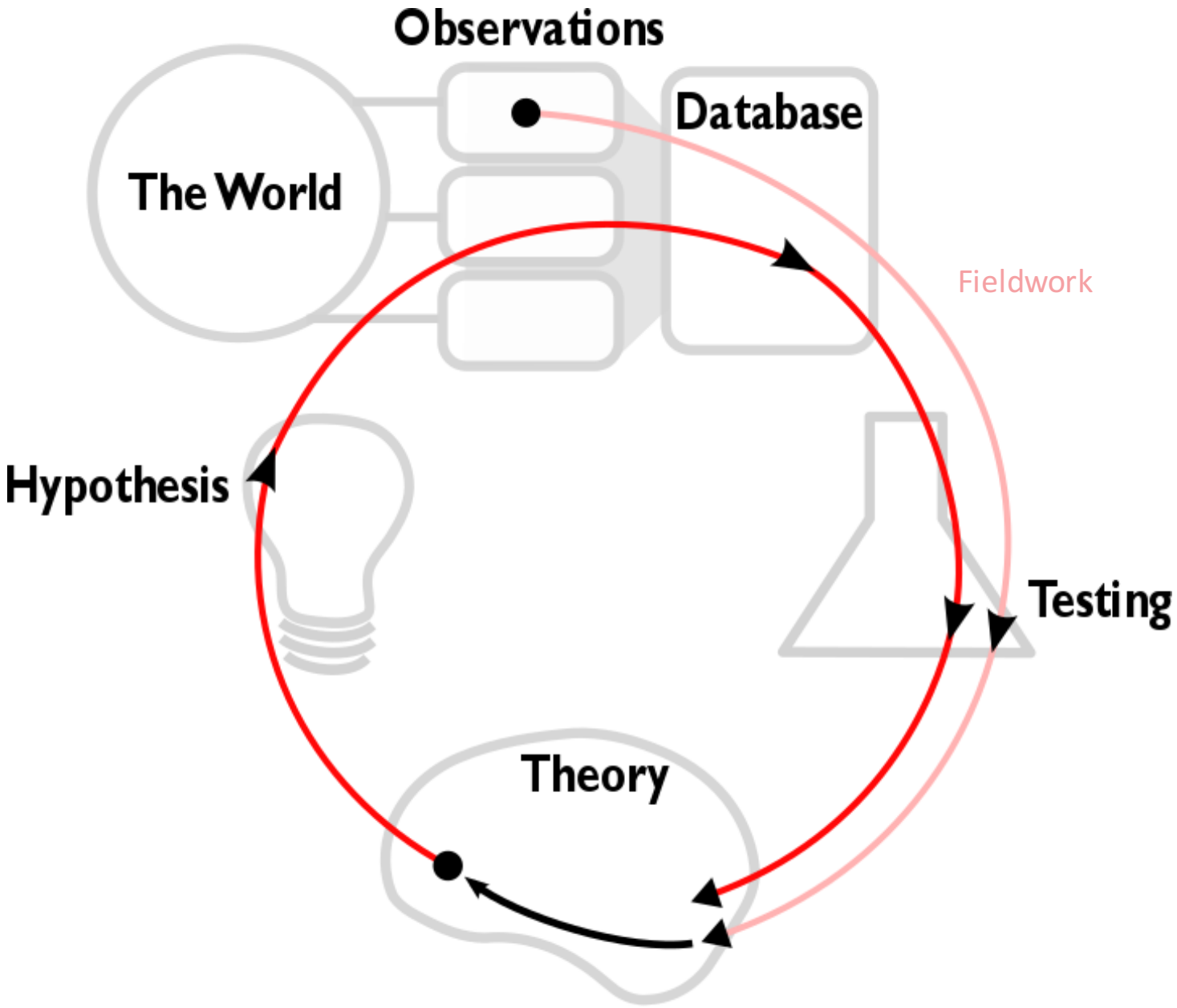
- Principle components
- Decision trees
- Causal graphs

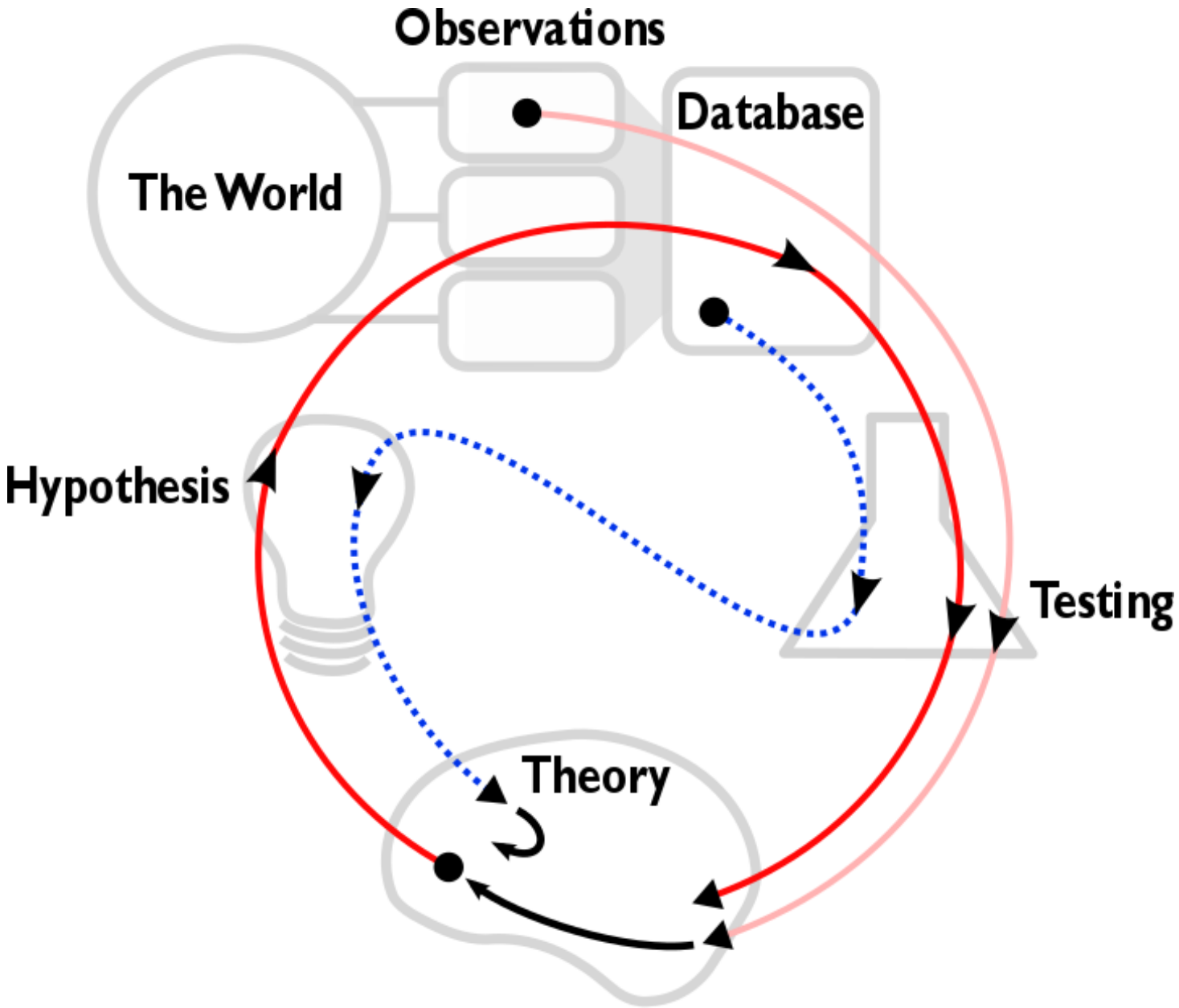

Acacia Trees
Tone
Genes
Individualism
Road Fatalities
Migration
Linguistic Diversity
Phoneme Inventory
Climate
Extra-marital Sex
Siestas
Morphological Complexity
Population size
Correlations Everywhere
(Roberts & Winters, 2013)

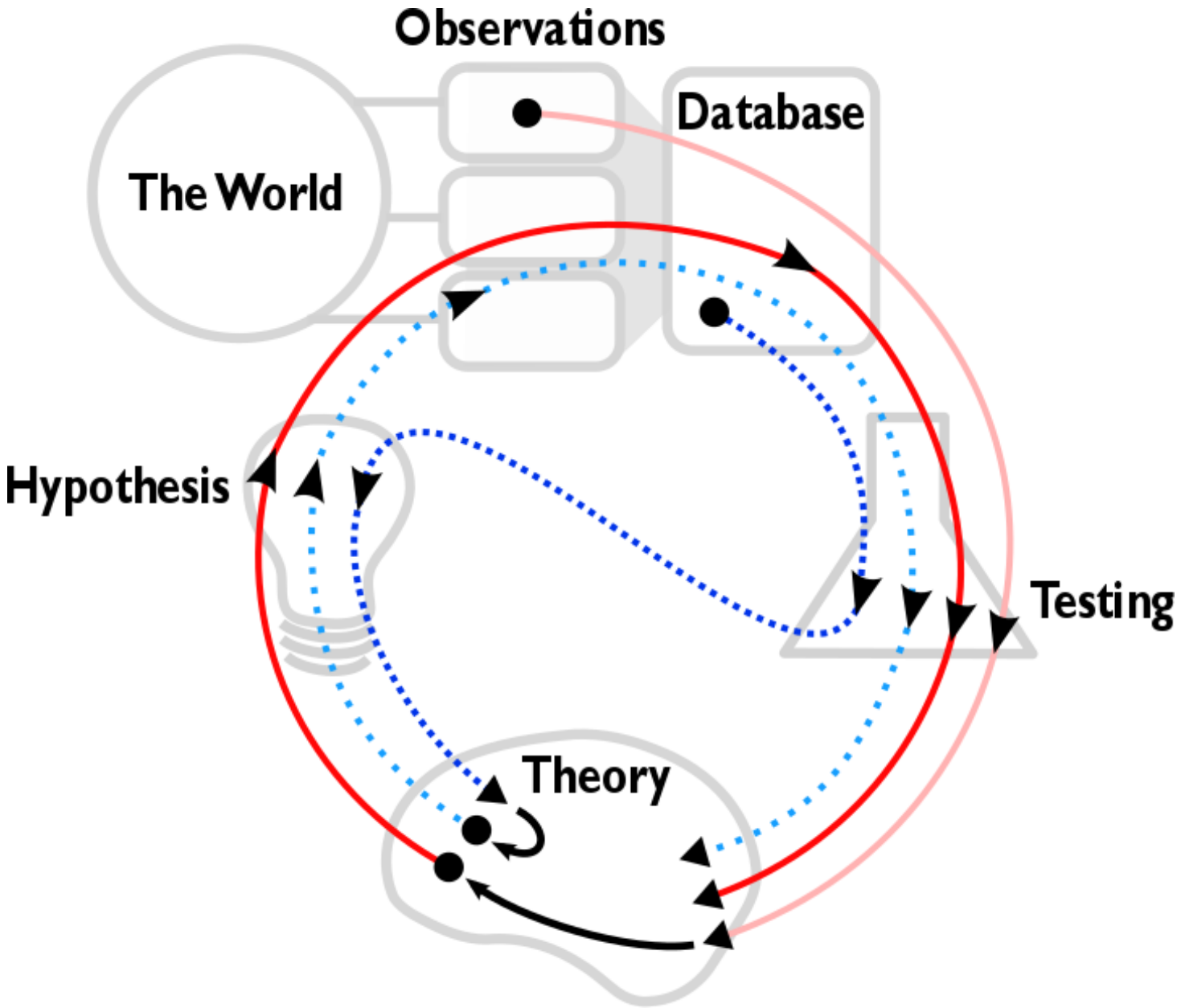

WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE

When should we search for patterns?

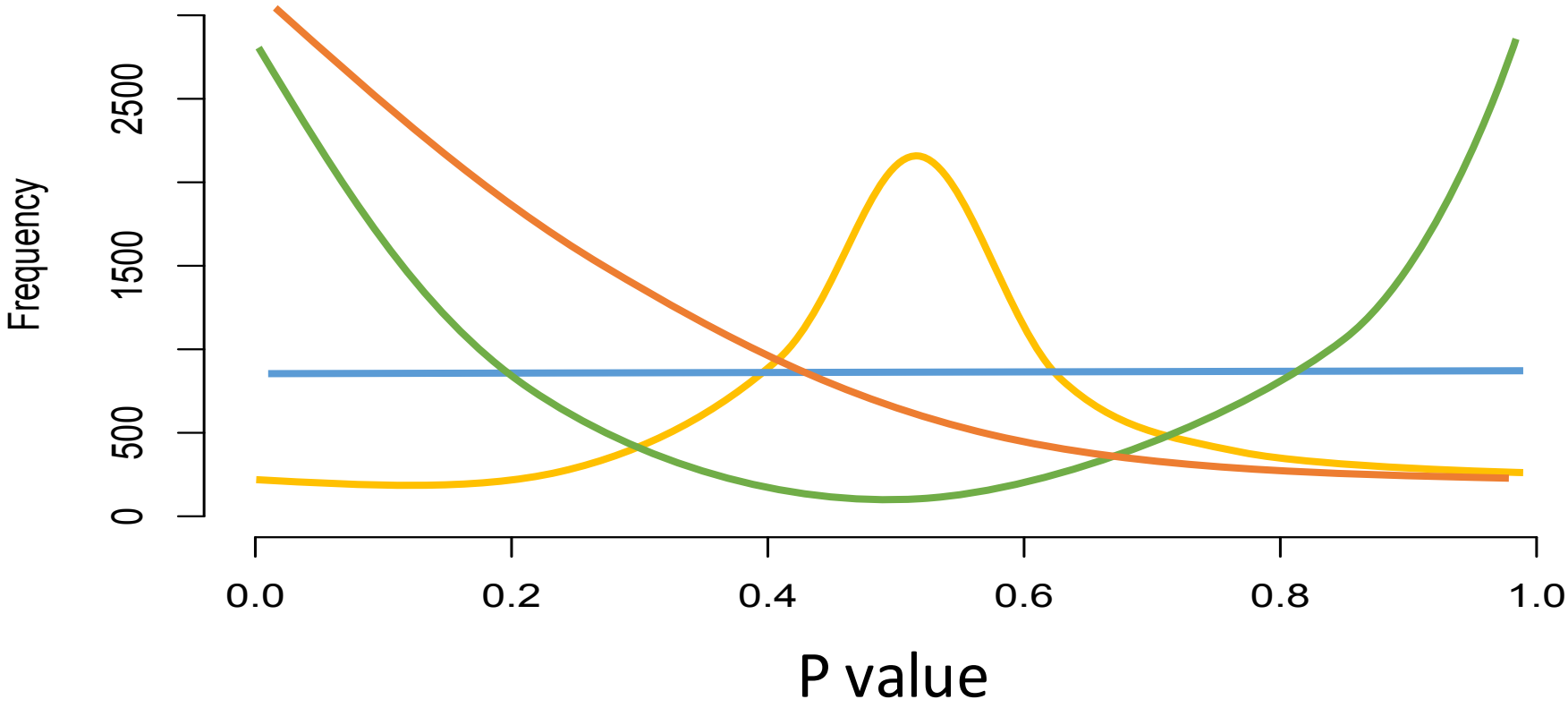
- To understand how confounding variables are related
- When there's no other way to visualise the data
- To contrast two general theories
- As feasibility studies


