Adventures in quantitative analyses

Seán Roberts

Overview

Today:

Tools for sharing data and analyses

Quantitative tools for testing hypotheses

Why Bother?

Quantitative methods let you ask more questions and communicate them more effectively to more people

How to think quantitatively: Hypotheses, data, tests

Galton's problem, and some solutions

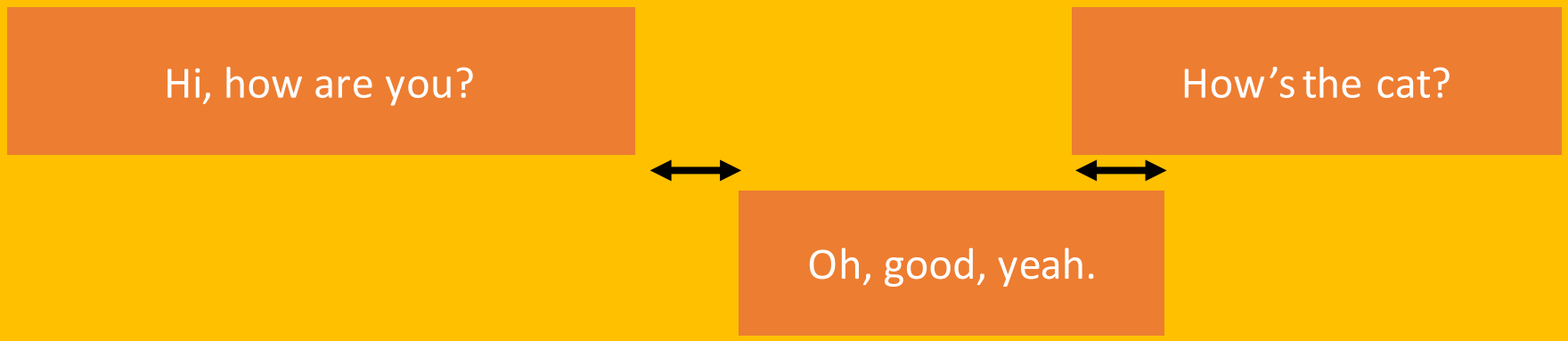
How to argue with data

Why bother?

Stress reduction

[repository demo]

Why Bother? Speed and flexibility



Variables:

Gap length

Concreteness (Brysbaert)

Frequency (Subtlex)

Speech rate (Praat)

Surprisal (Piantadosi)

Information density

Turn duration (pympi)

Number of Clauses (Switchboard)

Tree height

Frequency of speech act pair

Duration

Conversation time

Gender of speaker

Dialect of speaker

Orth/Phonetic transcriptions

Why Bother?

A typical response to quantitative methods

Roger Blench

Phylogenetic methods are:
Not reproducible
Not transparent
Tell us nothing new

Blench (2015) *New mathematical methods in linguistics constitute the greatest intellectual fraud in the discipline since Chomsky*

The scientific method

A way of demonstrating the validity of a theory about how something works,

which does not depend on our personal belief,

and we can **show to other people**.

Biases

Base rate neglect

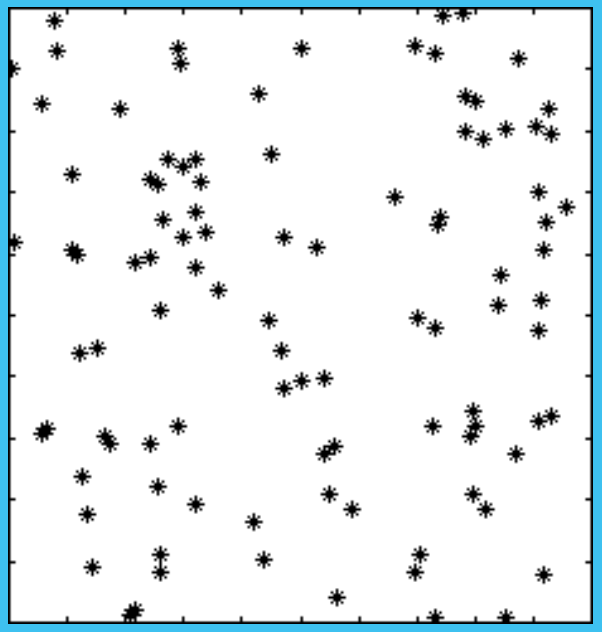
Clustering illusion

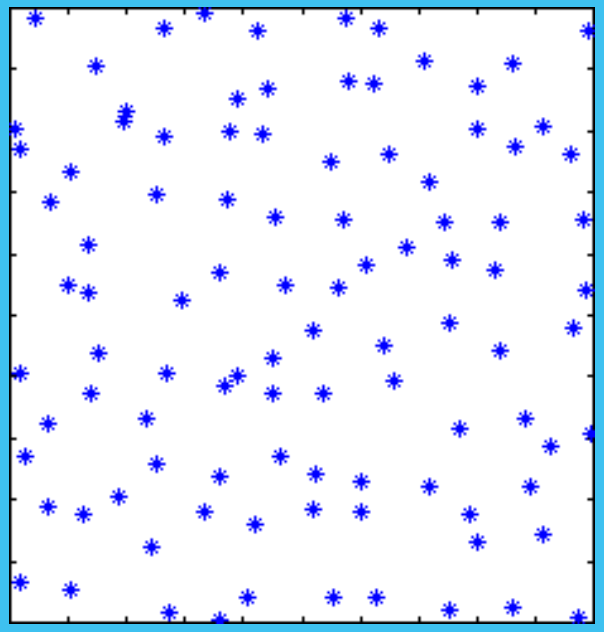
Galton's problem

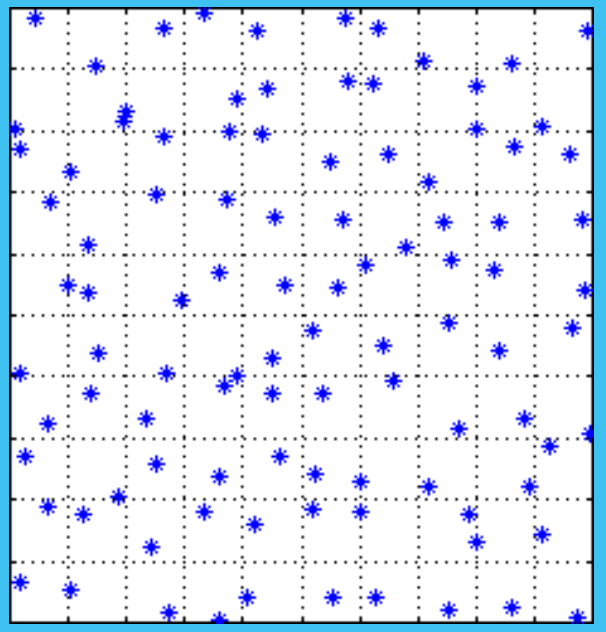
Confirmation bias

Hindsight bias

Illusion of validity

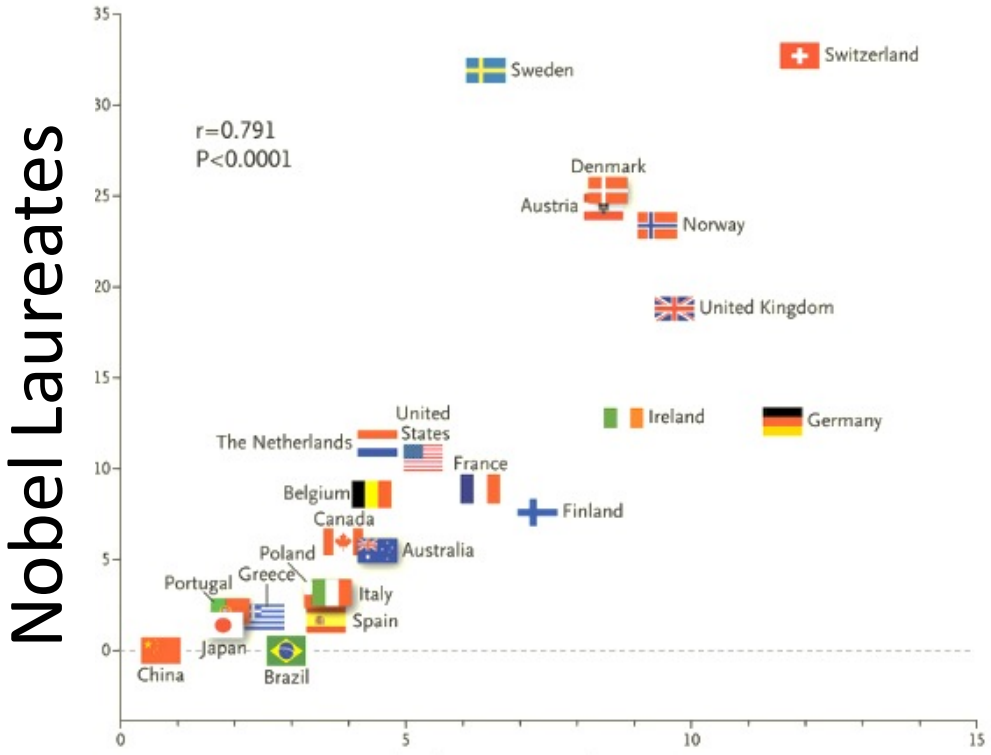




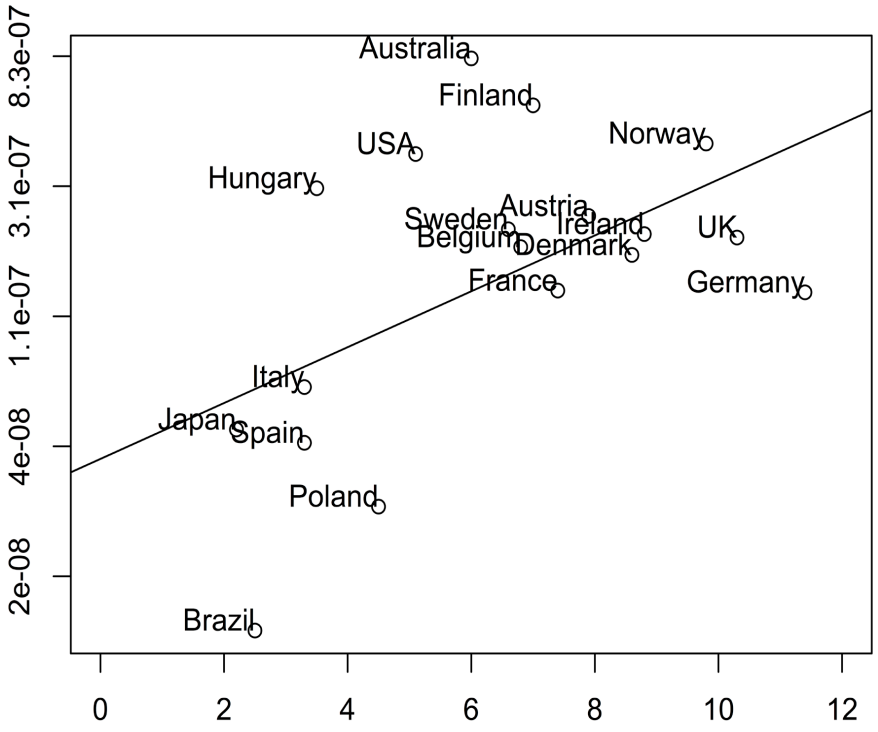


Chocolate

Messerli (2012)



Roberts & Winters (2013)



Scientific method

Scientific method aims to produce results that are:

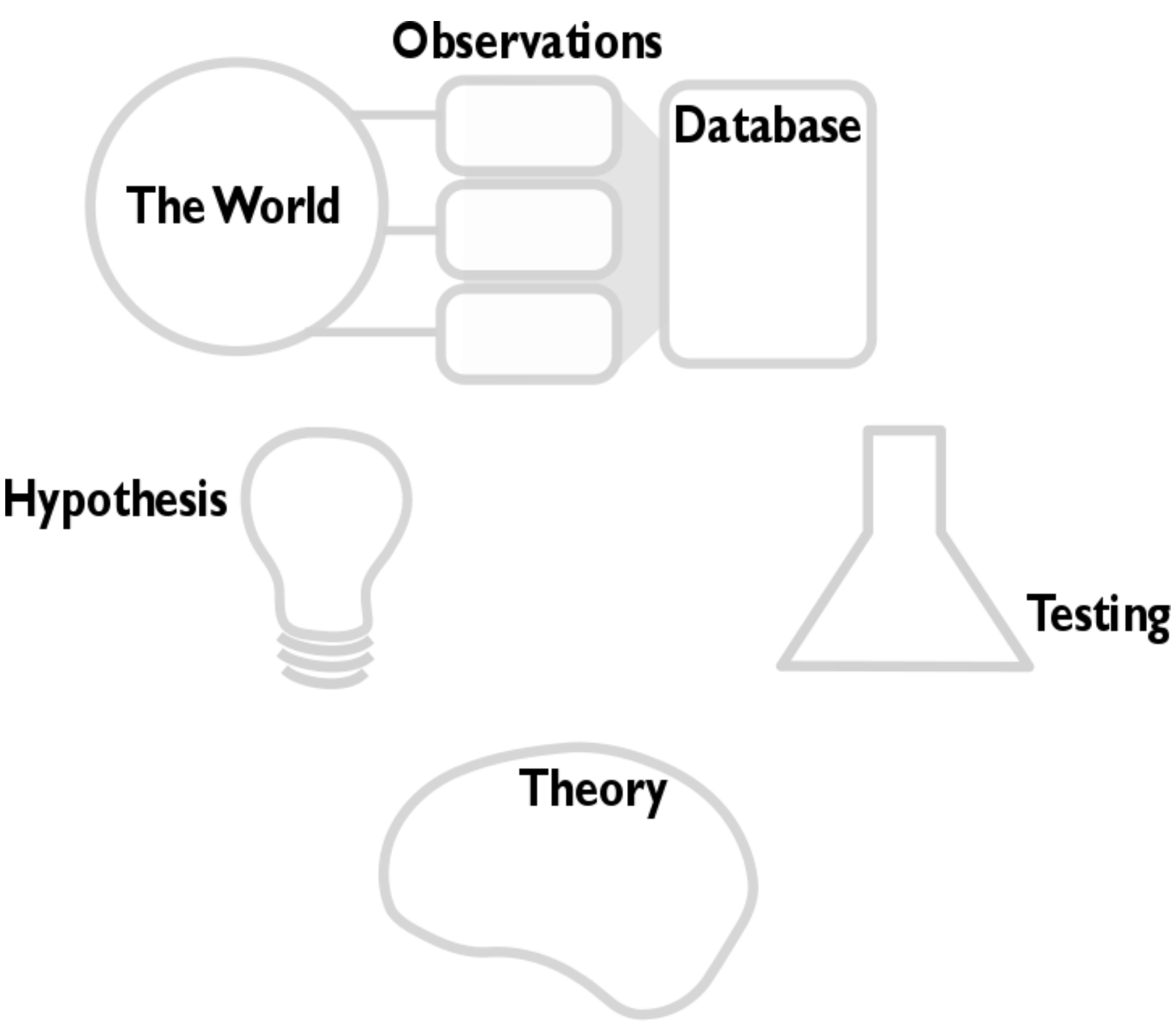
- Unbiased
- Transparent
- Reproducible
- Informative for theories

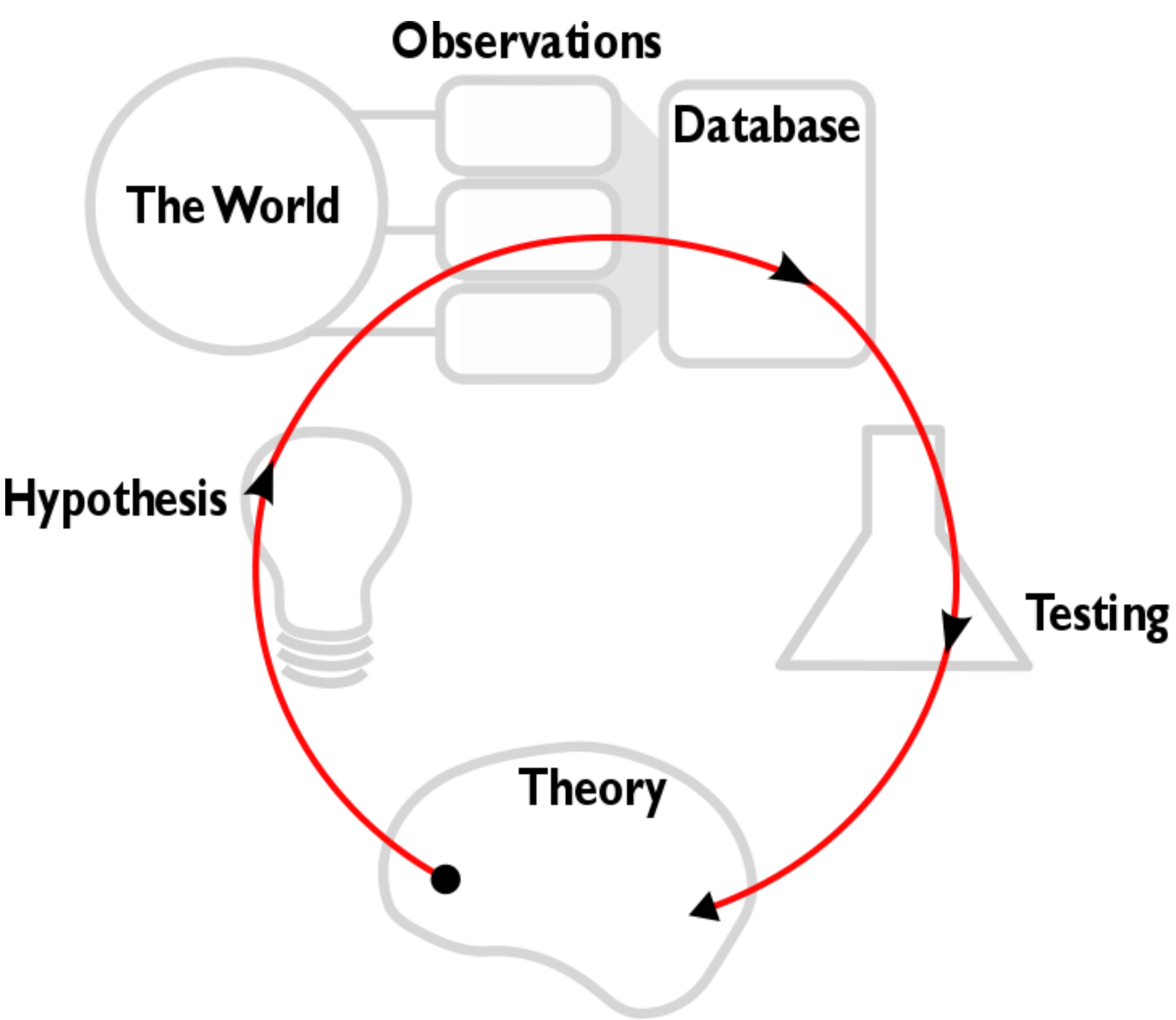
Leads to:

More interesting answers

More flexible analyses and discussion

Less stress


Observations
The World
Database
Hypothesis
Testing
Theory

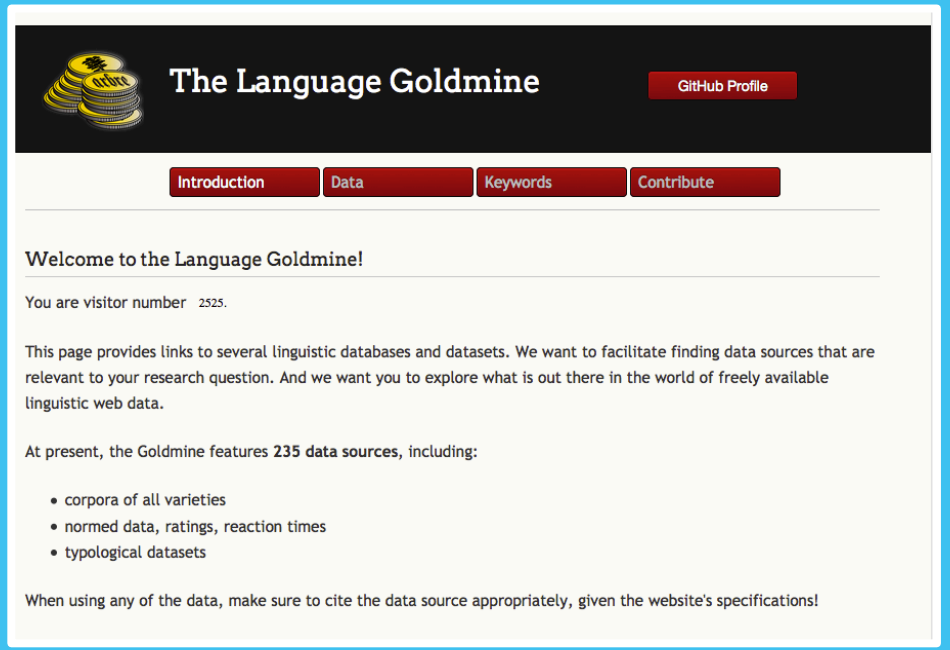

Observations
Database
The World
Hypothesis
Testing
Theory

Sources of data

The internet!

Basic grammar:	World Atlas of language structures	http://wals.info/
	GramBank	(in prep)
Language families:	Glottolog	http://glottolog.org/
Basic demographics:	Ethnologue	http://www.ethnologue.com/
Lexicons:	Austronesian Basic Vocabulary Database etc.	
Corpora:	SUBTLEX, Google Books	
Behavioural data:	Lexicon projects	http://crr.ugent.be/

The language goldmine:
A list of databases
http://languagegoldmine.com/



Identifier codes

Languages

Glottolog code

iso- codes

Glottocode	Name	Top-level family	ISO-639-3
Search	Search	Search	Search
aari1239	Aari	South Omotic	aiw
aasa1238	Aasax	Afro-Asiatic	aas
abad1241	Abadi	Austronesian	kbt
abag1245	Abaga	Nuclear Trans New Guinea	abg
abai1240	Abai Sungai	Austronesian	abf
abai1241	Abai Tubu-Abai Sembuak	Austronesian	
aban1242	Abanyom	Atlantic-Congo	abm
abar1238	Abar	Atlantic-Congo	mij

Meanings

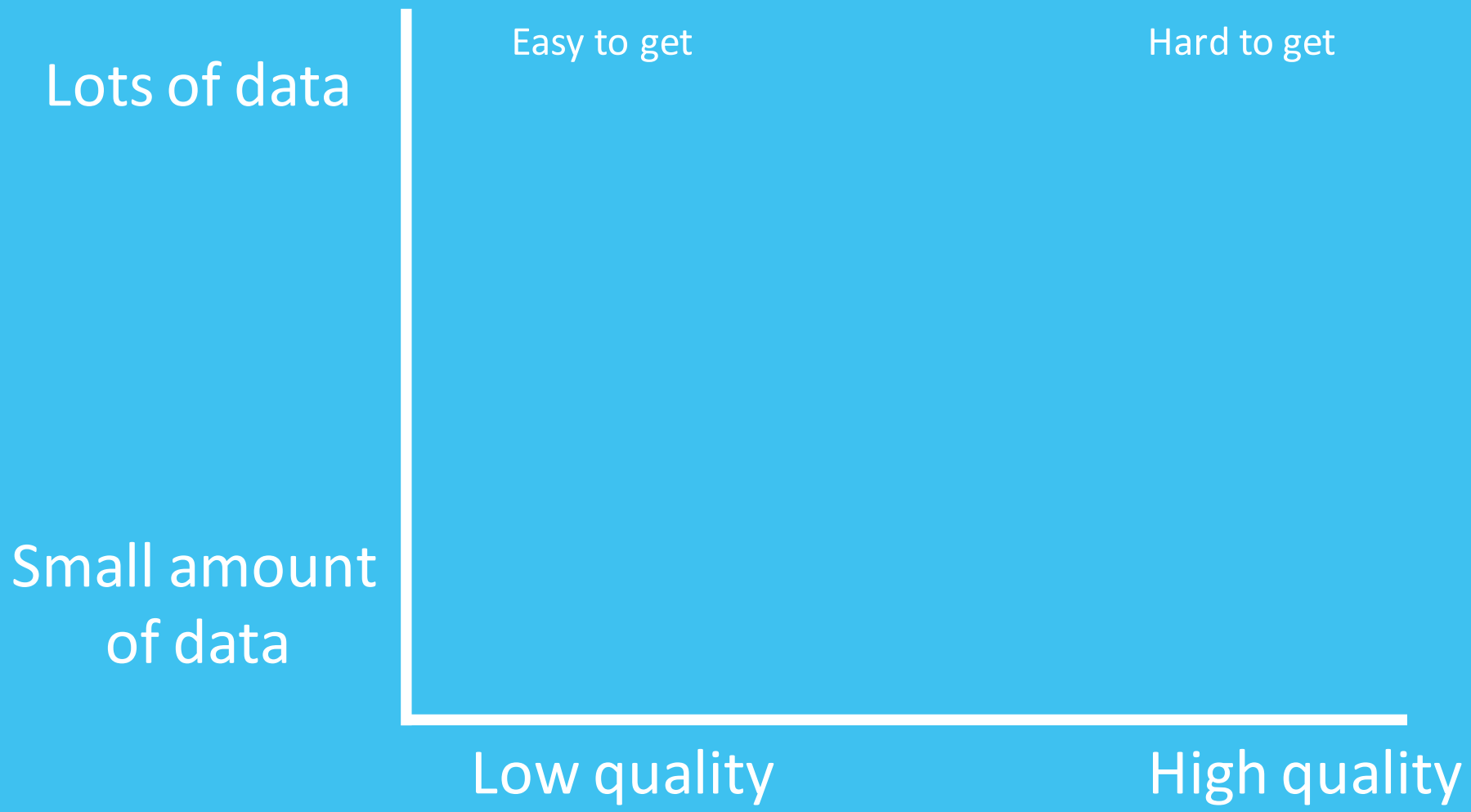
Loanword Typology code (LWT, based on IDS list)