Observations
Database
The World
Fieldwork
Hypothesis
Testing
Theory

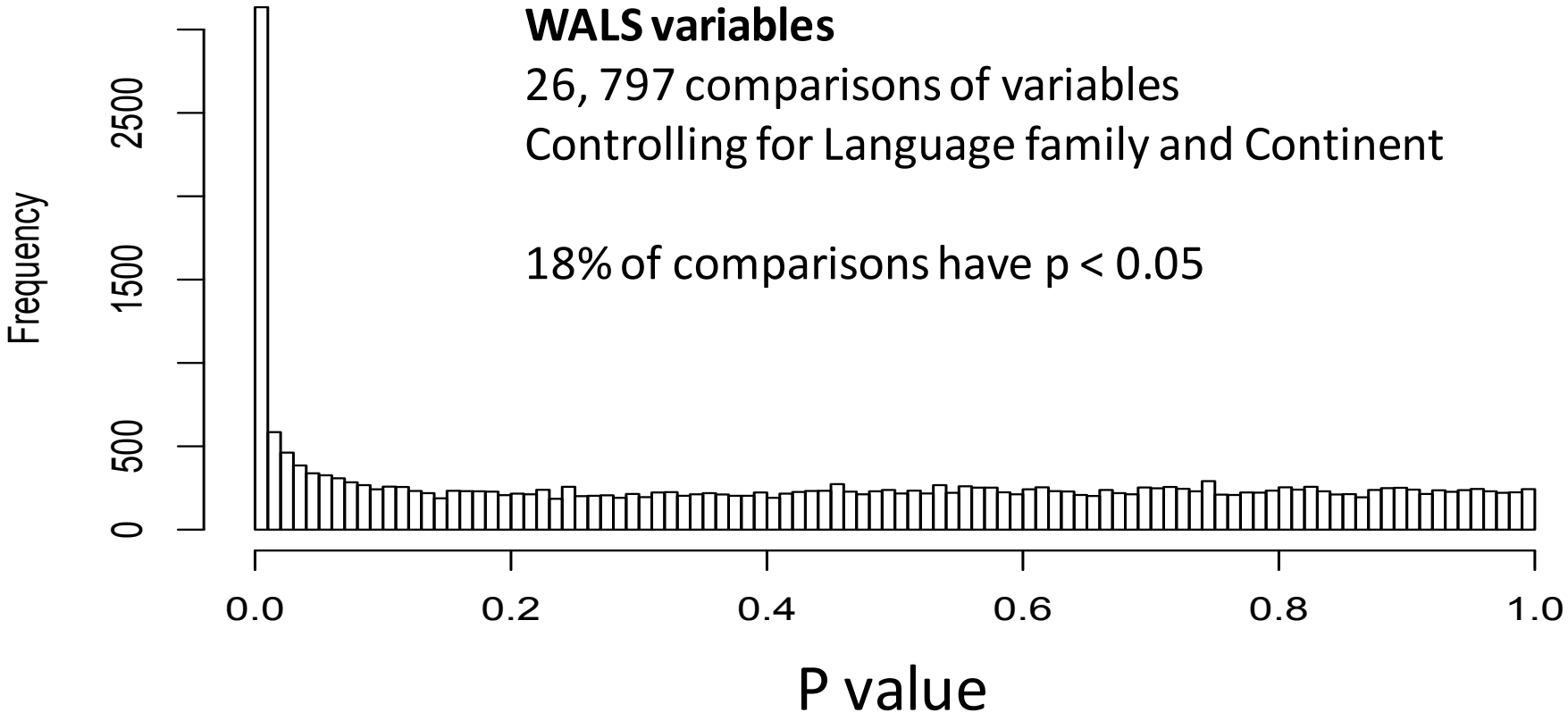

Observations
Database
The World
Hypothesis
Testing
Theory


Observations
Database
The World
Hypothesis
Testing
Theory

Serendipity



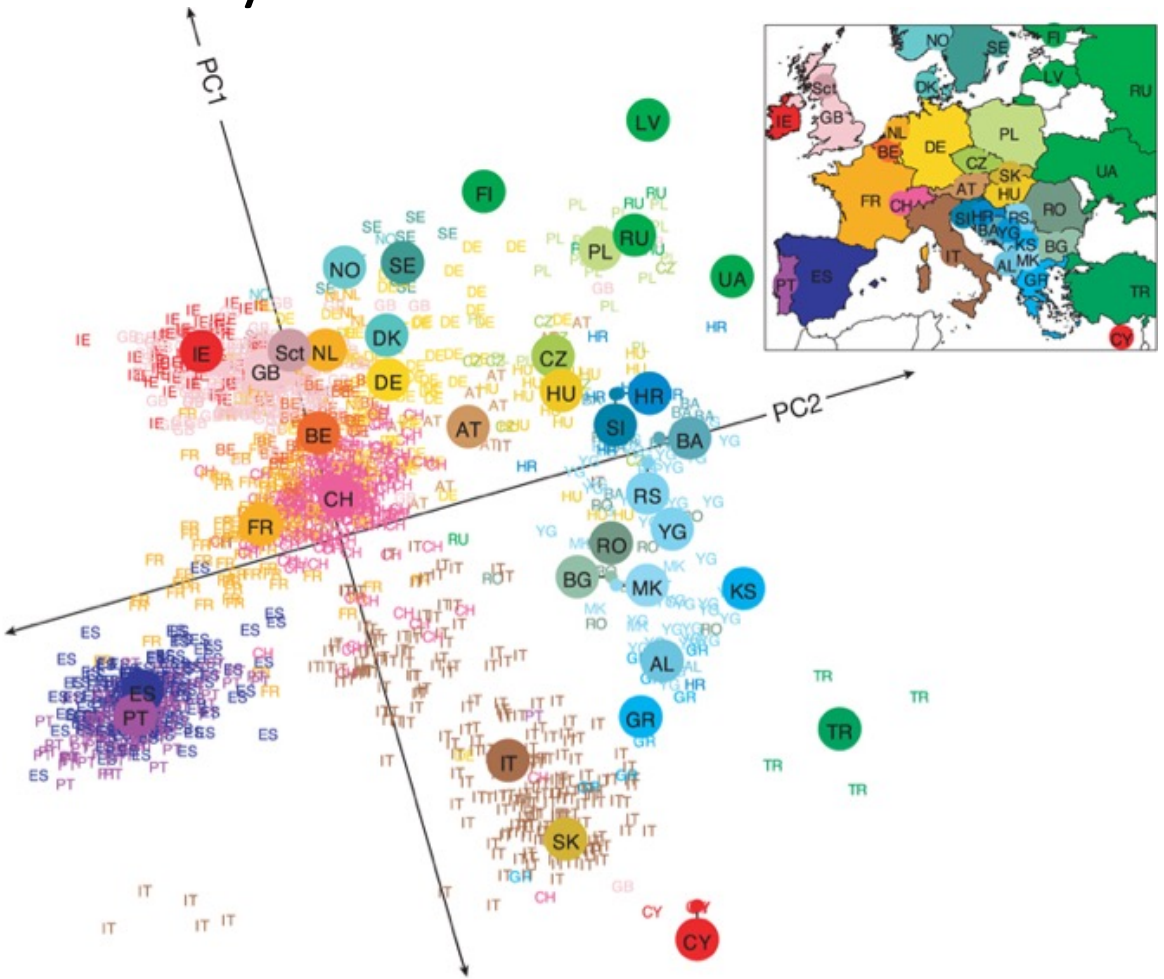
Serendipity



Principle component analysis

Compress many variables into two dimensions

In a way that preserves the actual distances between data points.



Example

Word Data

FindingPatterns/data/wordData.csv

- Log Frequency
- Age of acquisition
- Length
- Age
- POS
- Arousal
- Dominance
- Valence
- Concreteness

```
scaled.d = scale(d)
pca <- prcomp(scaled.d)
biplot(pca)
```

analyseWords_PCA.R

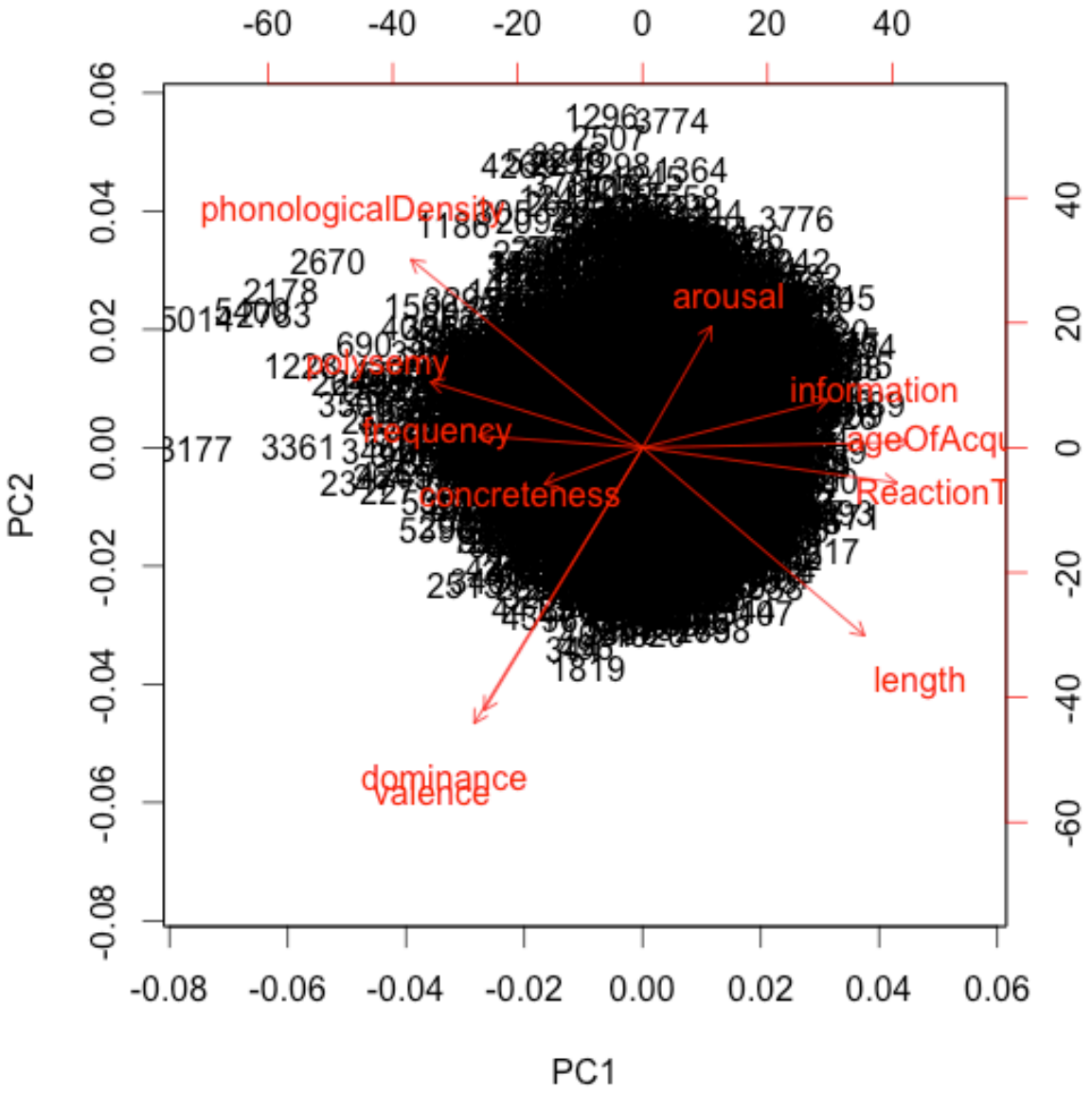
https://github.com/shh-dlce/qmss-2017/blob/master/FindingPatterns/analysis/analyseWords_PCA.R

Variance inflation factor

```
library(usdm)
usdm::vif(d)
```

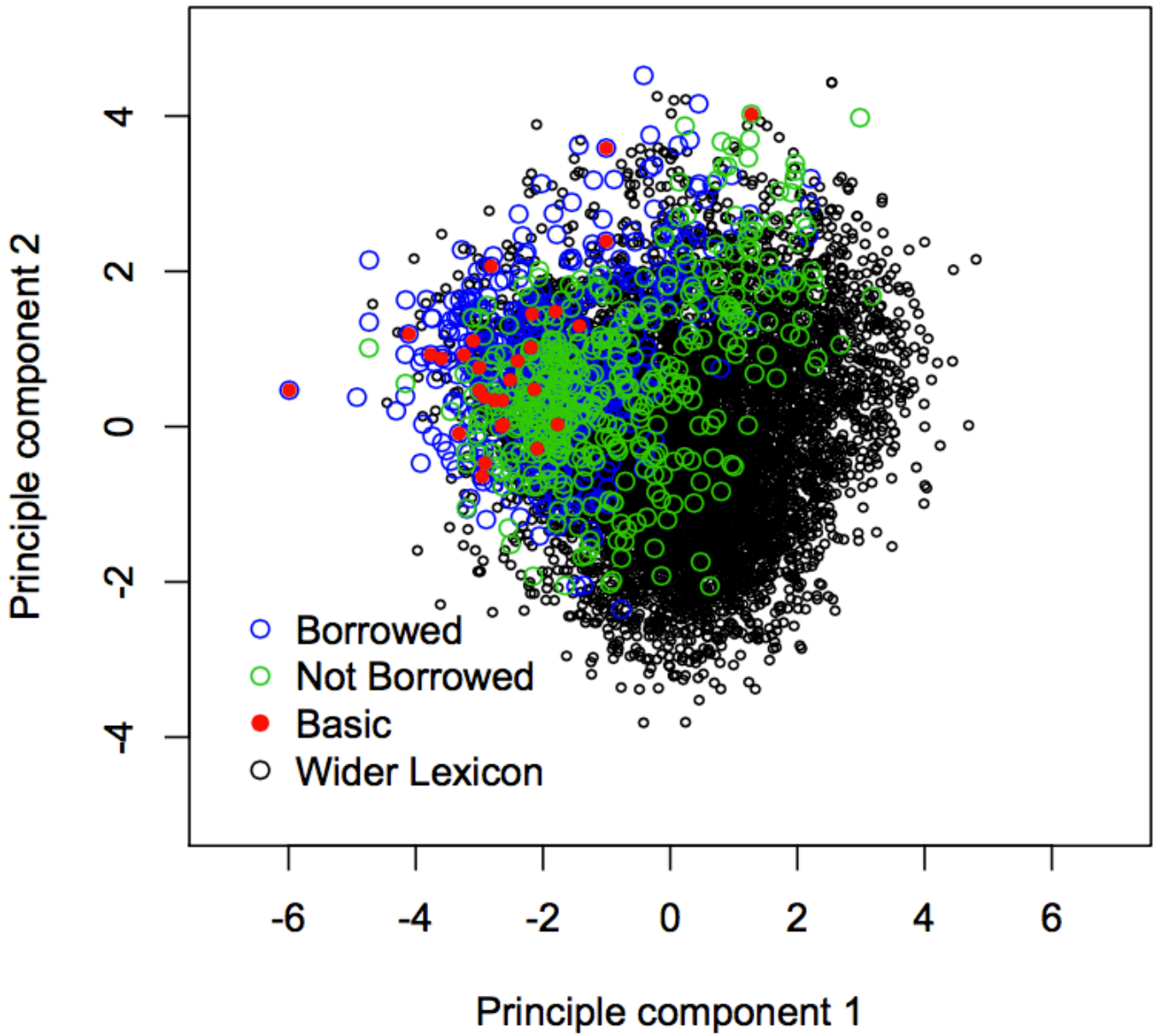
Rule of thumb: Multicolinearity is high if VIF > 10

(variable is 3.1 times as large as it would be if that predictor variable were uncorrelated with the other predictor variables)


-60
-40
-20
0
20
40
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.08
PC2
40
20
0
-20
-40
-60
-0.08
-0.06
-0.04
-0.02
0.00
0.02
0.04
0.06
PC1
phonologicalDensity
arousal
polysemy
information
frequency
concreteness
length
dominance
valence

Word Data

Log Frequency
Age of acquisition
Length
Age
POS
Arousal
Dominance
Valence
Concreteness

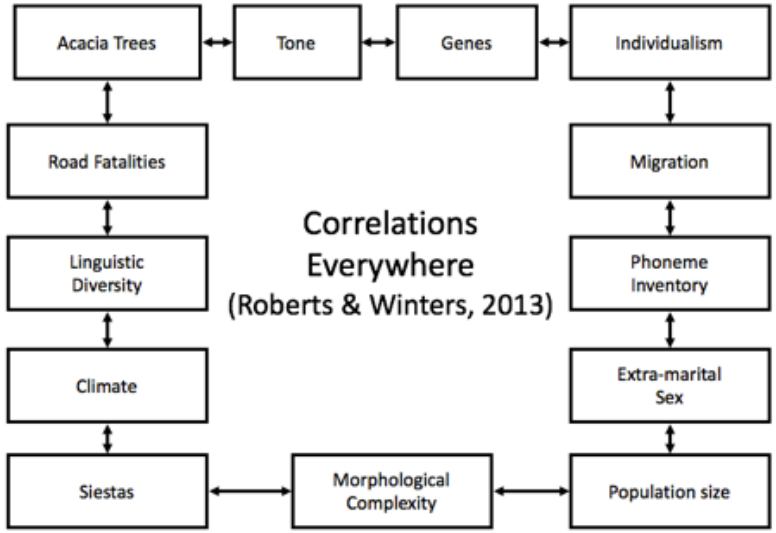


Problem

I have many possible variables I could use, which one is the best?

There are many possible confounding variables, which ones should I include?

Does my effect might only apply in certain parts of the data?



Decision trees

Find the most efficient set of questions for identifying clusters in the data

One dependent variable vs many independent variables

Data Type: Continuous, Categorical

Assumptions: No missing data

Advantages

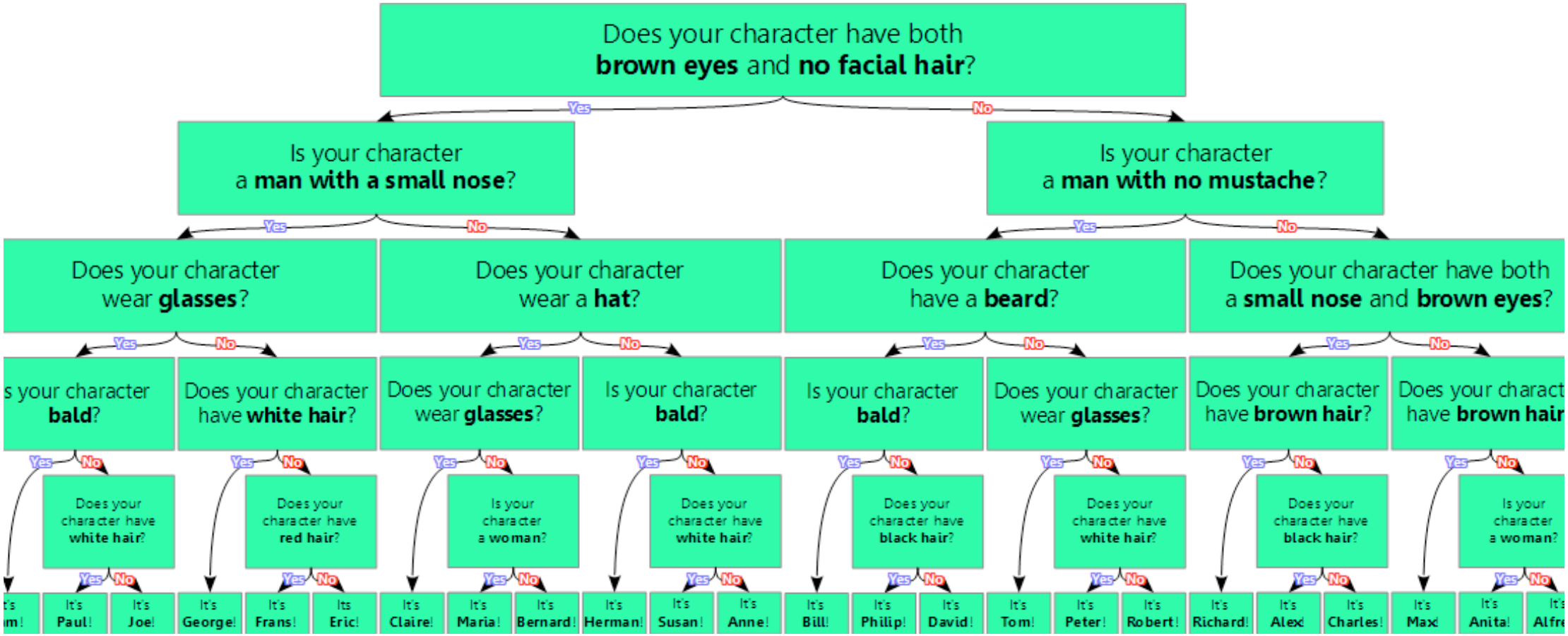
Flexible, easy to implement

Can handle multiple variables, multicolinearity

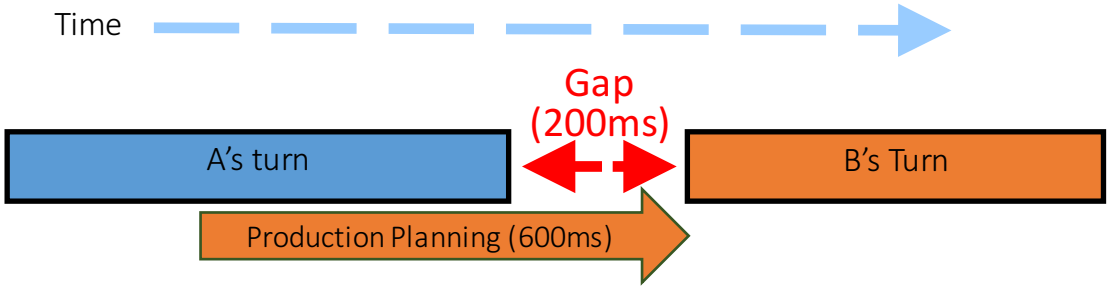
Easy to interpret


David
Eric
Frans
George
Herman
Joe
Maria
Paul
Peter
Philip
Richard
Robert
Sam
Susan
Tom
Robert
MB


MB
GAMES
GUESS WHO?
THE FLIP 'N' FIND
FACE GAME
6+


Does your character have both **brown eyes** and **no facial hair**?
Yes
No
Is your character a **man with a small nose**?
Is your character a **man with no mustache**?
Yes
No
Yes
No
Does your character wear **glasses**?
Does your character wear a **hat**?
Does your character have a **beard**?
Does your character have both a **small nose** and **brown eyes**?
Yes
No
Yes
No
Yes
No
Yes
No
s your character **bald**?
Does your character have **white hair**?
Does your character wear **glasses**?
Is your character **bald**?
Is your character **bald**?
Does your character wear **glasses**?
Does your character have **brown hair**?
Does your charact have **brown hair**
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Does your character have **white hair**?
Does your character have **red hair**?
Is your character **a woman**?
Does your character have **white hair**?
Does your character have **black hair**?
Does your character have **white hair**?
Does your character have **black hair**?
Is your character **a woman**?
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
t's **m**!
It's **Paul**!
It's **Joe**!
It's **George**!
It's **Frans**!
It's **Eric**!
It's **Claire**!
It's **Maria**!
It's **Bernard**!
It's **Herman**!
It's **Susan**!
It's **Anne**!
It's **Bill**!
It's **Philip**!
It's **David**!
It's **Tom**!
It's **Peter**!
It's **Robert**!
It's **Richard**!
It's **Alex**!
It's **Charles**!
It's **Max**!
It's **Anita**!
It's **Alfre**

Rapid Turn Taking



Response time relies on processing

Response time relies on pragmatics

Antje Meyer

Stephen Levinson

Switchboard corpus


Francisco Torreira


Stephen Levinson

348 conversations between 231 speakers, ~ 31 h
19,754 Turn transitions

Variables:

For turn before and after the gap:

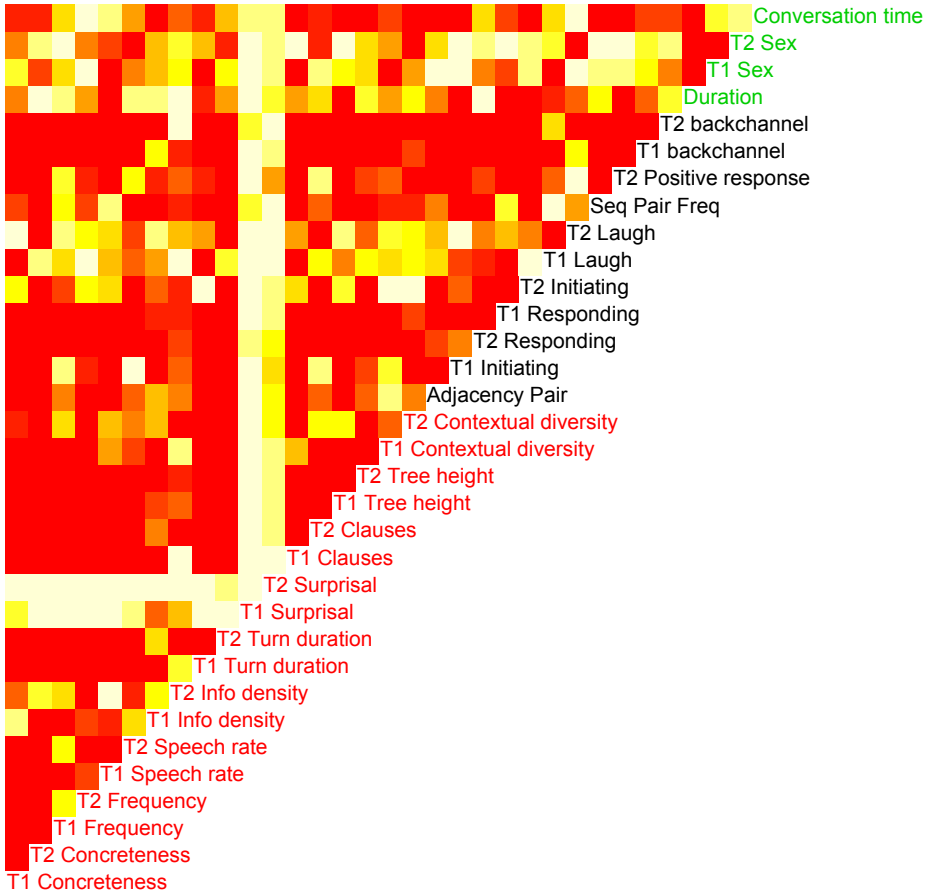
Concreteness (Brysbaert)
Frequency (Subtlex)
Speech rate (Praat)
Surprisal (Piantadosi)
Information density
Turn duration (pympi)
Number of Clauses
Syntax tree height
Turn duration

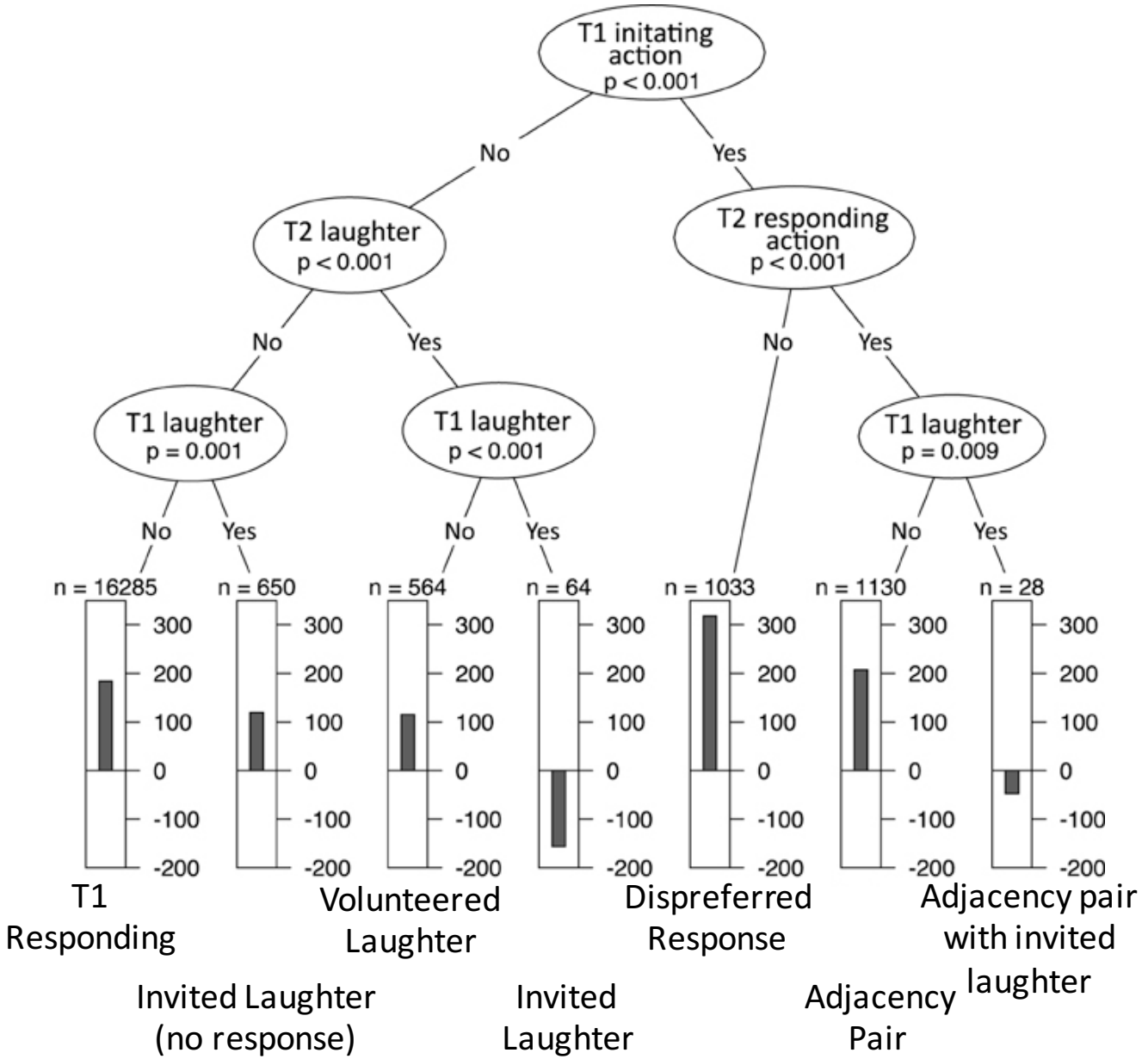
Pragmatic action (initiating, responding)
Sequential position
Frequency of speech act pair
Laughter
Valence of response