http://wold.clld.org/

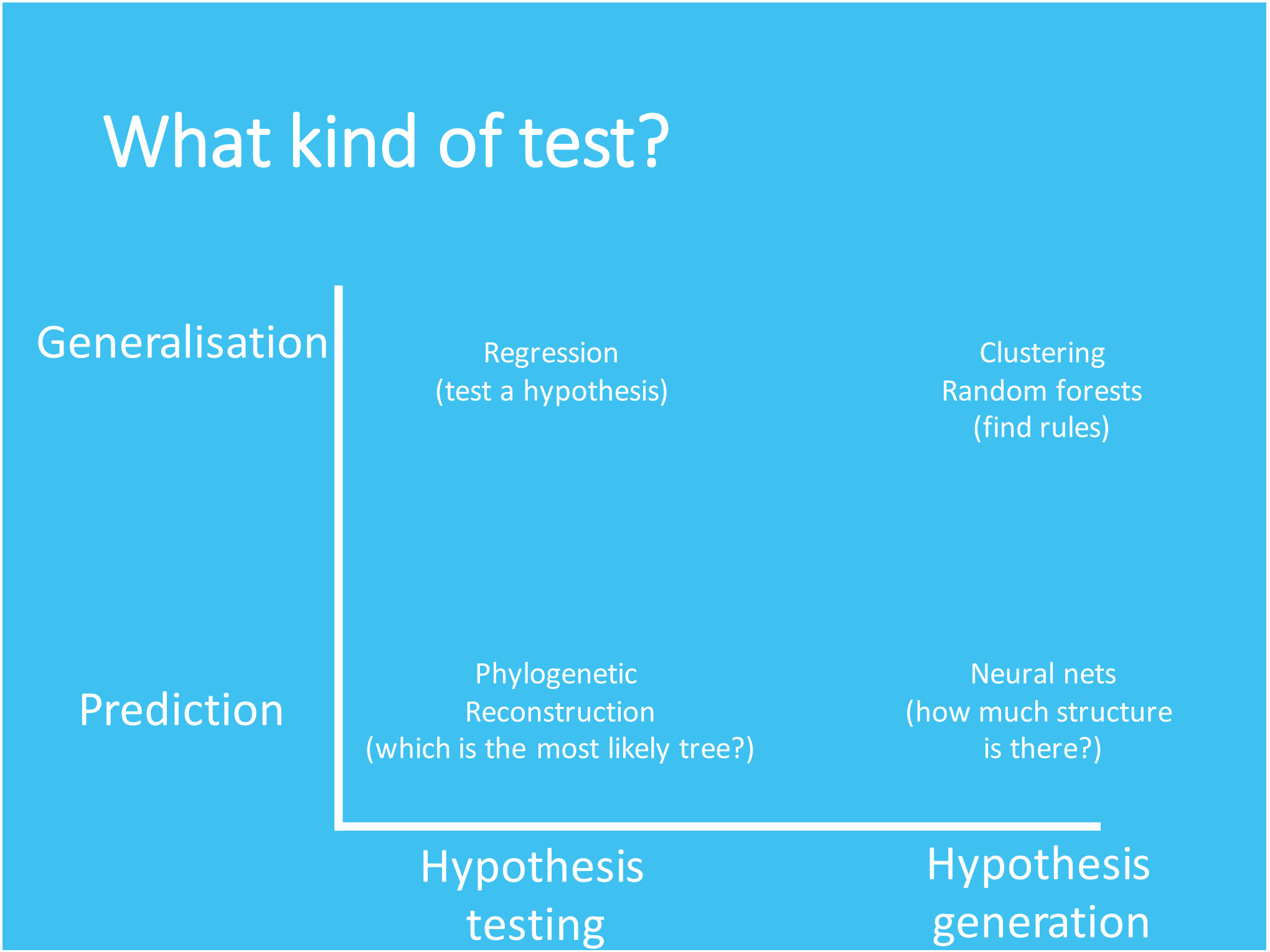
Concepticon:

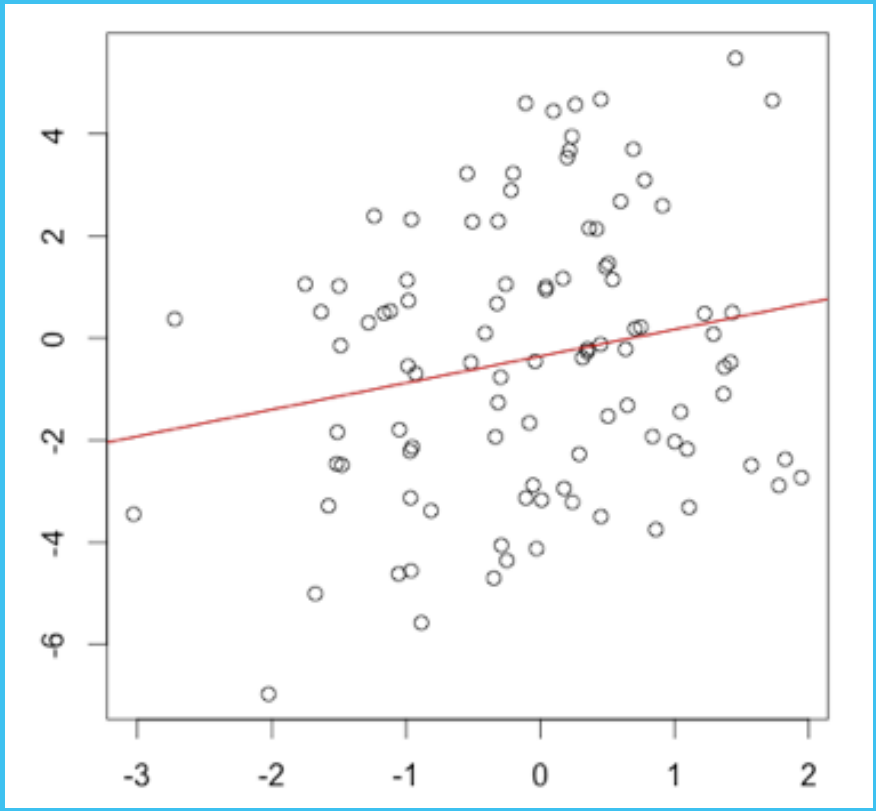
http://concepticon.clld.org/

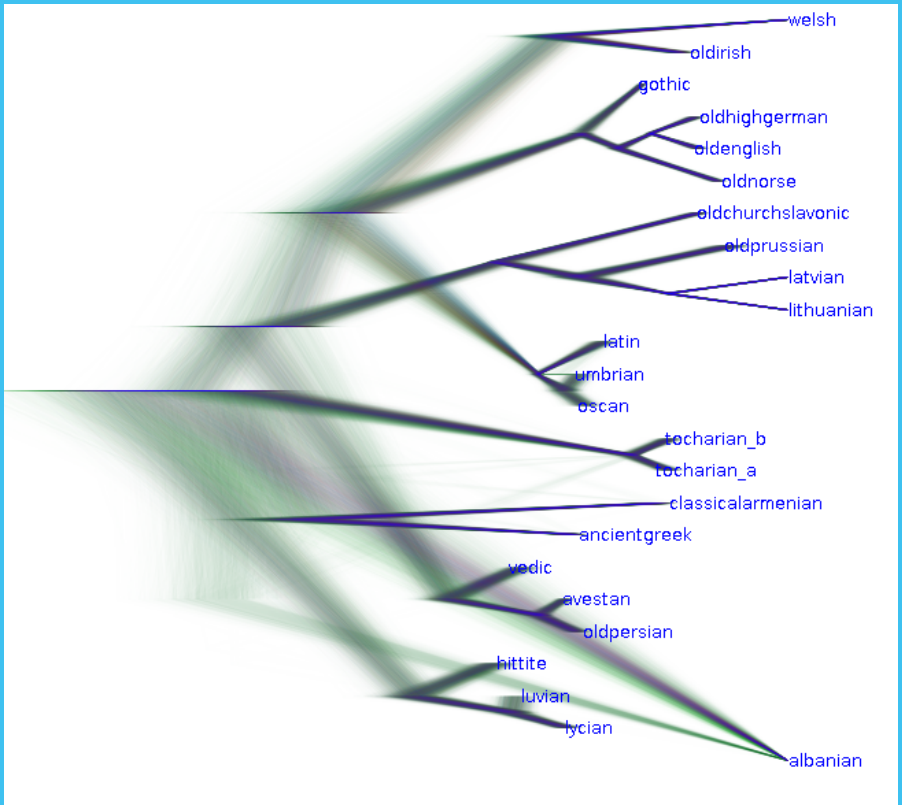
What kind of data?