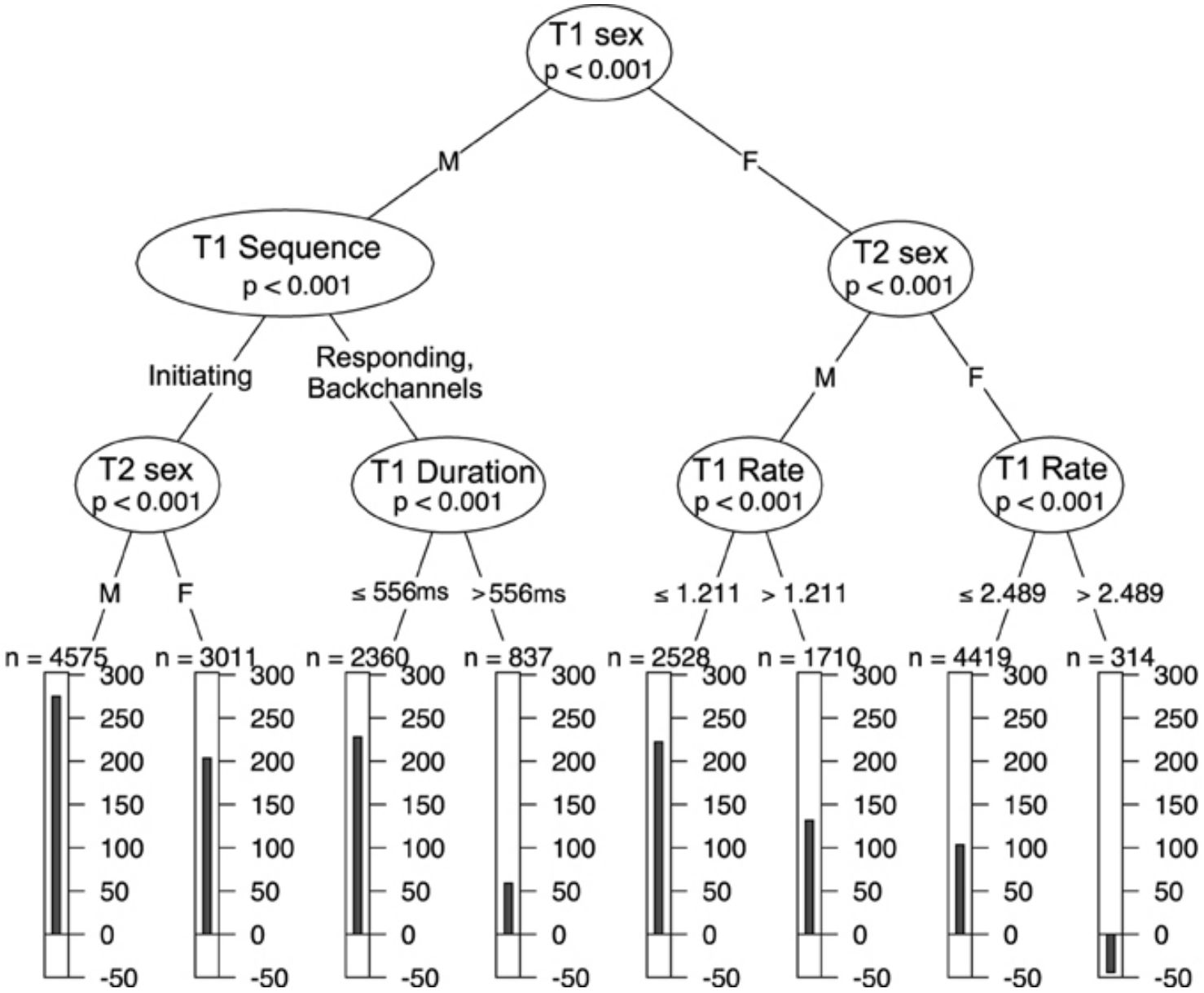
Conversation time
Gender of speaker
Dialect of speaker

Multicolinearity

74% of variable pairs are significantly correlated with $p < 0.05$




T1 initating action
p < 0.001
No
Yes
T2 laughter
p < 0.001
T2 responding action
p < 0.001
No
Yes
No
Yes
T1 laughter
p = 0.001
T1 laughter
p < 0.001
T1 laughter
p = 0.009
No
Yes
No
Yes
No
Yes
n = 16285
n = 650
n = 564
n = 64
n = 1033
n = 1130
n = 28
300
200
100
0
-100
-200
T1 Responding
Invited Laughter (no response)
Volunteered Laughter
Invited Laughter
Dispreferred Response
Adjacency Pair
Adjacency pair with invited laughter


T1 sex
p < 0.001
M
F
T1 Sequence
p < 0.001
T2 sex
p < 0.001
Initiating
Responding, Backchannels
M
F
T2 sex
p < 0.001
T1 Duration
p < 0.001
T1 Rate
p < 0.001
T1 Rate
p < 0.001
M
F
≤ 556ms
>556ms
≤ 1.211
> 1.211
≤ 2.489
> 2.489
n = 4575
n = 3011
n = 2360
n = 837
n = 2528
n = 1710
n = 4419
n = 314
300
250
200
150
100
50
0
-50

Linguistic diversity

132 countries:

Greenberg diversity index

Population size

Population density

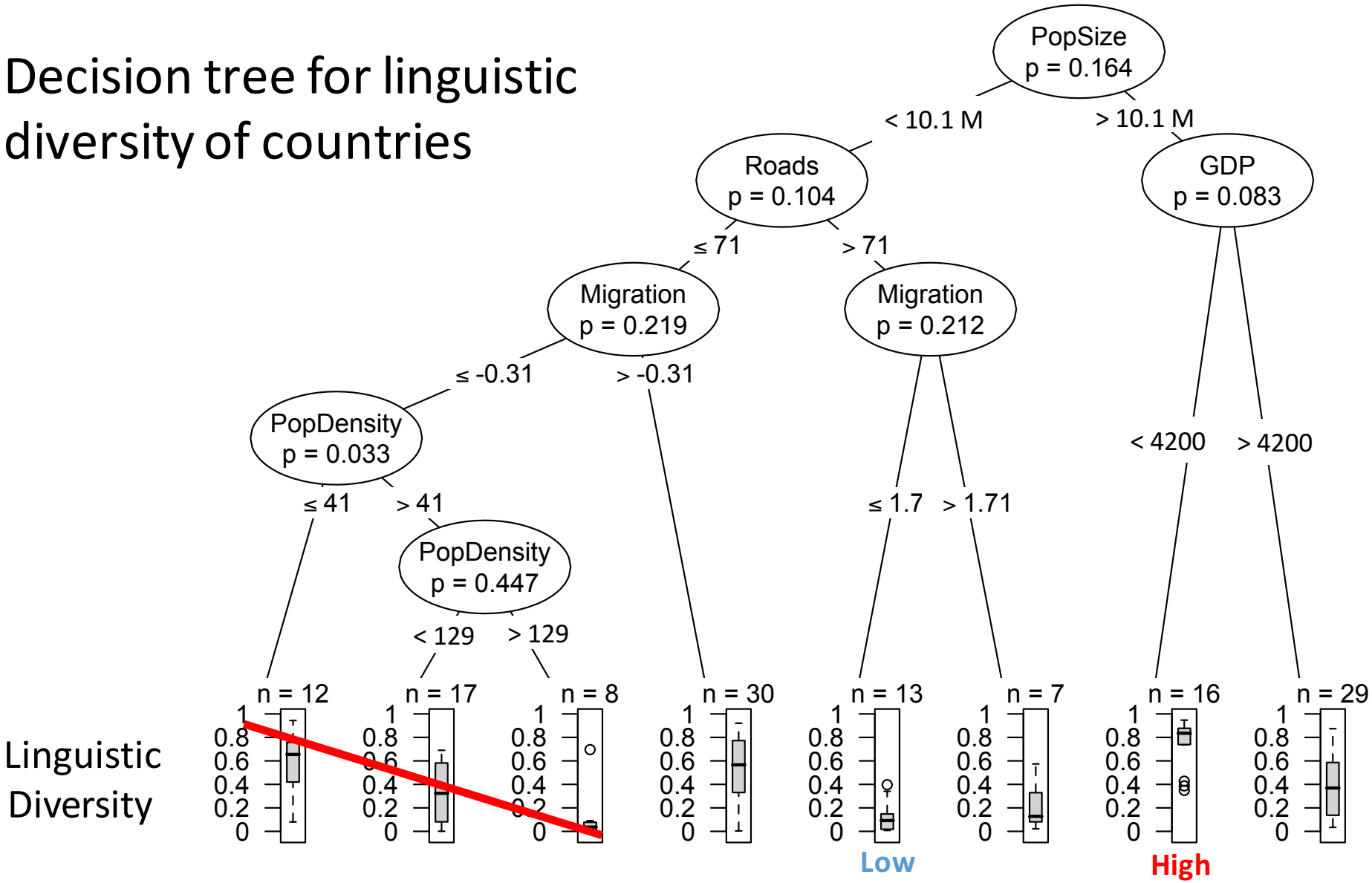
GDP

Migration rates

Km roads

Prediction: population density predicts diversity

Decision tree for linguistic diversity of countries



Building a tree

The strength of association between each predictor variable and FTO is determined by a statistical test of independence.

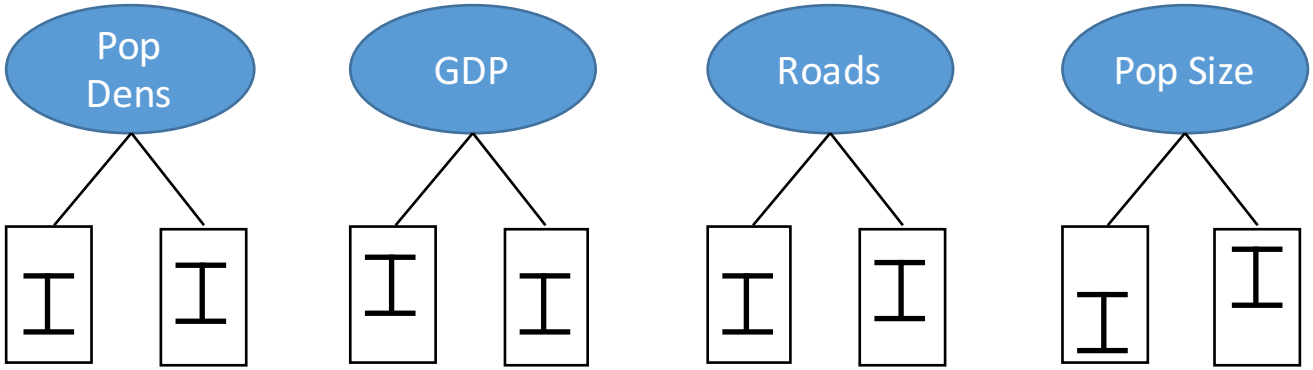
The variable with the strongest association is chosen as the first node in the tree.

The data is divided according to this variable into two sub-sets.

The process repeats recursively with each sub-set until all predictor variables are statistically independent from FTO in each leaf of the tree.

Building a tree

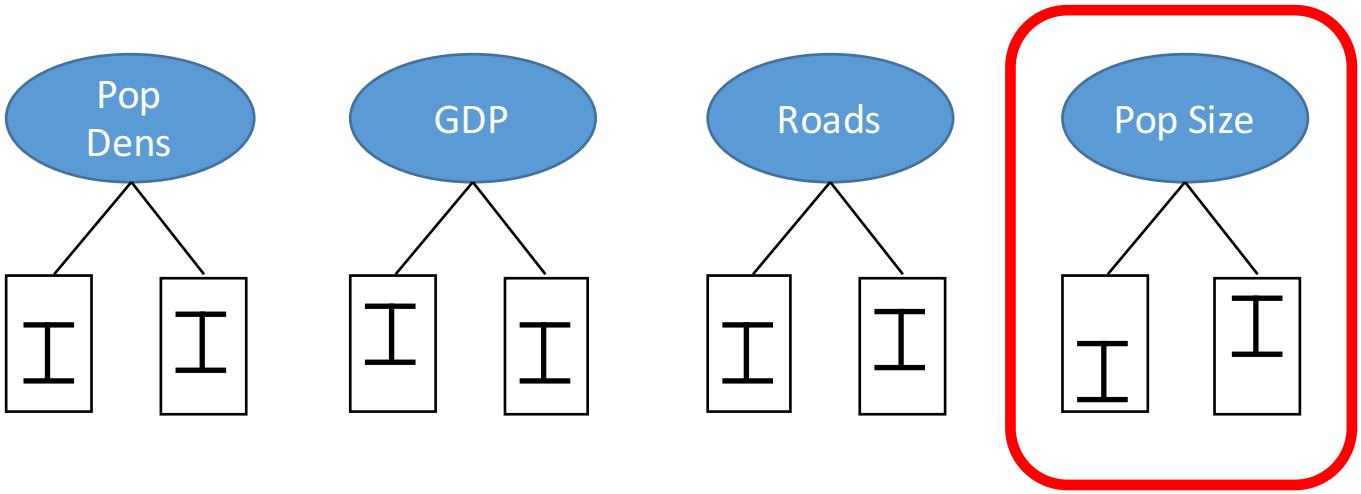
The strength of association between each predictor variable and FTO is determined by a statistical test of independence.



Building a tree

The strength of association between each predictor variable and FTO is determined by a statistical test of independence.

The variable with the strongest association is chosen as the first node in the tree.

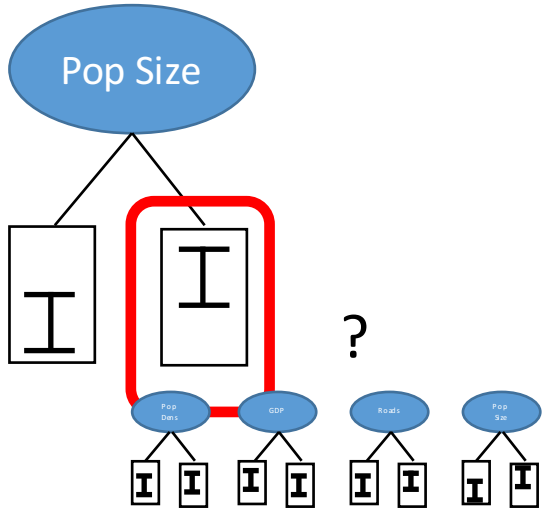


Building a tree

The strength of association between each predictor variable and FTO is determined by a statistical test of independence.

The variable with the strongest association is chosen as the first node in the tree.

The data is divided according to this variable into two sub-sets.
The process repeats recursively with each sub-set until all predictor variables are statistically independent from FTO in each leaf of the tree.



Variable importance

Problem: Single trees are sensitive

- Choice of first variable can change the order of the next ones

Solution:

Generate a ‘forest’ of trees from randomly selected sub-sets of variables.

Aggregate the variable importance:

- How much influence a variable has over the fit of the real data to the predicted data

Importance measures

Standard mean decrease in classification accuracy when a variable is permuted (see Breiman, 2001).

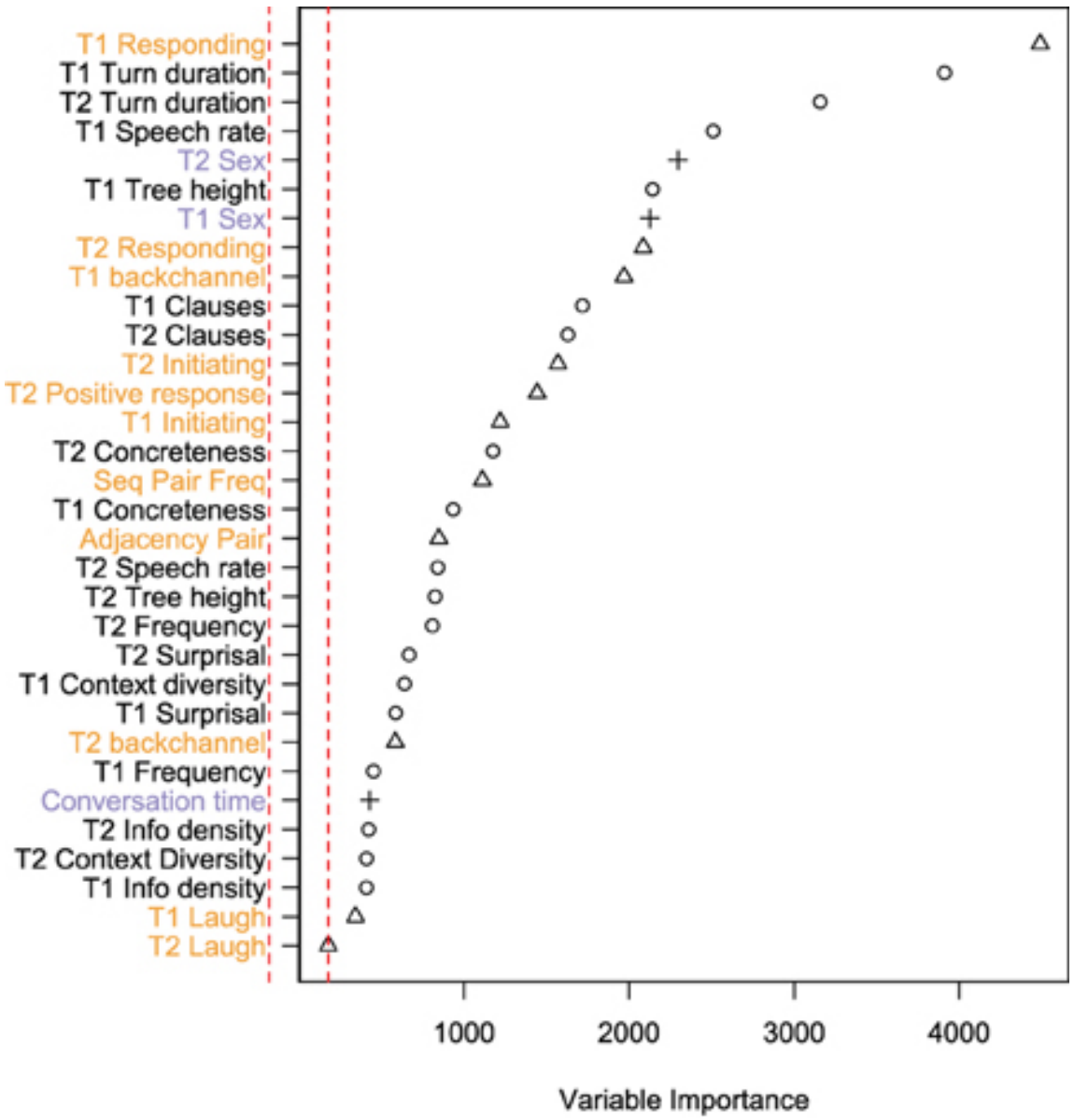
For each tree in the forest, the prediction error (mean squared error) is calculated by comparing the true values of FTO to the values predicted by the tree.

Permute the test variable, re-calculate prediction error.

The difference between the two errors gives a measure of how influential the variable is for prediction of FTO.

The difference in errors are calculated for all trees.

The importance measure is then the mean of these differences normalized by the standard deviation of the differences.


T1 Responding
T1 Turn duration
T2 Turn duration
T1 Speech rate
T2 Sex
T1 Tree height
T1 Sex
T2 Responding
T1 backchannel
T1 Clauses
T2 Clauses
T2 Initiating
T2 Positive response
T1 Initiating
T2 Concreteness
Seq Pair Freq
T1 Concreteness
Adjacency Pair
T2 Speech rate
T2 Tree height
T2 Frequency
T2 Surprisal
T1 Context diversity
T1 Surprisal
T2 backchannel
T1 Frequency
Conversation time
T2 Info density
T2 Context Diversity
T1 Info density
T1 Laugh
T2 Laugh
1000
2000
3000
4000
Variable Importance

Decision trees

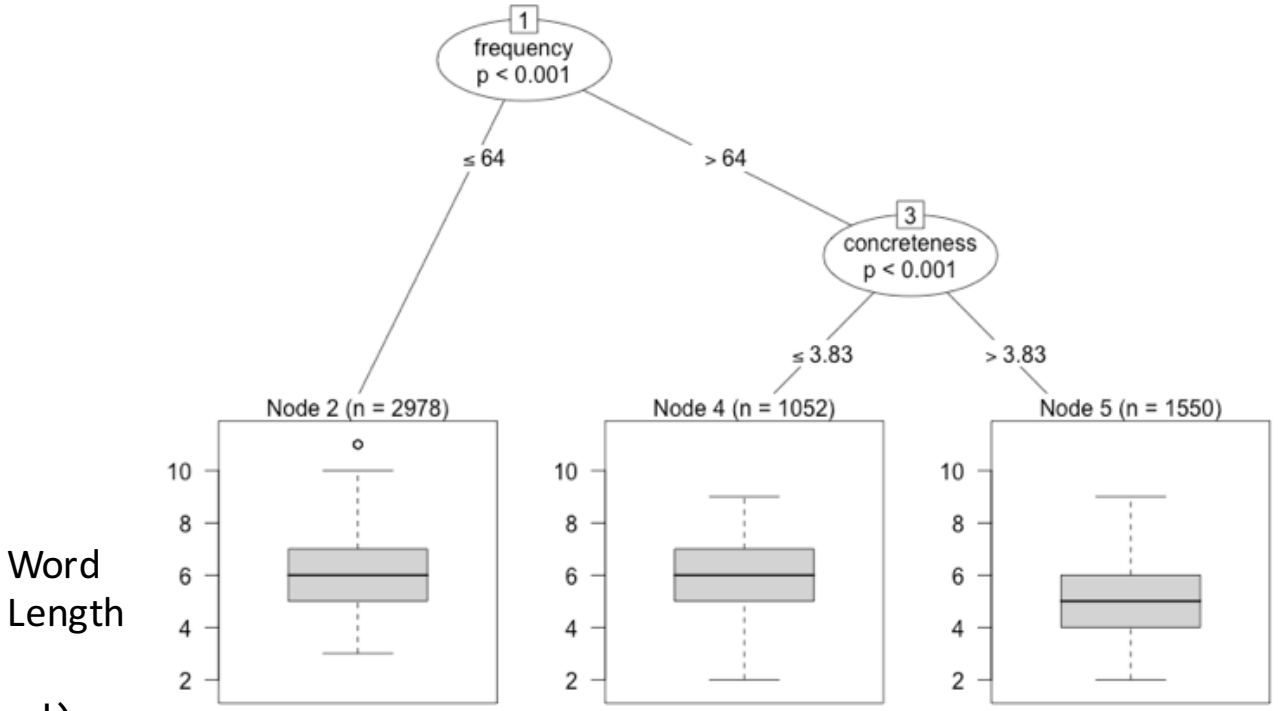
```
# make a single tree:
ct = ctree(dur~.,data=dx)
# make a tree with just 3 branches:
ct2 = ctree(dur~.,data=dx, controls=ctree_control(maxdepth=3))
# plot it (this will make a decision tree, like in our paper)
plot(ct)
#or a bit more tidy:
plot(ct,inner_panel=node_inner(ct,id=F),terminal_panel=node_barplot)


# This can be pretty difficult to read if you have lots of factors, you can look
at the details like this:
ctreex@tree
```

Random Forests

```
# build a random forest:
# see ?cforest_unbiased for list of parameters
data.controls <- cforest_unbiased(ntree=1000, mtry=5)
# run the random forest
d.cforest <- cforest(dur~.,data=dx, controls=data.controls)
# work out ranking of variables
d.varimp <- varimp(d.cforest, conditional=T)
print(sort(d.varimp))
```

Example

Properties of words:
Predict word Length




```
library(party)
tree2 = ctree(length ~ ., data=d)
plot(tree2)
```

analyseWords_DecisionTrees.R

https://github.com/shh-dlce/qmss-2017/blob/master/FindingPatterns/analysis/analyseWords_DecisionTrees.R

Example

Predict intensity of agriculture by:

Trance, High Gods, Games

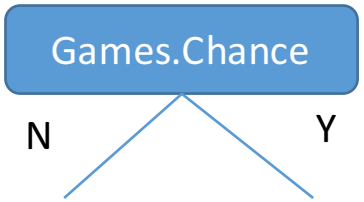
Control for structure (e.g. language family) using random effects:

Use REEMtree package

Other

Trance,
Trance (due to possession only),
Trance and Possession

Controlling for language family

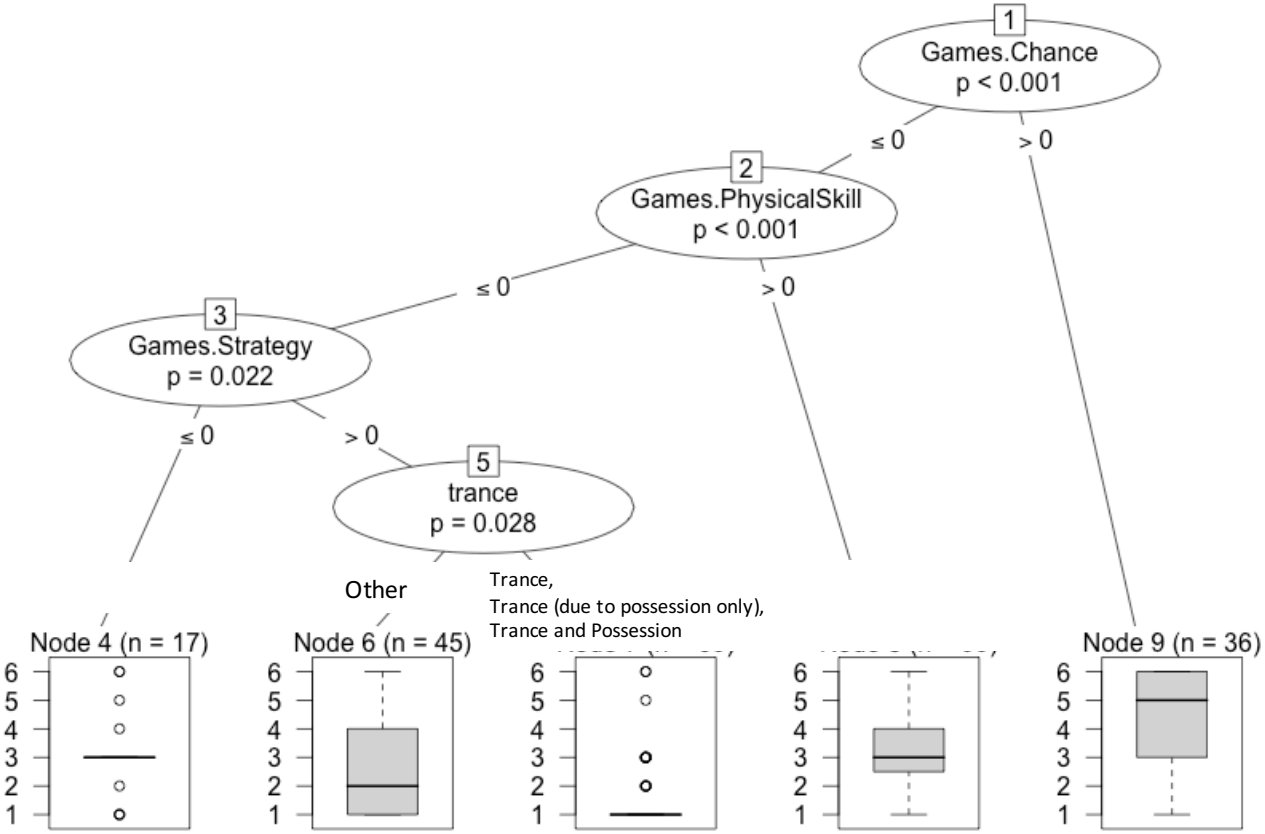


Argiculture Intensity = 2.3 3.9

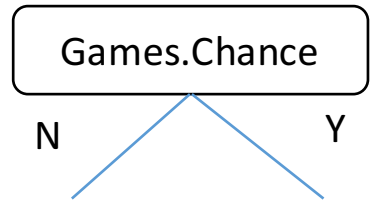
```
reem.tree = REEMtree(Agriculture ~ .,  data=d,
        random = ~1|Language.family)
```

https://github.com/shh-dlce/qmss-2017/blob/master/FindingPatterns/analysis/analsyseTrance_RandomEffects.R

Without controlling for language family



Controlling for language family



Agriculture Intensity = 2.3 3.9

Producing unbiased hypotheses

What time does it start?