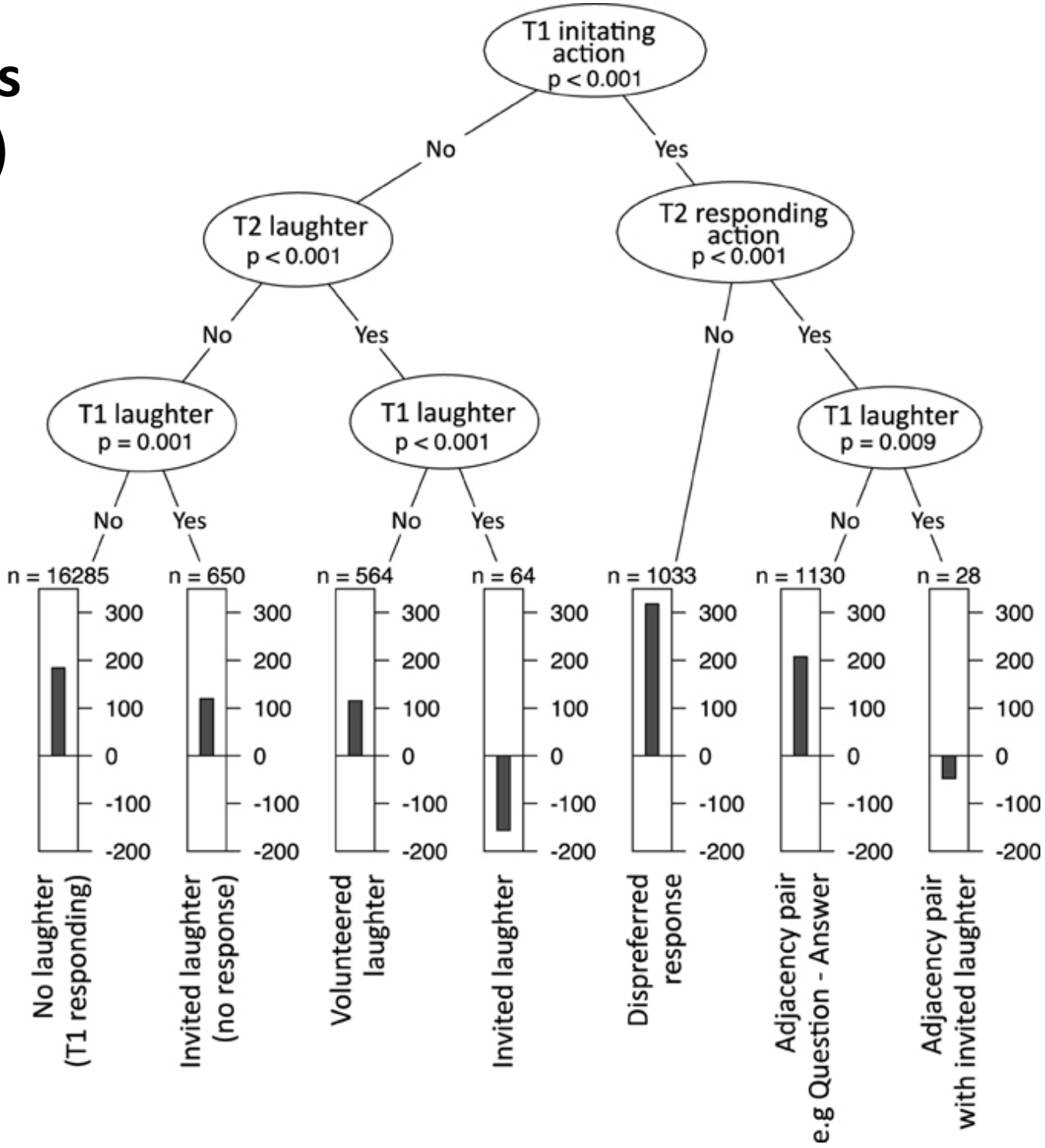
What kind of test?



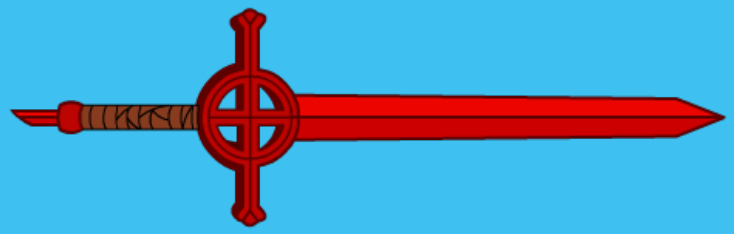

4
2
0
-2
-4
-6
-3
-2
-1
0
1
2


welsh
oldirish
gothic
oldhighgerman
oldenglish
oldnorse
oldchurchslavonic
oldprussian
latvian
lithuanian
latin
umbrian
oscan
tocharian_b
tocharian_a
classicalarmenian
ancientgreek
vedic
avestan
oldpersian
hittite
luvian
lycian
albanian

Random Forests
(Hour of power)

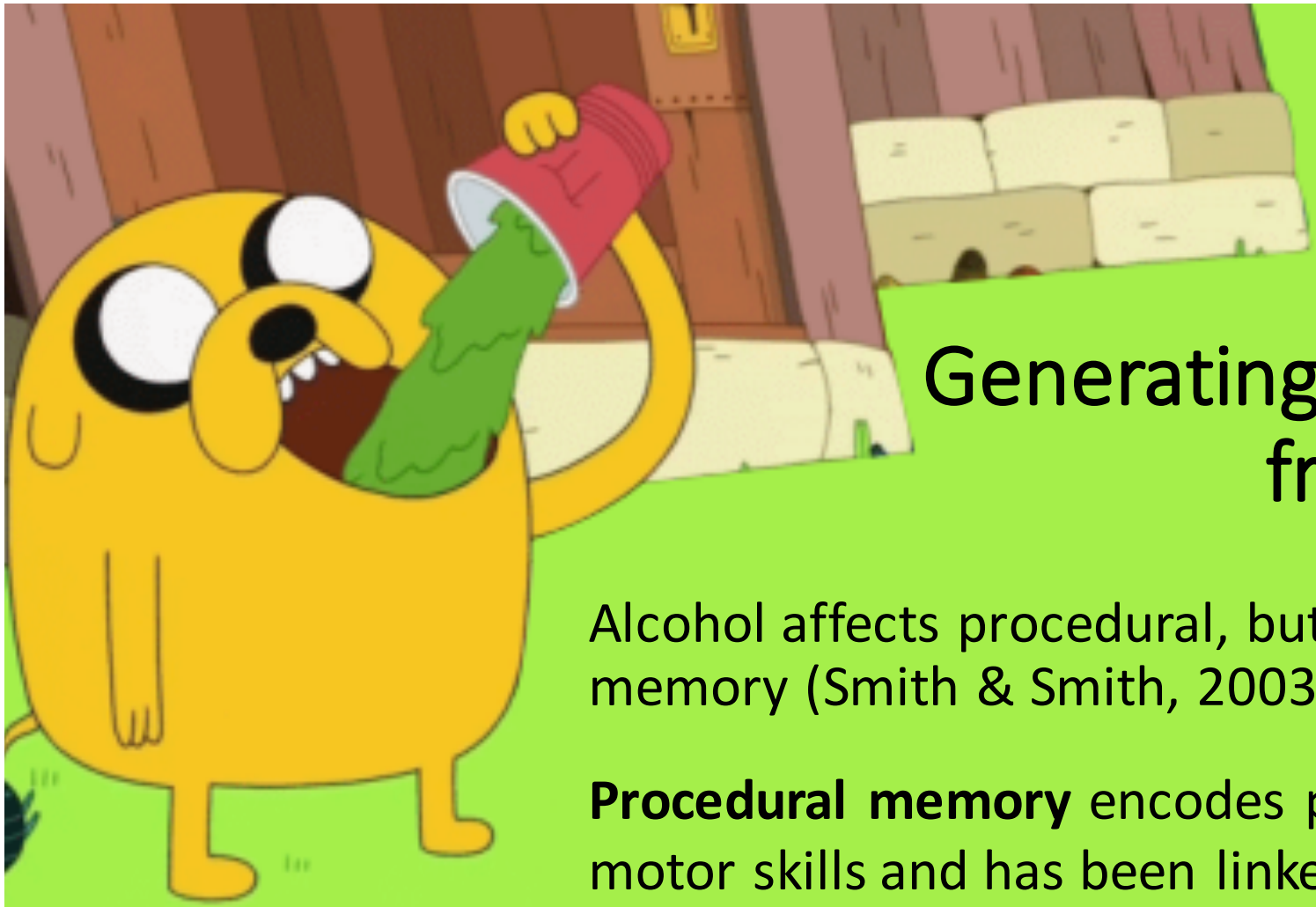


Approaches



Validity

Robustness



Generating hypotheses from theories

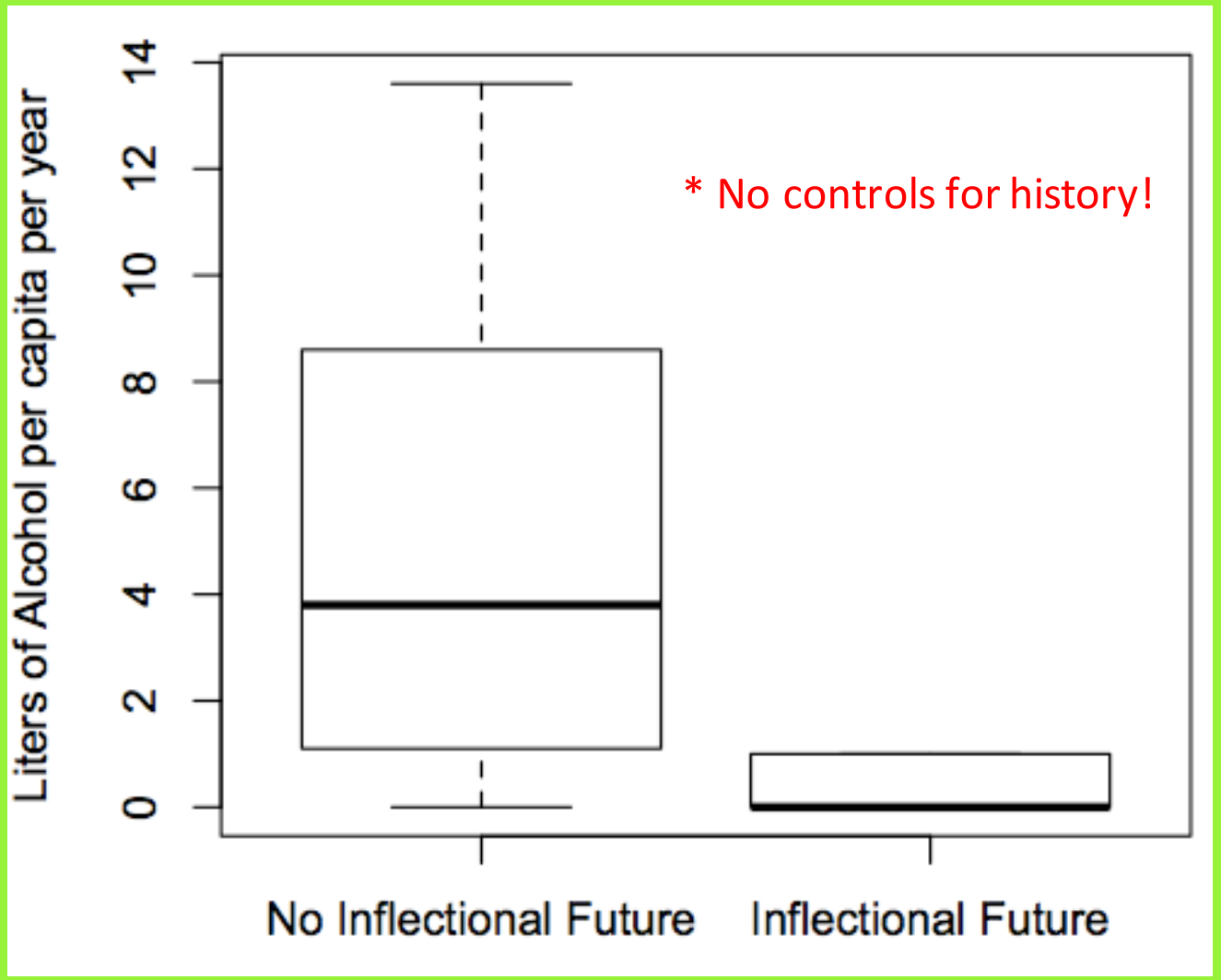
Alcohol affects procedural, but not declarative memory (Smith & Smith, 2003)

Procedural memory encodes procedures and motor skills and has been linked to grammar, morphology and pronunciation

Declarative memory stores facts and is used for retrieving lexical items

Smith C, & Smith D (2003). Ingestion of ethanol just prior to sleep onset impairs memory for procedural but not declarative tasks. *Sleep, 26* (2), 185-91 PMID: 12683478

Cultures which drink more alcohol rely on lexical encoding more than morphological encoding



Linking hypotheses to data

Ritual sacrifice and climate: Witch trials

Competing hypotheses:

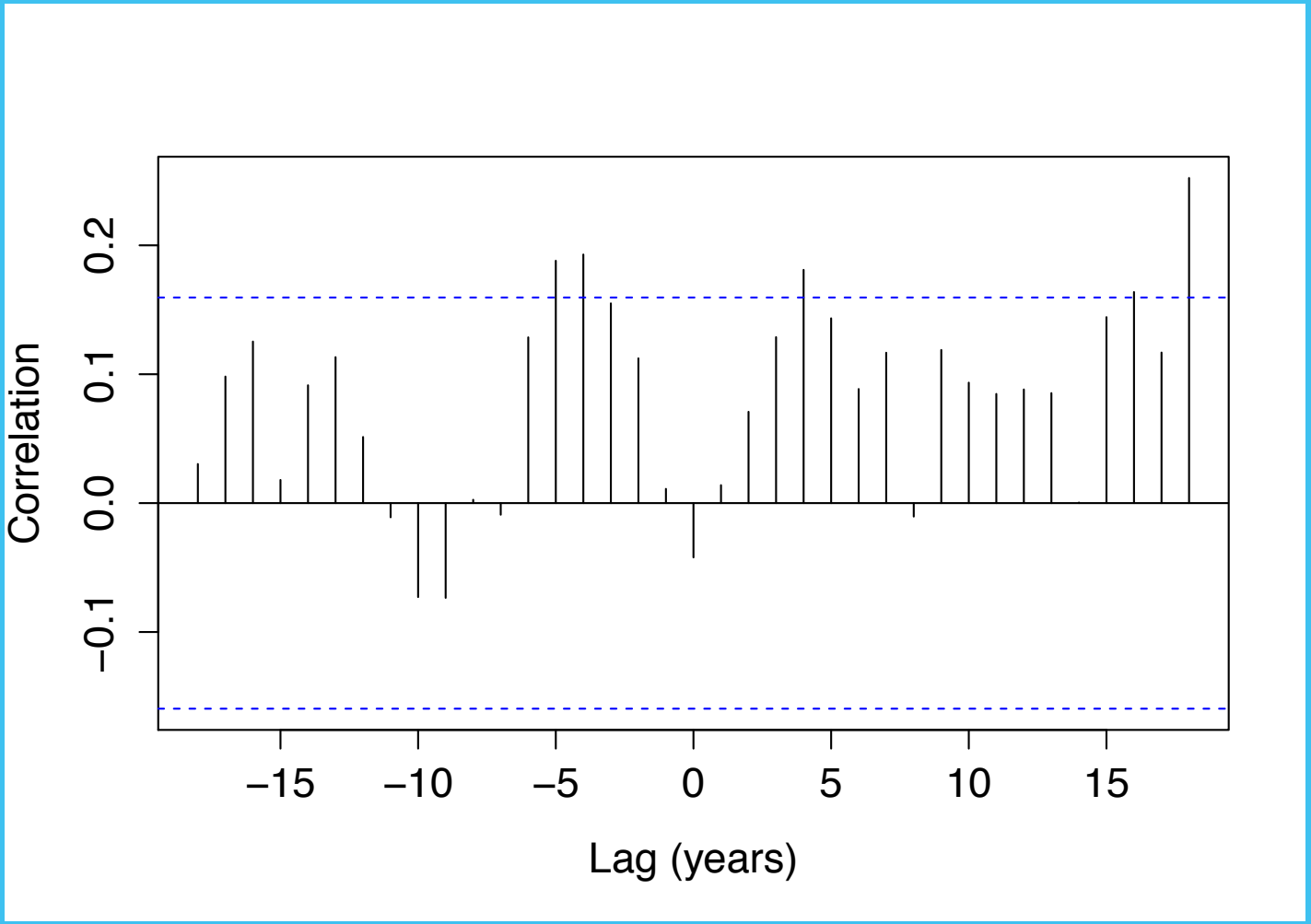
H0 No link between crop failure and witches

H1 Crop failure leads to social tension, which causes people to accuse others of being witches

Data:

List of witch trials in Essex, 1560-1700

Palmer Drought Severity Index for Essex, 1500 – 1700



What's the right baseline?

Coin toss

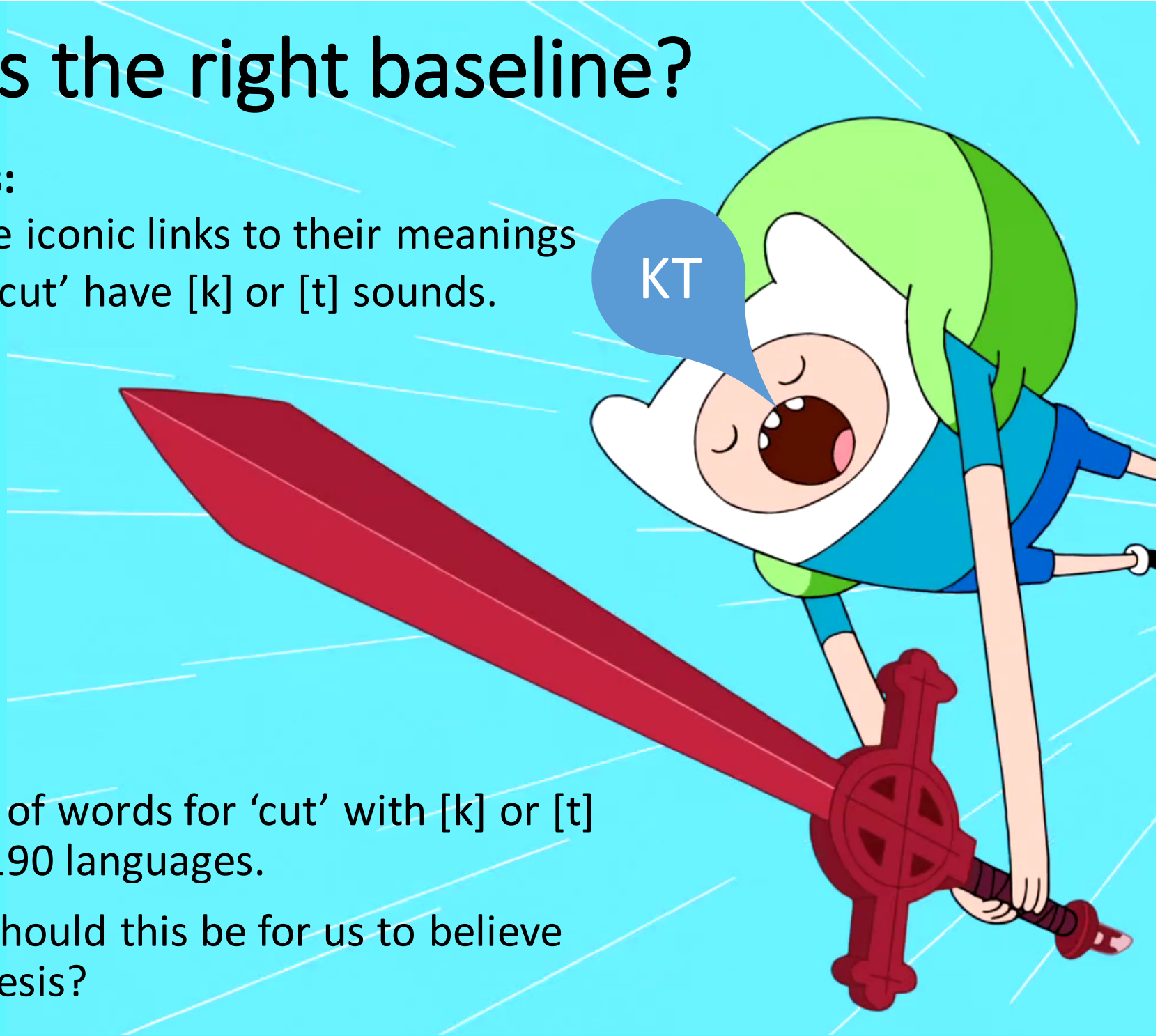
What's the proportion of heads we expect by chance?

What's the right baseline?

Hypothesis:

Words have iconic links to their meanings

Words for 'cut' have [k] or [t] sounds.



Data:

Proportion of words for 'cut' with [k] or [t] sounds in 190 languages.

How high should this be for us to believe the hypothesis?

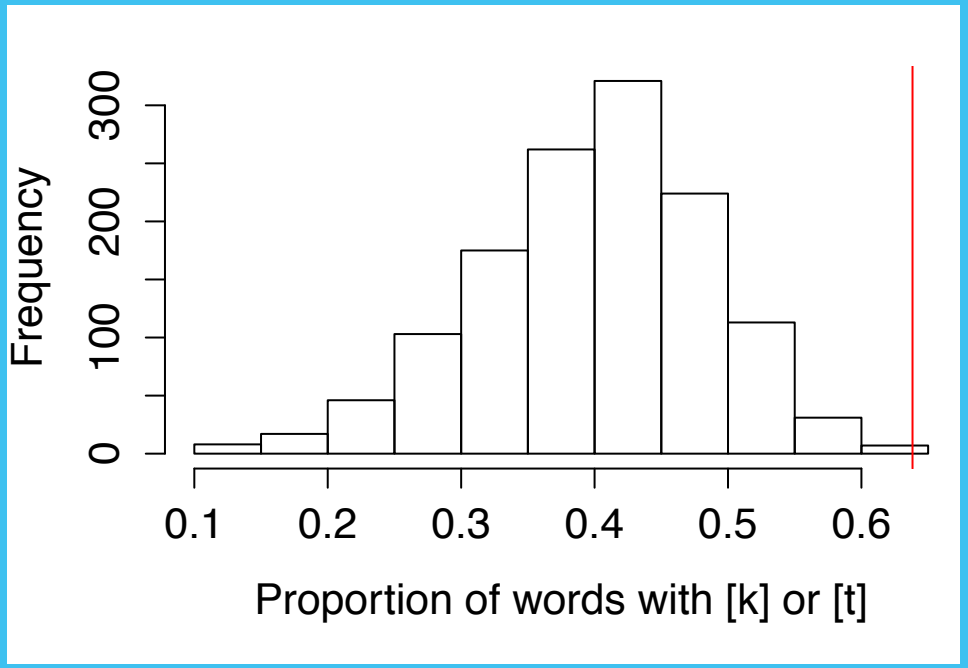
[demo]

Using other concepts as a baseline

Use proportion of words with [k] and [t] for other concepts. (data from IDS, WOLD, Spraakbanken)

[kt] in Cut words = 63%

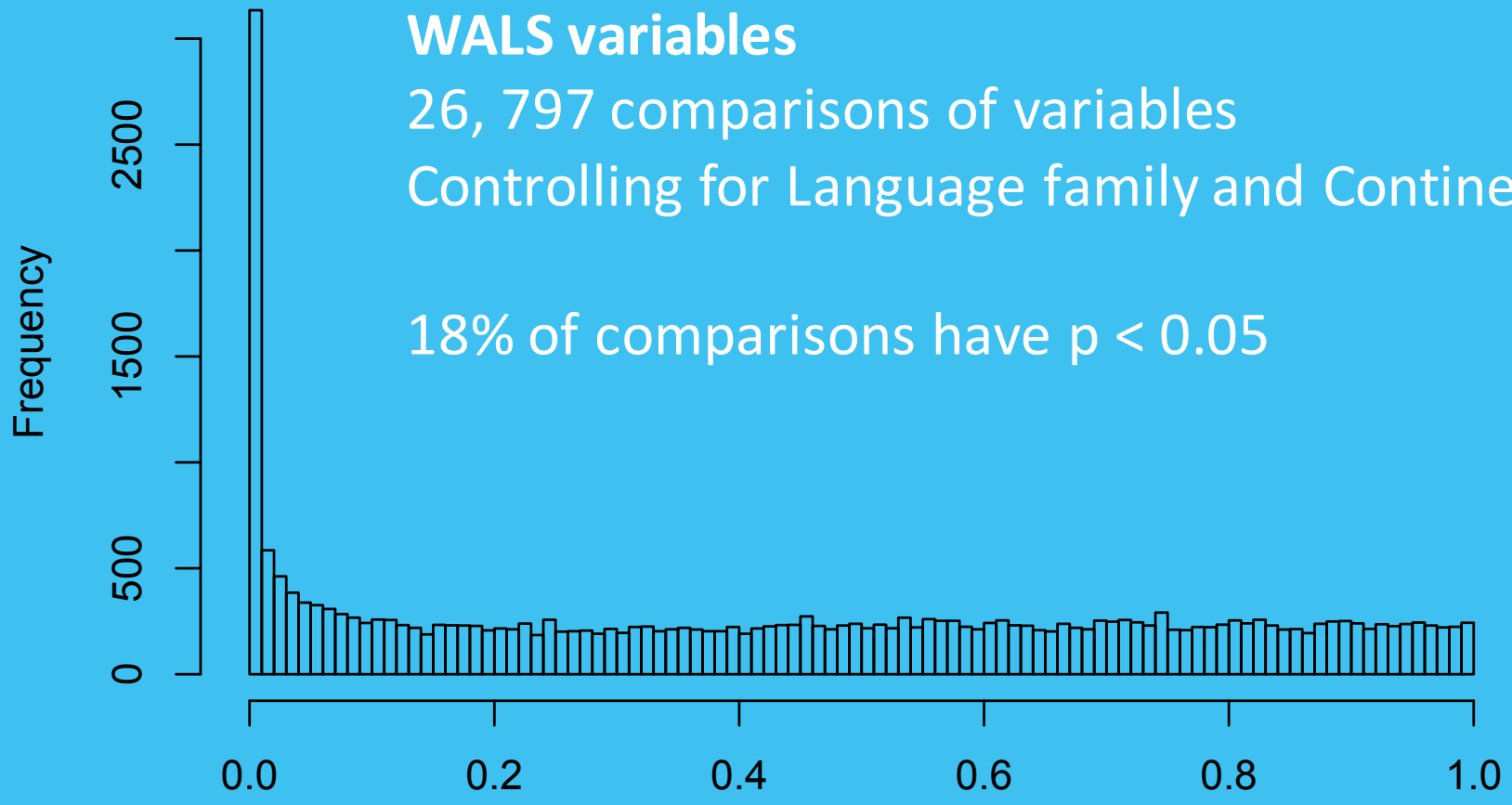
For 1306 other concepts:



Cut has more words with [kt] than 99.85% of other concepts
(p = 0.0015, z = 2.74)

Only 2 other concepts have more [kt] words: ‘basket’ and ‘break’.