Four o'clock

English	Latvian
haw	ka:
haw mɛni	tsik
haw mətʃ	tsik
wət	kas
wɛn	kad
wɛr	kur
wɪtʃ	kurʃ
hu	kas
waj	ˈka:pe:ts

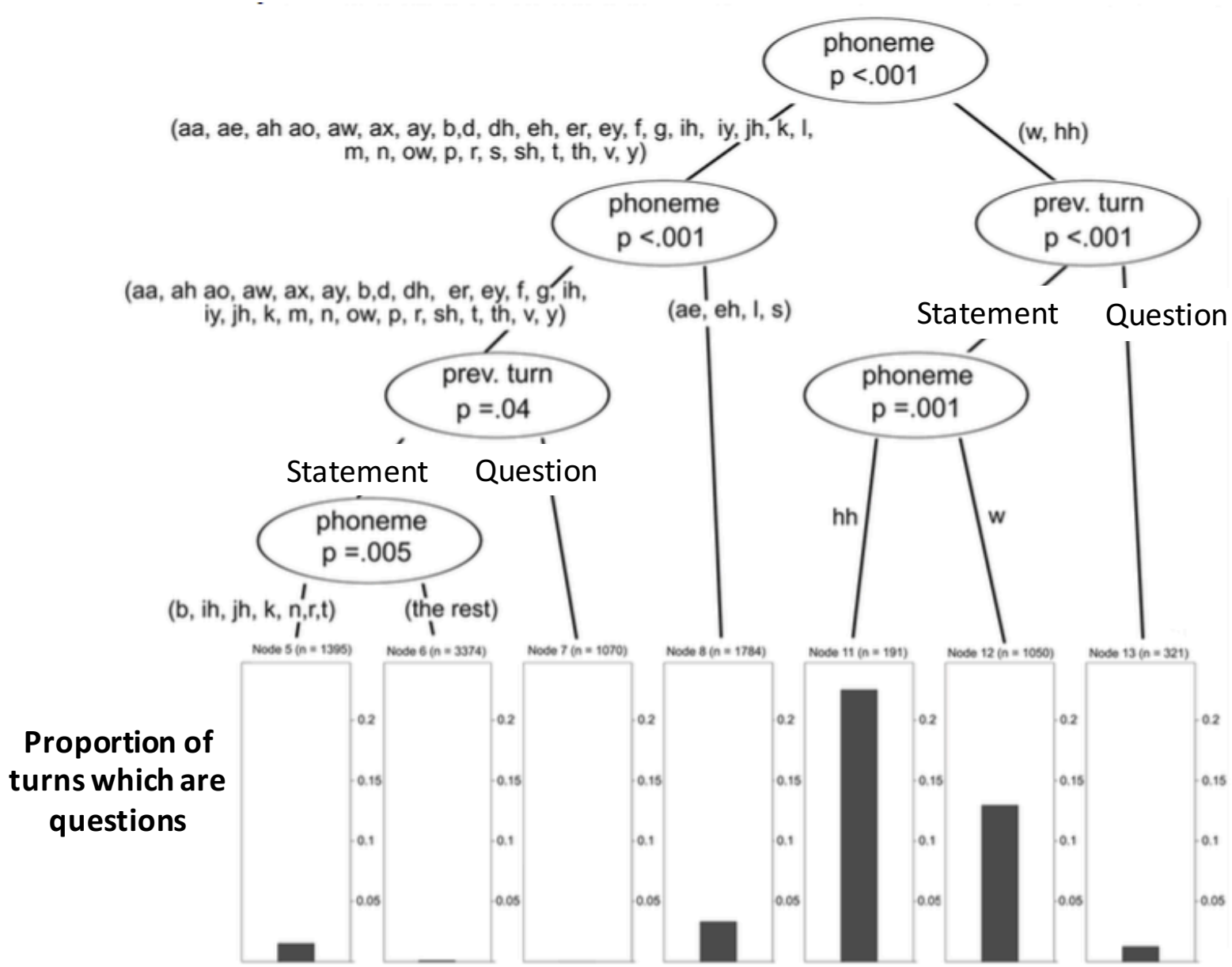
/w/ and /h/ are cues that a question is coming

Data: 10,000 turn pairs from English telephone conversations

- Is the turn a question?
- First phoneme of turn
- Type of previous turn

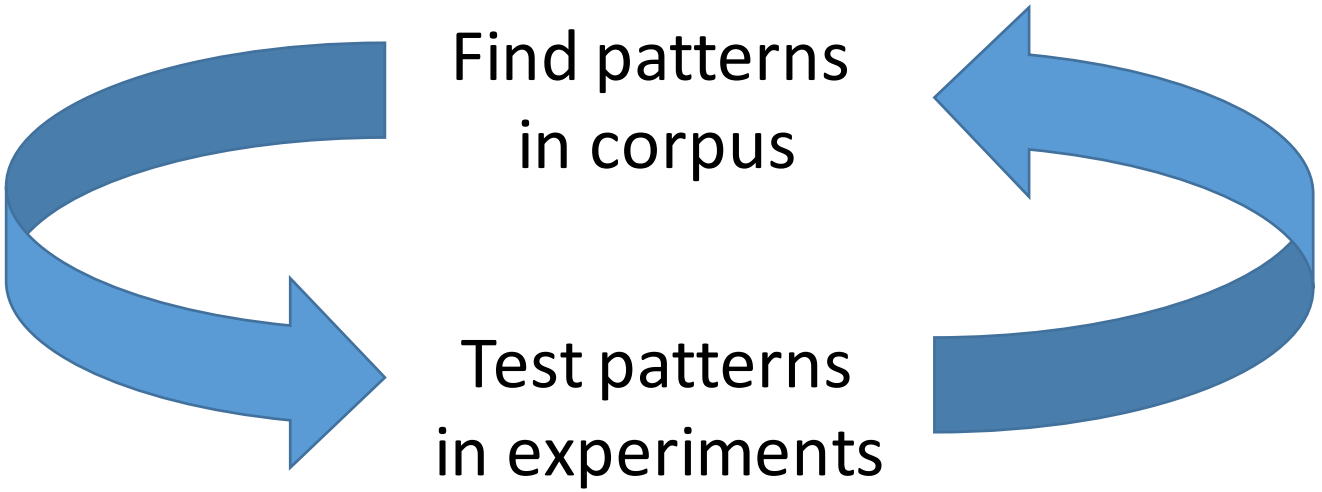
Prediction: A decision tree will divide turns starting with /w,h/ from turns starting with other phonemes

Slonimska & Roberts (in press)



Proportion of turns which are questions

Virtuous cycle

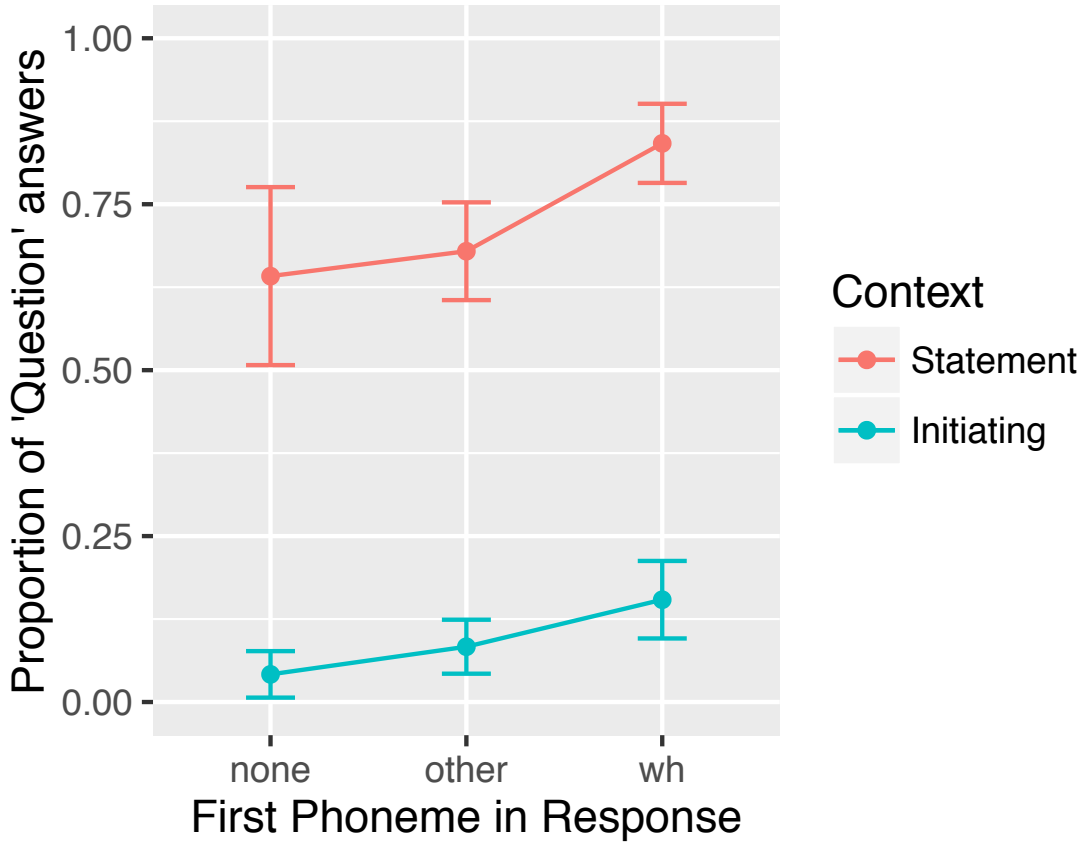


Forced choice task

I went on holiday.

W-

Will the turn be a question?



Slonimska & Roberts (2017)

Prediction

Make a prediction based on new data:

```
predictions  = predict(d.cforest,newDatainDataFrame)
```

If you run this without new data, you'll just try to predict the values that the model was trained on. This can be insightful, too.

Compare predictions to actual values (a measure of model fit):

```
cor.test(predictions,  dx$dur)
```

You can also run a forest with a random 90% of the data, then predict the other 10% and see how well the model fits

More details

Reading:

Roberts, S. G., Torreira, F., and Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: a corpus study. Frontiers in Psychology, 6.

Tagliamonte, S. A., and Baayen, R. H. (2012). Models, forests, and trees of york english: was/were variation as a case study for statistical practice. Lang. Variat. Change 24, 135–178. doi: 10.1017/S0954394512000129

Decision trees with random effects:

REEMtree package in R

Causal Graphs

Description

Infer the most likely graph of causal relations between variables from observational data

Advantages

Handles large numbers of variables

Can handle complex relationships

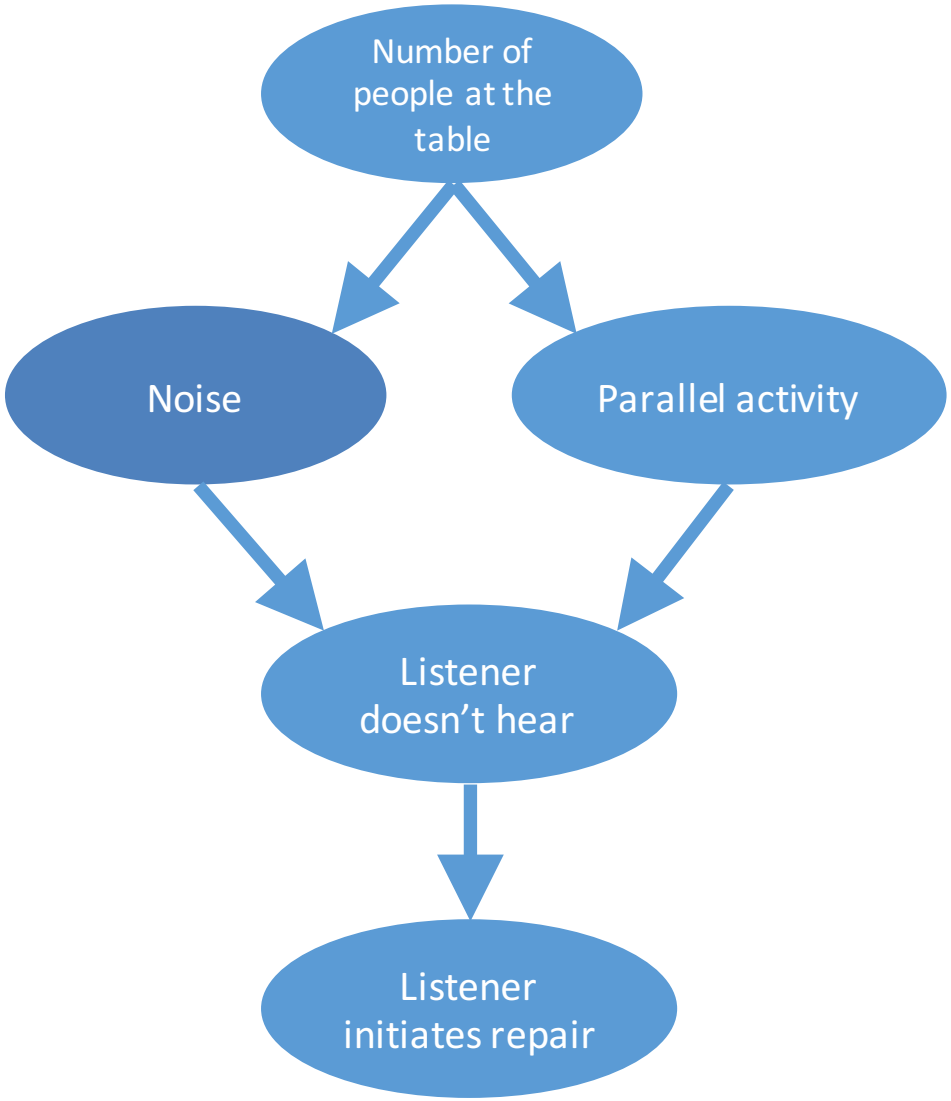
Easy to interpret

Assumptions

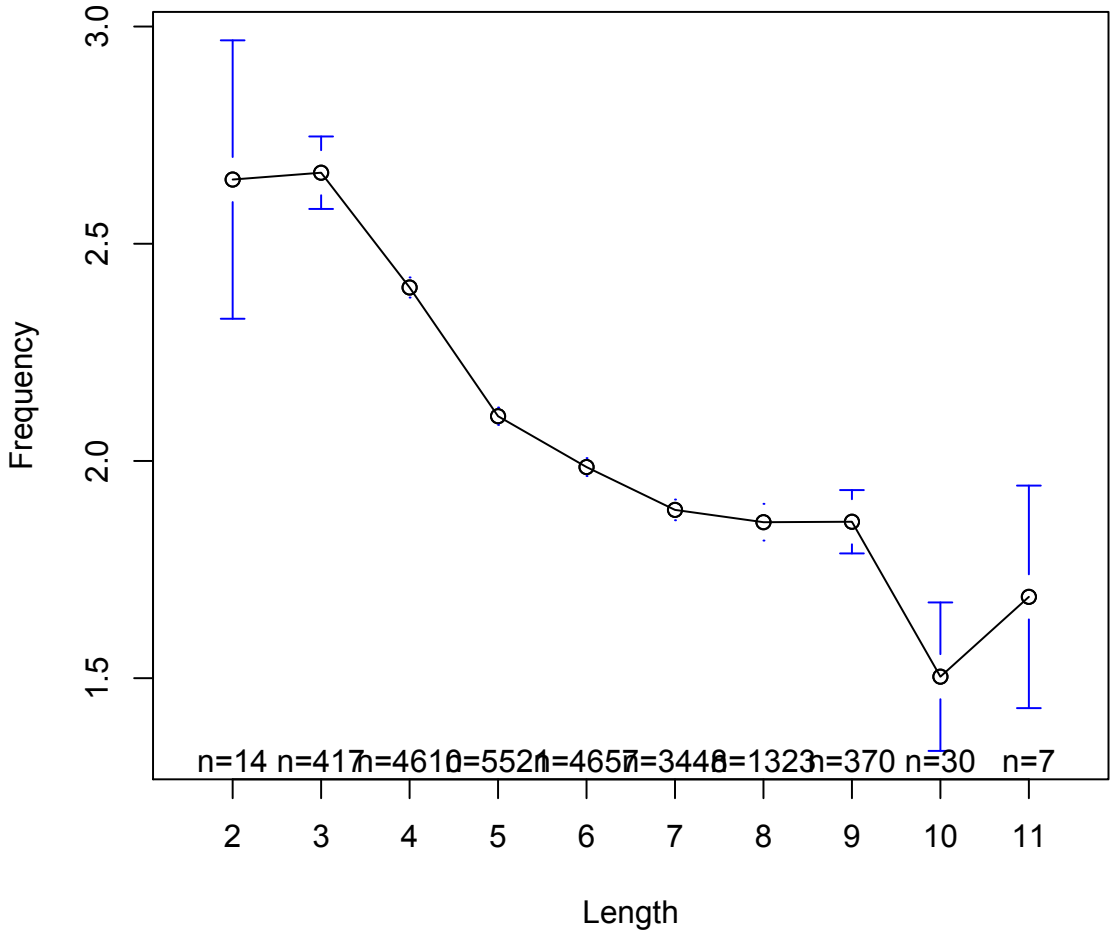
We have measured all relevant variables (closed world)

The real graph is directed and acyclic

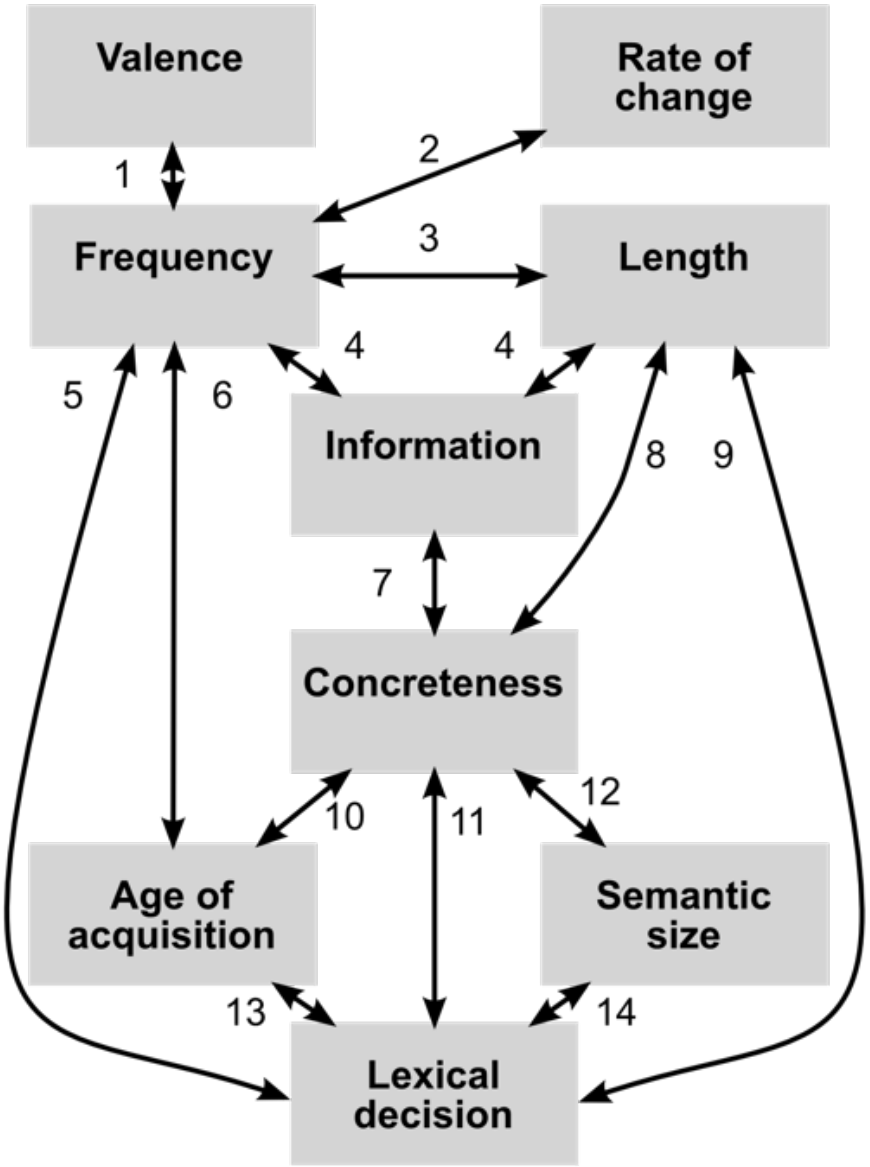
A conversation in a crowded restaurant



Frequency and Length



Properties of words



1. Boucher & Osgood (1969)
2. Pagel et al. (2007)
3. Zipf (1936)
4. Piantadosi et al. (2011)
5. Balota et al., 2004
6. Kuperman et al., 2013
7,8. Piantadosi et al. (2011b)
9. Hudson & Bergman, 1985
10. Reilly & Jacobs, 2007
11,12. Yao et al. (2013)
13. Walker & Hulme (1999)
14. Sereno et al. (2009)

Inferring causal graphs

PC algorithm (Sprites et al., 2000; Kalisch et al., 2012)

Start with fully unconnected graph

For each pair of variables:

Try to find evidence that the variables are independent:

no correlation,

or correlation is explained by a set of other variables

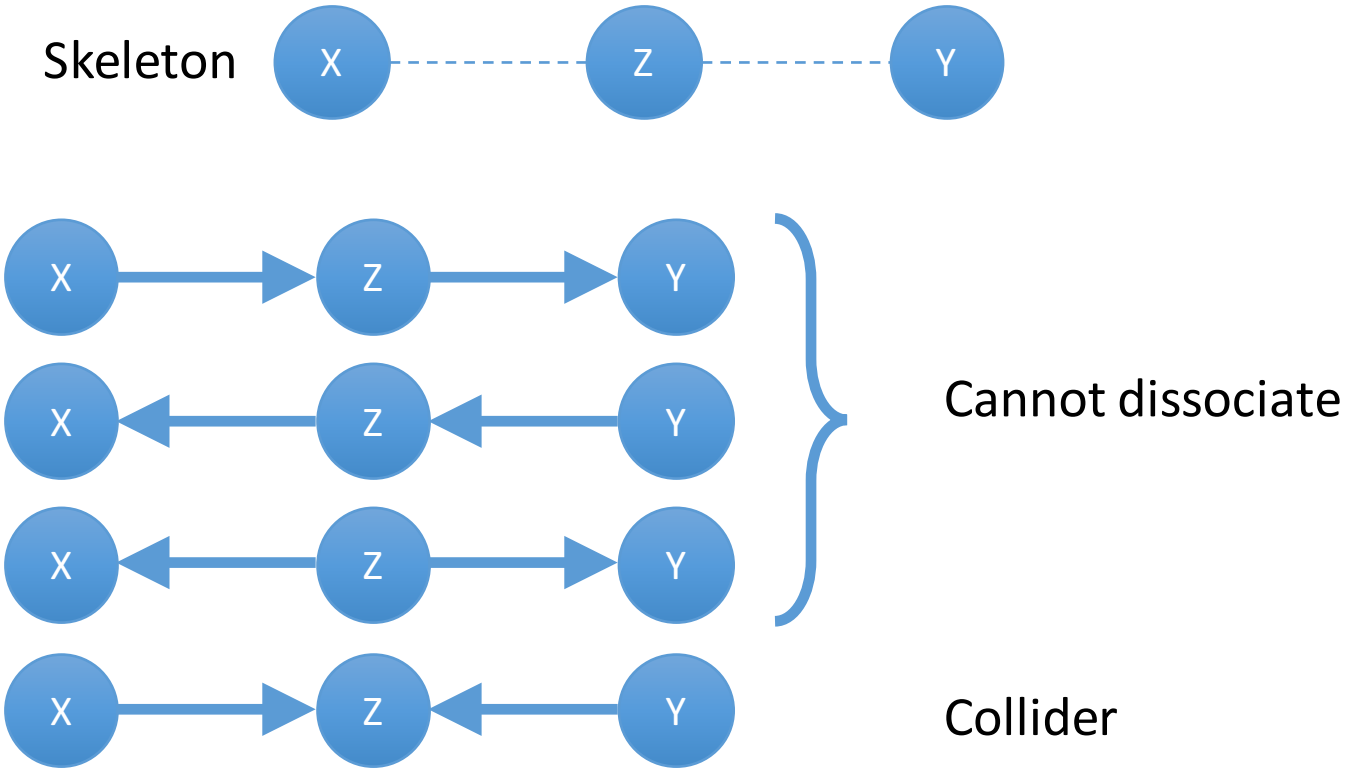
Any statistical test can be used (e.g. conditional independence)

If variables are not independent, add an edge.

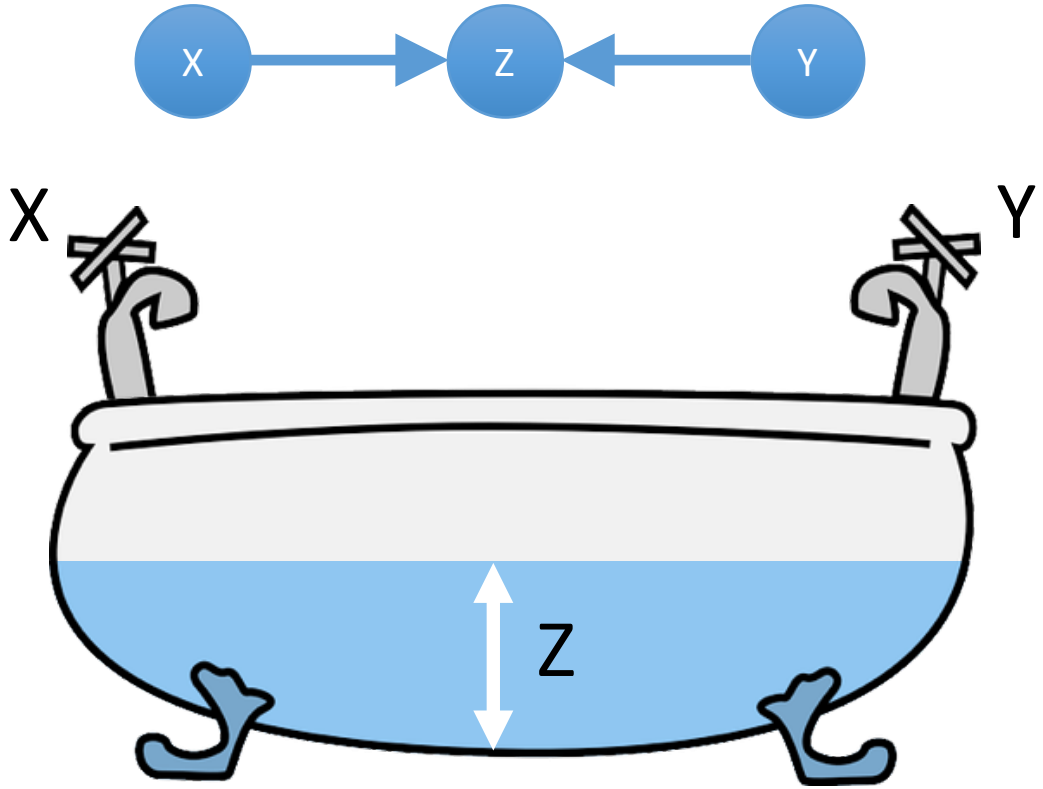
The PC algorithm is an efficient way of performing only the tests which need to be done.