Serendipity

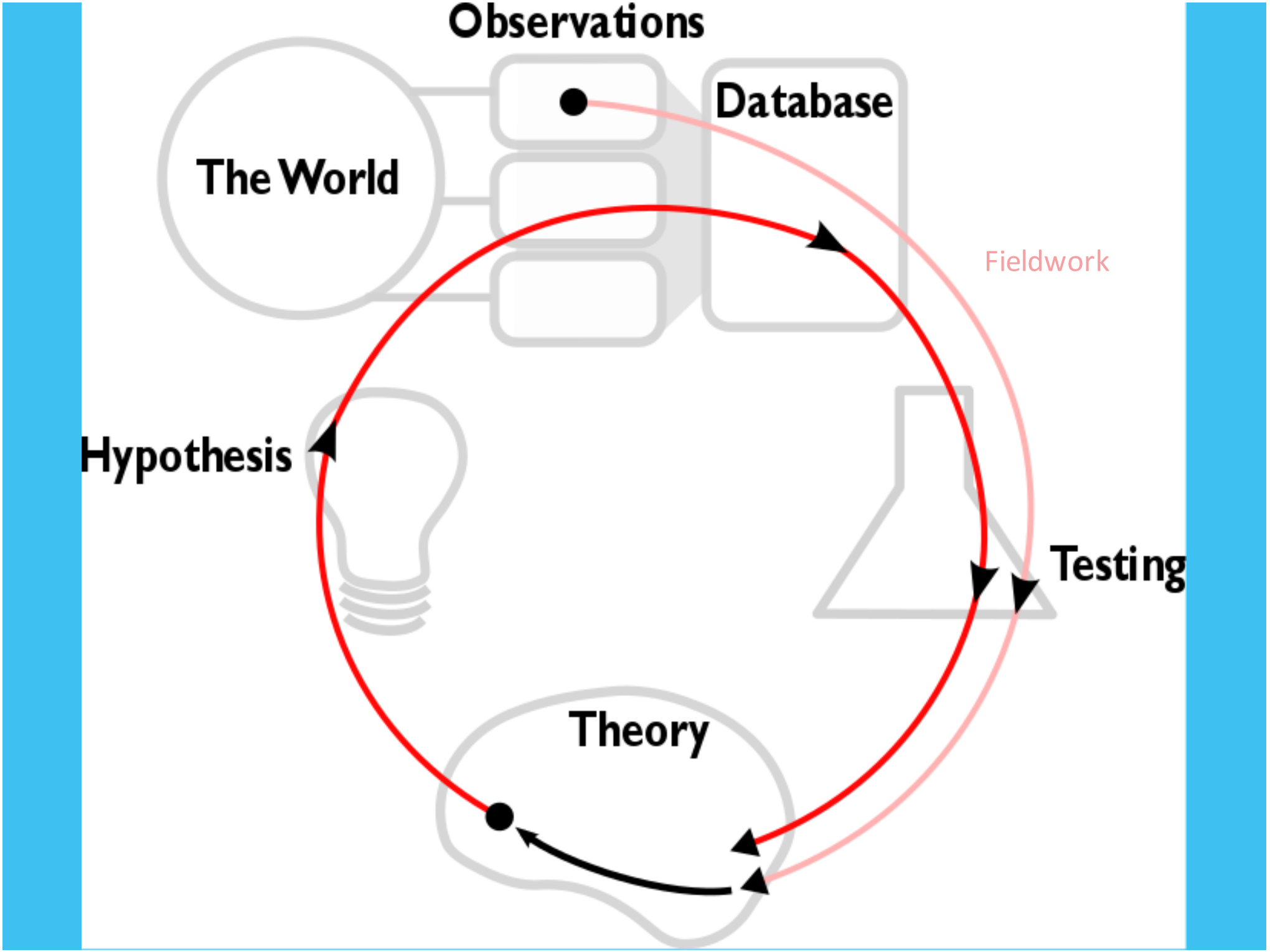


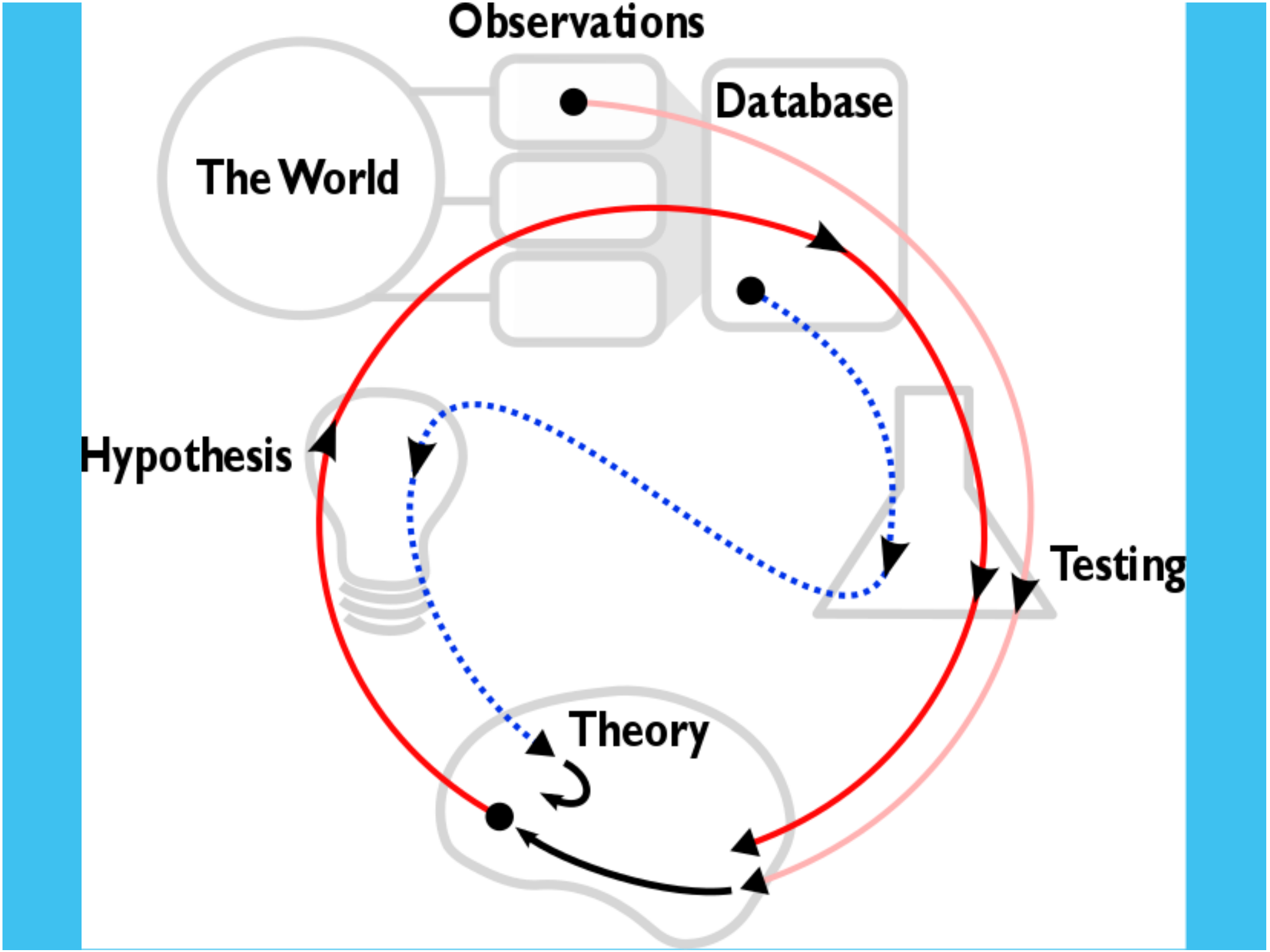
WALS variables

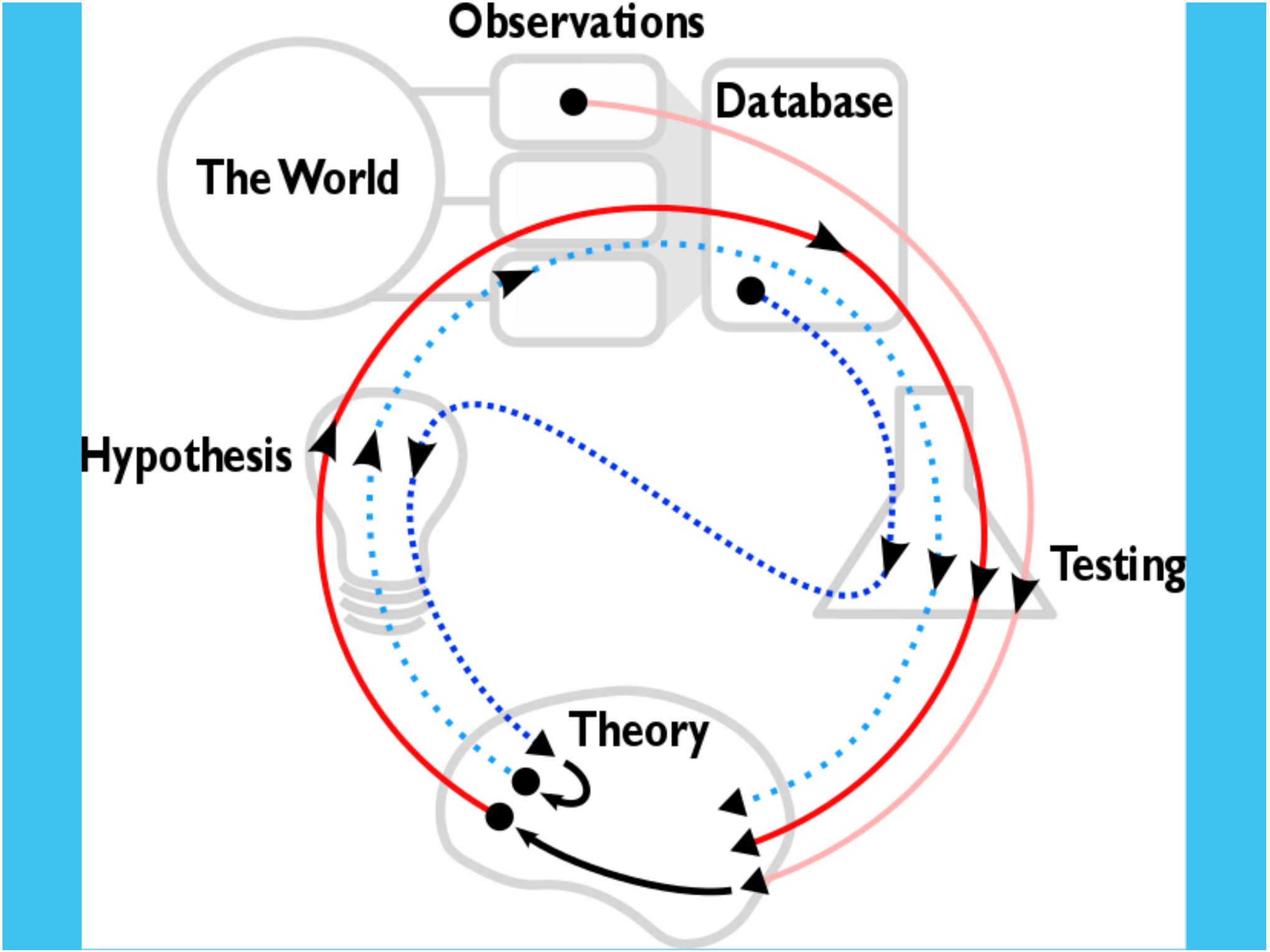
26, 797 comparisons of variables

Controlling for Language family and Continent

18% of comparisons have $p < 0.05$

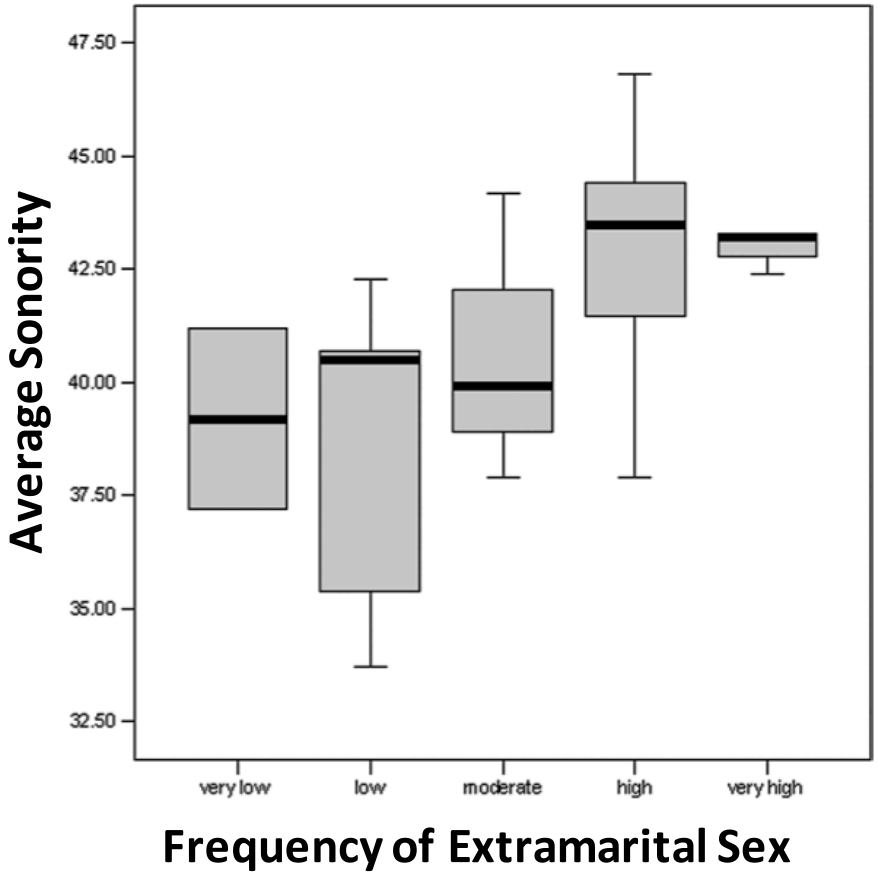

Observations
Database
The World
Fieldwork
Hypothesis
Testing
Theory


Observations
Database
The World
Hypothesis
Testing
Theory


Observations
Database
The World
Hypothesis
Testing
Theory

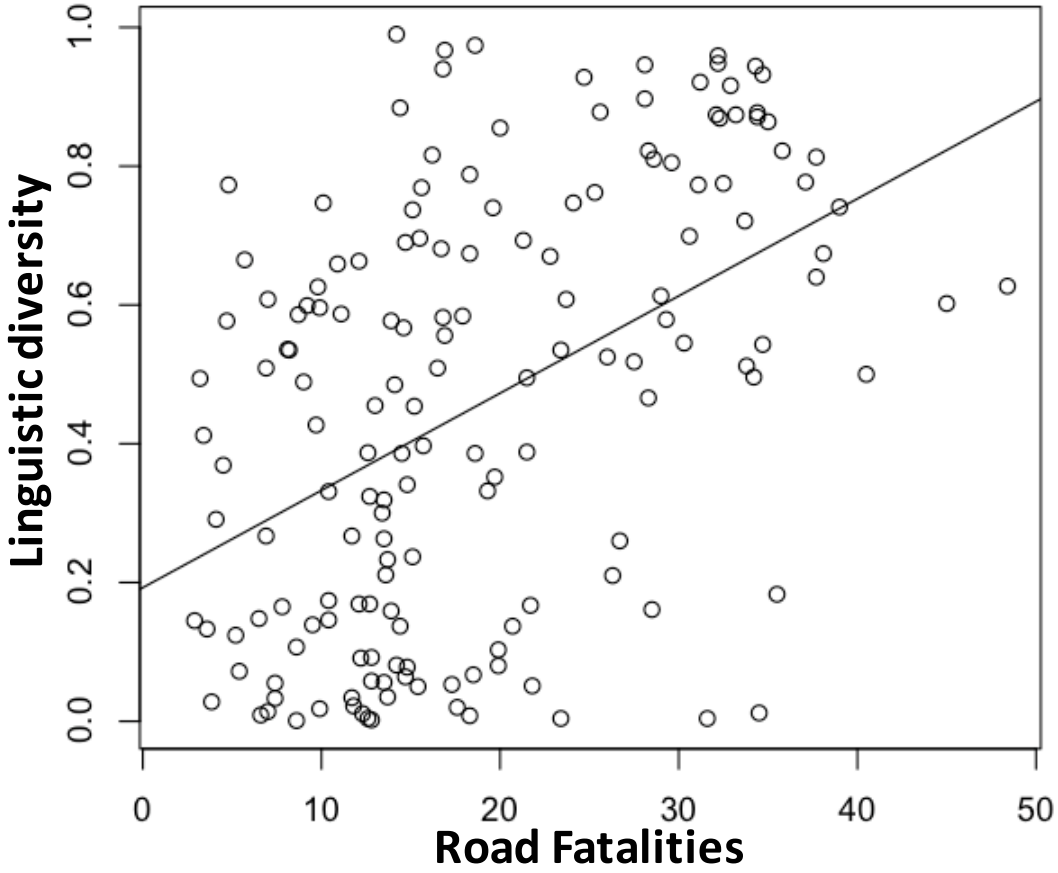
[break]

Galton’s Problem



r = 0.51, p = 0.01

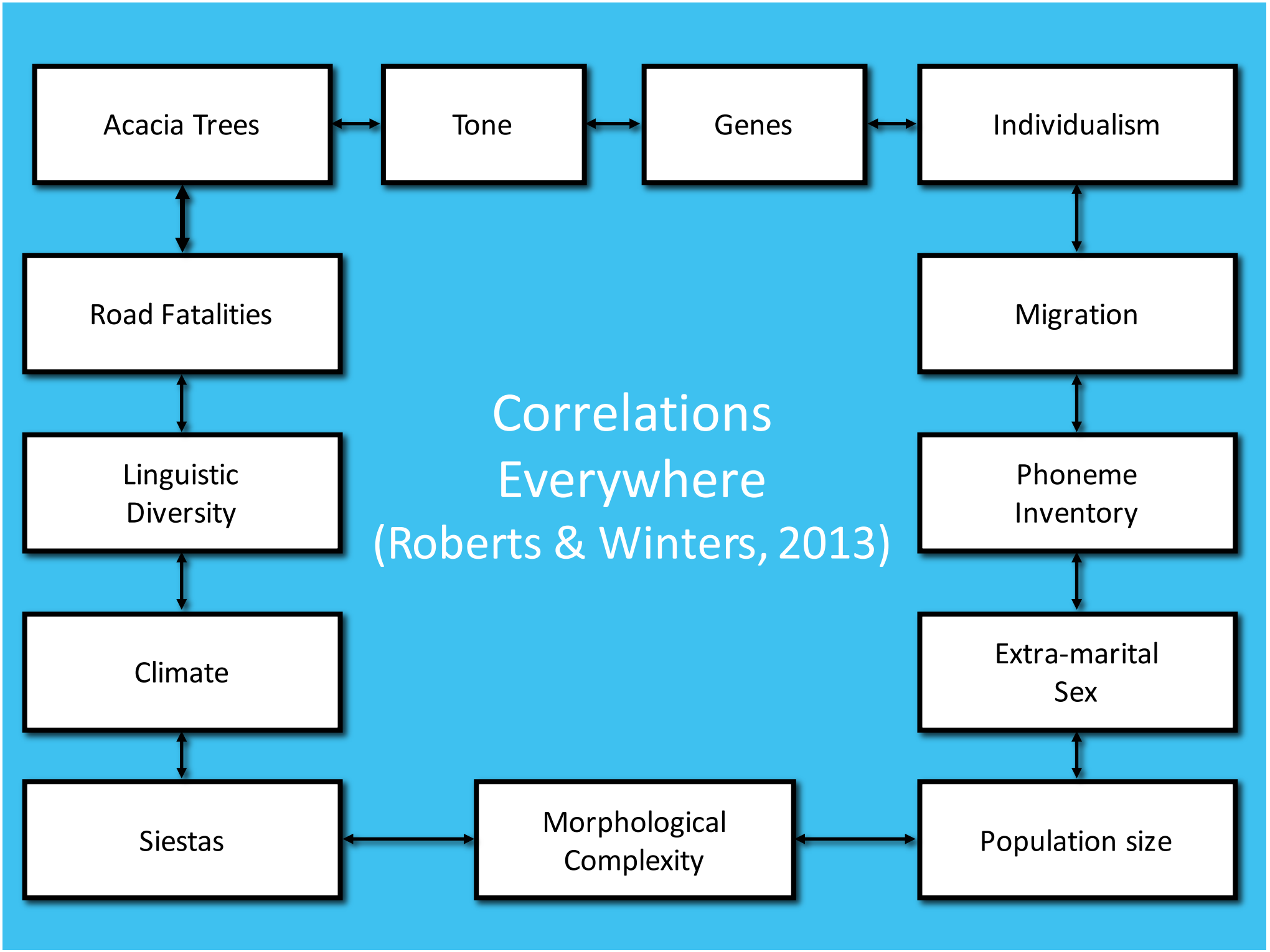
Ember & Ember (2007)