Results in a ‘skeleton’ graph

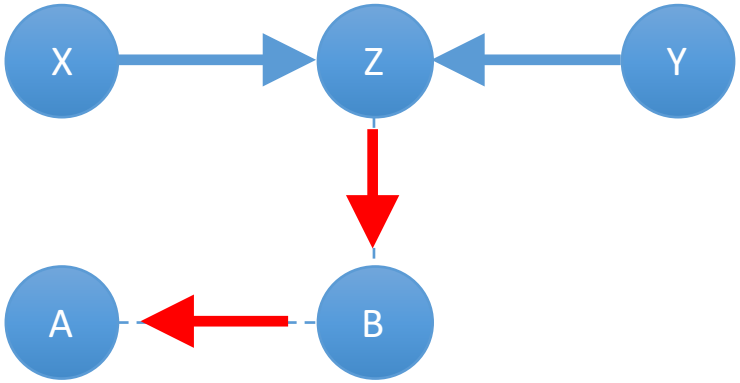
Orienting the edges



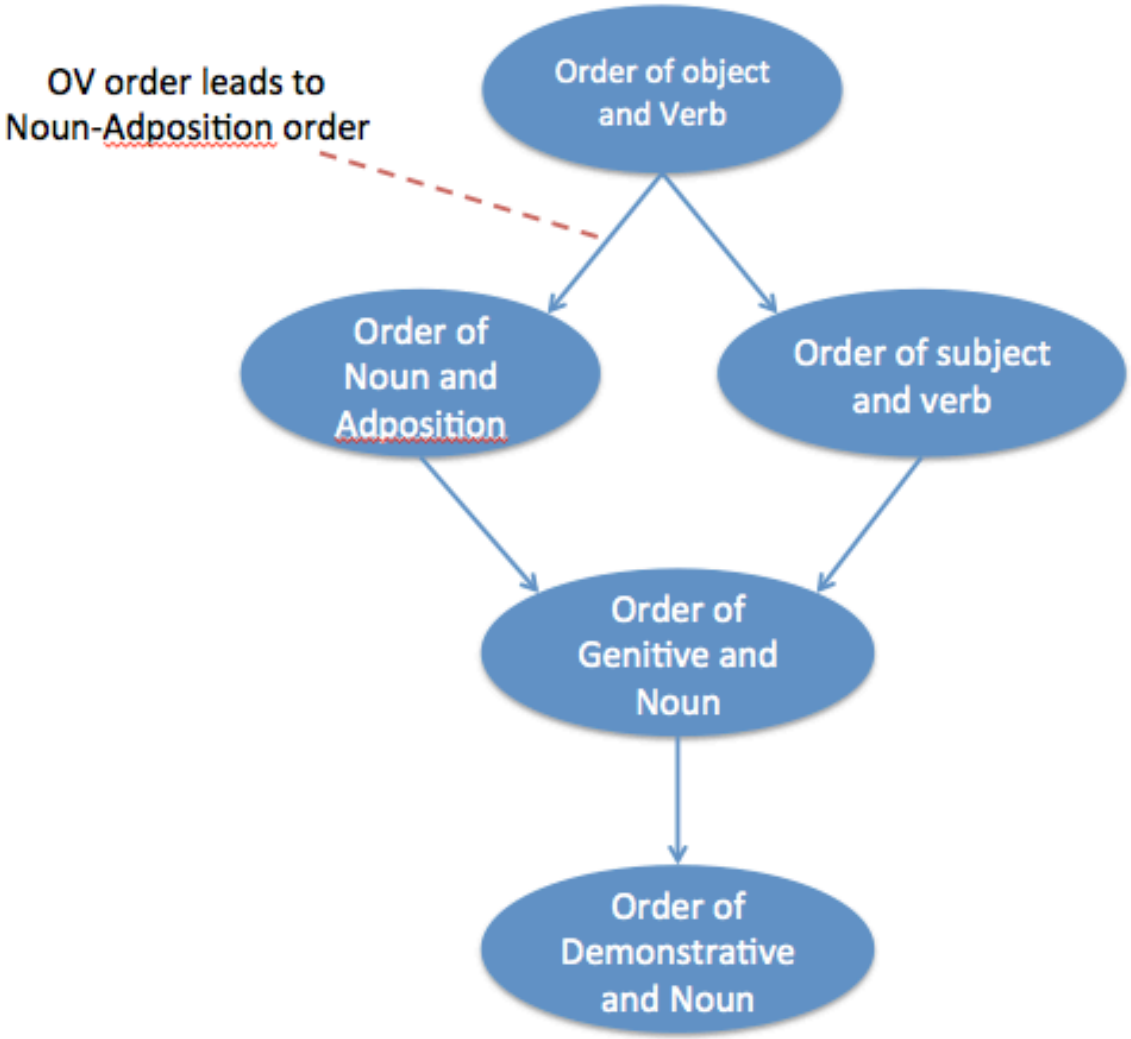
Colliders



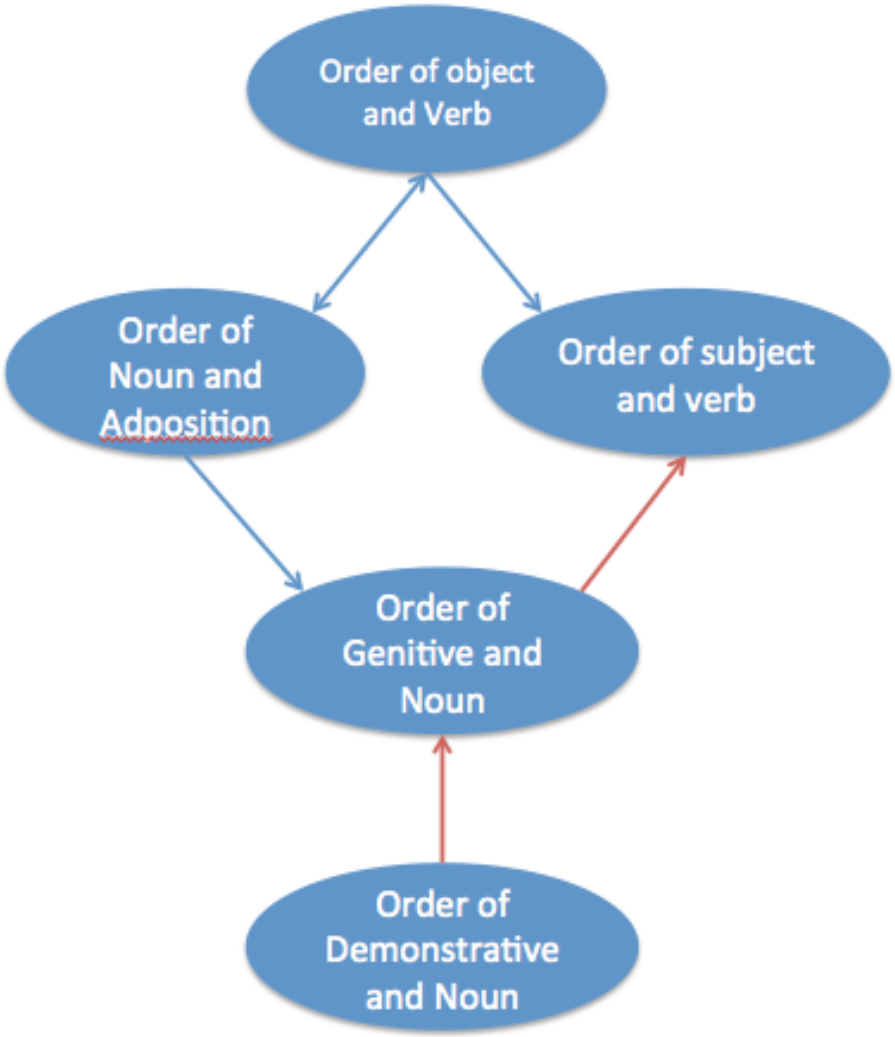
Orienting the edges



Greenberg



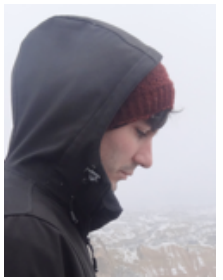
Estimated from data



Case study: Properties of words

- Length (CELEX)
- Frequency (form and lemma) (CELEX)
- Part of Speech (CELEX)
- Age of Acquisition (Kuperman et al. 2013)
- Valence, Arousal, Dominance (Warriner et al. 2013)
- Concreteness (Brysbaert et al., in press)
- Lexical decision time (Keuleers et al., 2012)
- Orthographic neighbourhood (OLD20)
- Phonological neighbourhood (IPhOD)
- Contextual diversity (Subtlex)
- Familiarity (Wordnet Polysemy)
- Surprisal (information) (Piantadosi et al.)

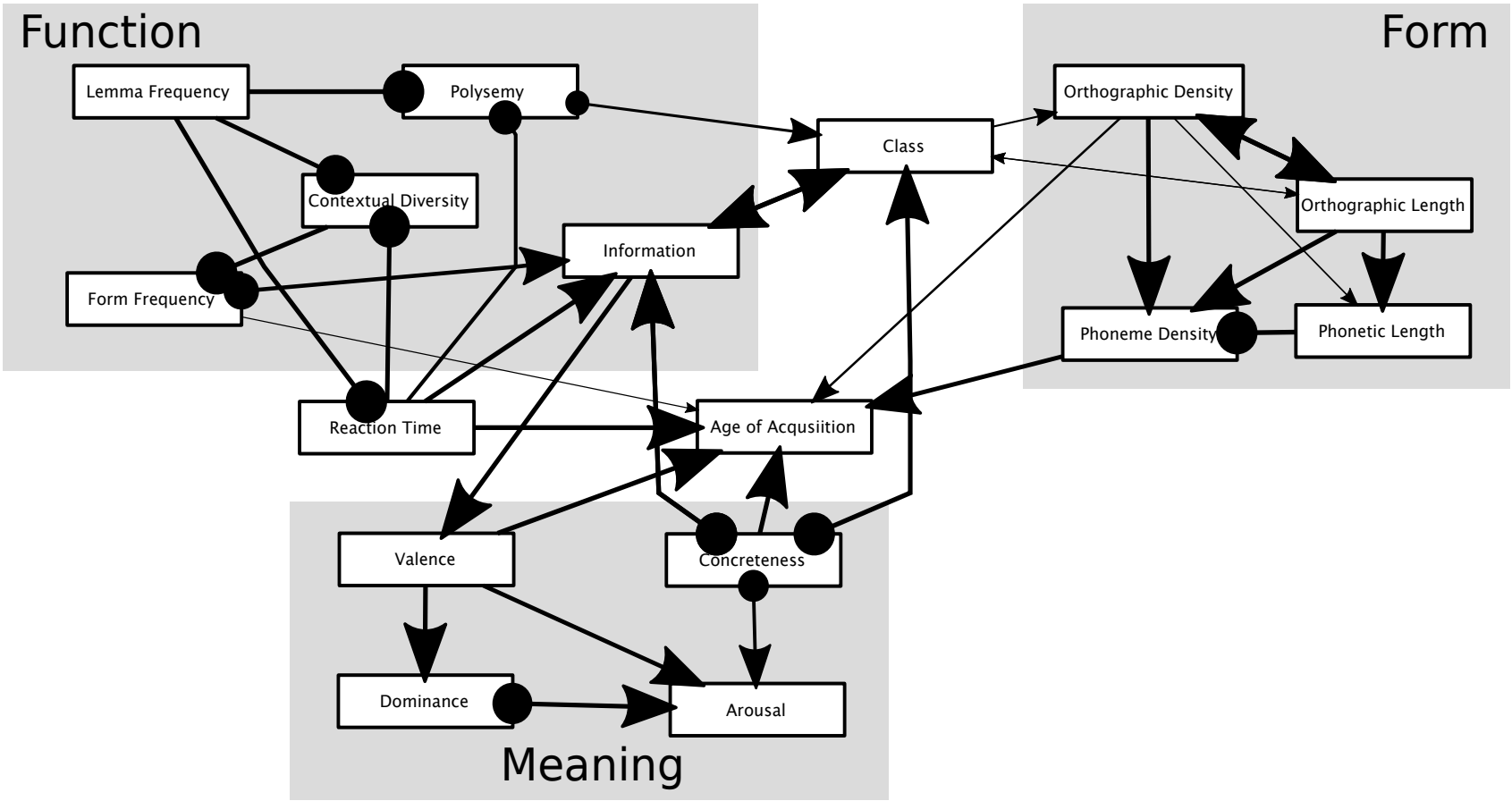
Marloes Maathuis



Damián Blasi

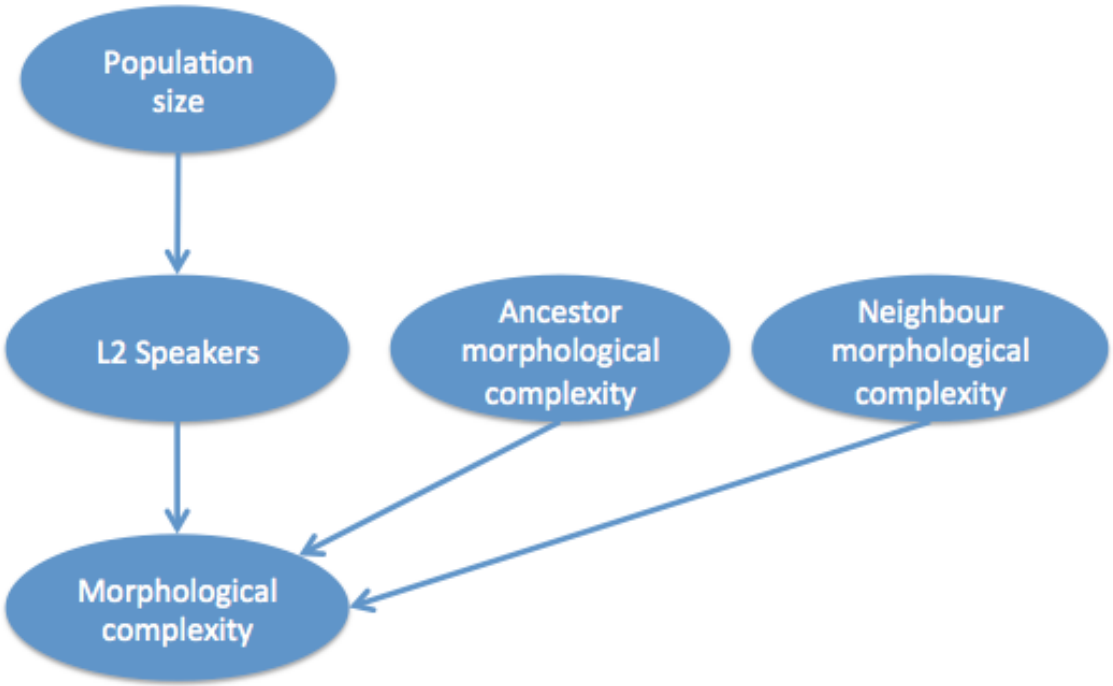
Emmanuel Keuleers

10,000 English words

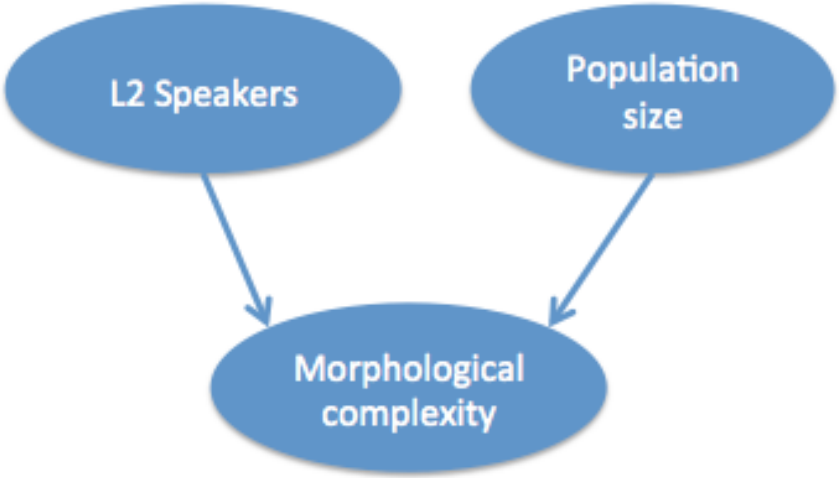

Function
Form
Meaning
Lemma Frequency
Polysemy
Class
Orthographic Density
Contextual Diversity
Orthographic Length
Information
Form Frequency
Phoneme Density
Phonetic Length
Reaction Time
Age of Acqusiition
Valence
Concreteness
Dominance
Arousal

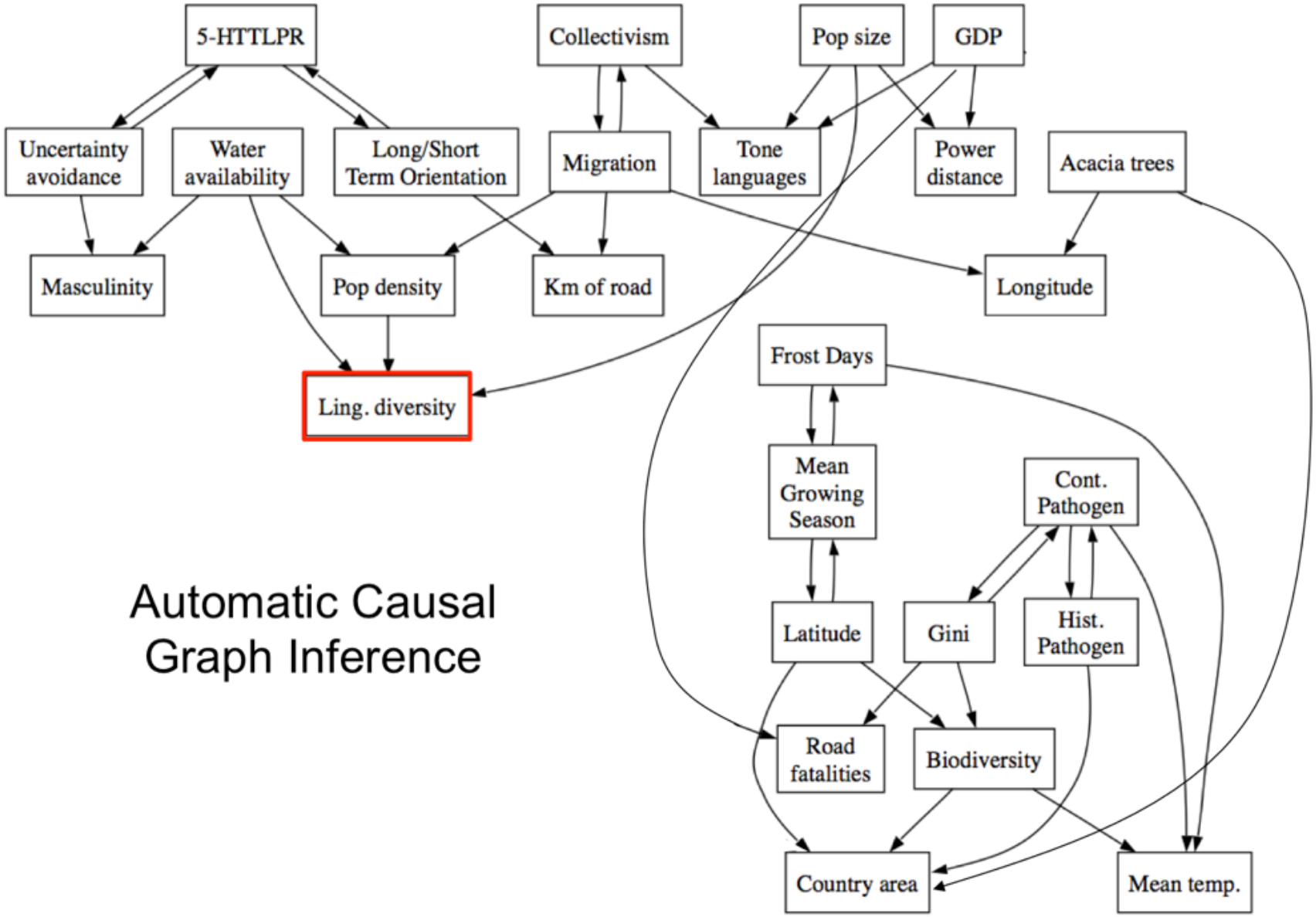
Code for Causal Graphs

Bentz & Winter (2014)



Estimated from data





Automatic Causal Graph Inference