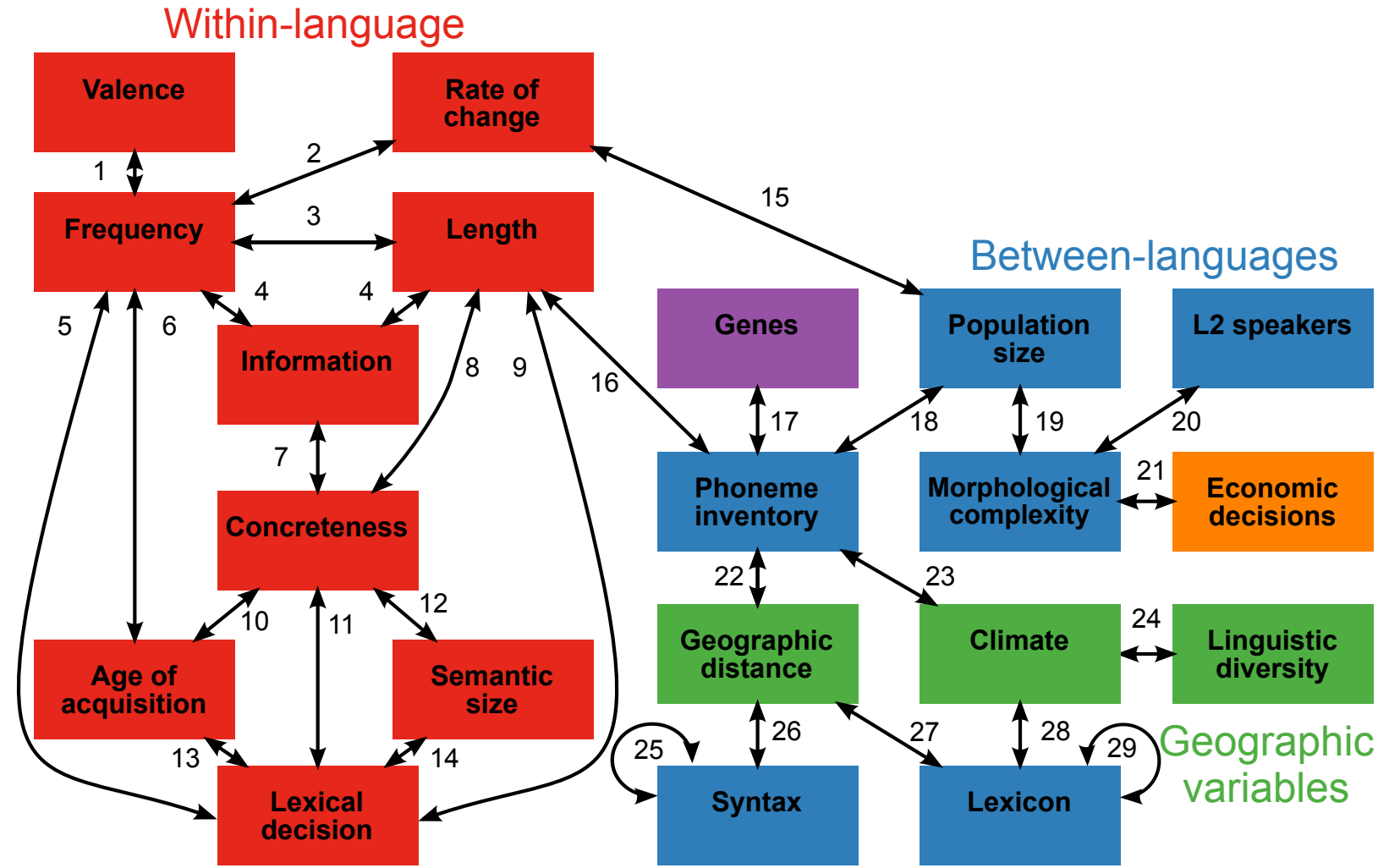


F (97,10) = 4.18, p<0.0001

Controlling for:

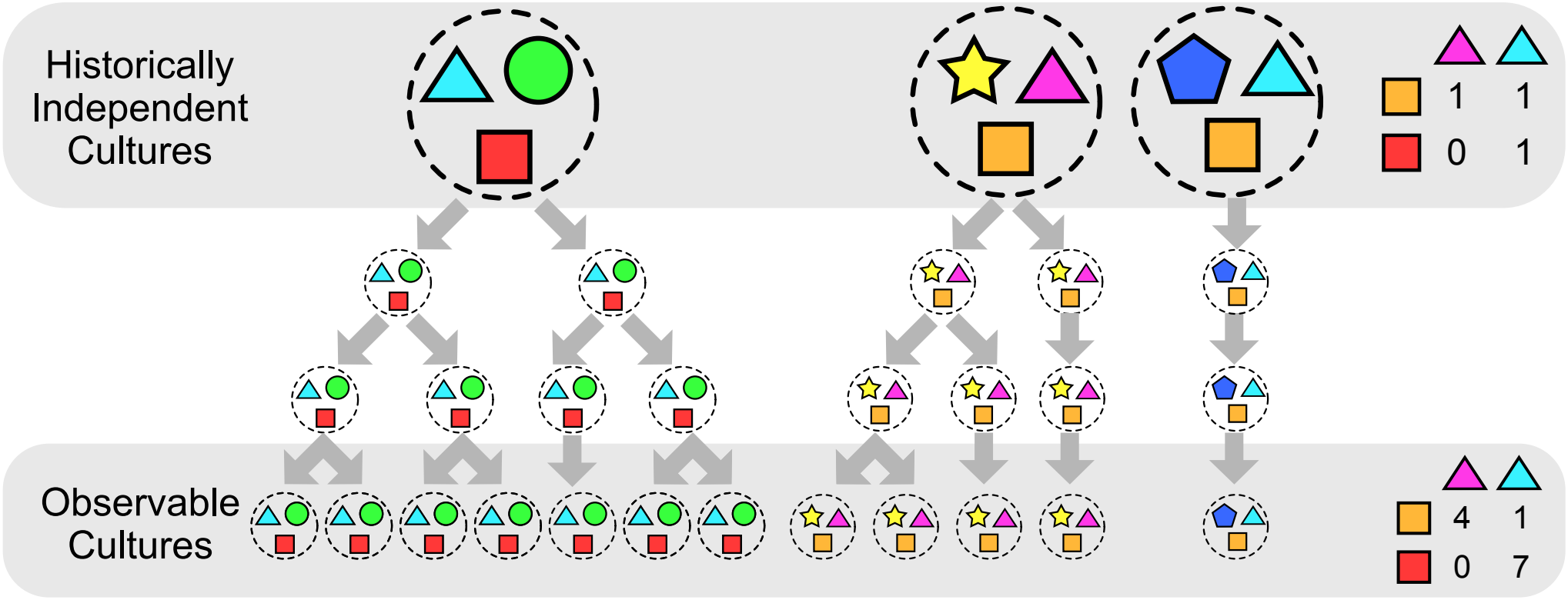
- Per-capita GDP
- Country nominal GDP
- Population density
- Migration
- Inside / outside Africa
- Distance from the equator


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)

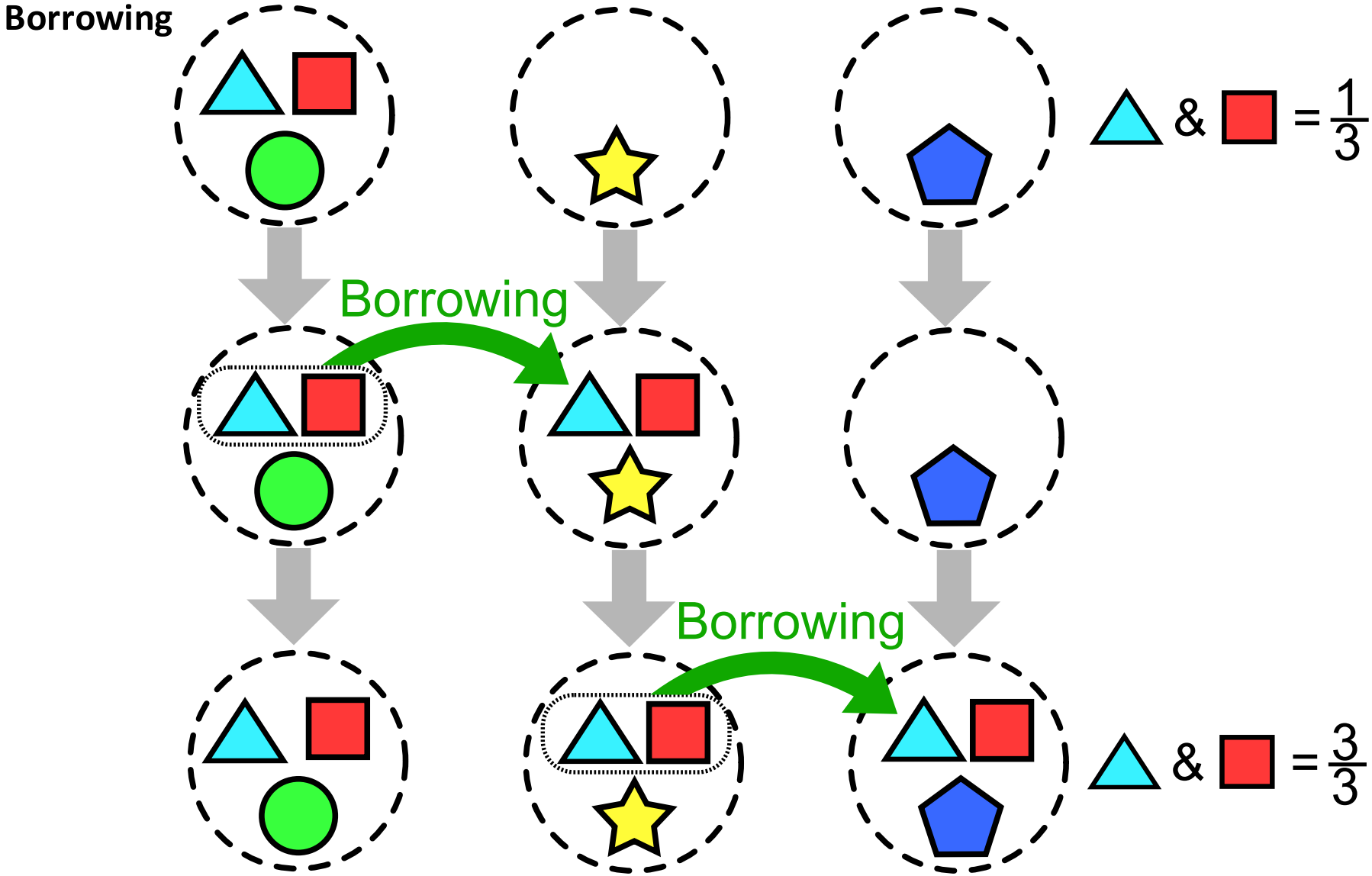


1. Boucher & Osgood (1969)
2. Pagel et al. (2007)
3. Zipf (1936)
4. Piantadosi et al. (2011)
5. Balota et al., 2004
6. Kuperman et al., 2013
7,8. Piantadosi et al. (2011b)
9. Hudson & Bergman, 1985
10. Reilly & Jacobs, 2007
11,12. Yao et al. (2013)
13. Walker & Hulme (1999)
14. Sereno et al. (2009)
15. Jordan & Currie (submitted)
16. Nettle (1999)
17. Dediu & Ladd (2007)
18. Hay & Bauer (2007)
19. Lupyan & Dale (2010)
20. Bentz & Winter (2013)
21. Chen (2013)
22. Atkinson (2011)
23. Everett (2013)
24. Nettle (1999)
25. Dunn et al. (2011)
26. Spruit (2006)
27. Gray et al. (2009)
28. Lindsay & Brown (2004)
29. Majid et al. (2008)

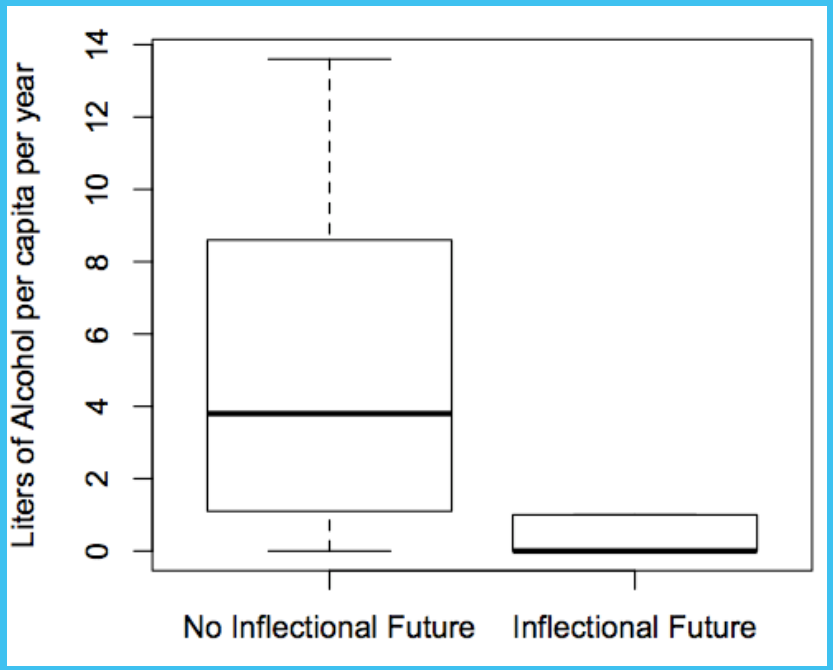
Historical relationships inflate correlations



Correlations due to borrowing



Correlations due to the structure of the data



Future			Alcohol
Y	Language 1	Country 1	0.9
N	Language 2		
N	Language 3		
N	Language 4	Country 2	4.5
Y	Language 5		
Y	Language 6		
N	Language 7	Country 3	2.1

Contingency tables are not enough!

		Adposition	
		Postpositions	**Prepositions**
Word Order	**VO**	6	76
	SV	374	11

Solutions

Independent samples

Regression with random effects

Regression with phylogenetic tree

Phylogenetic simulation

Arguing against analyses

Keith Chen

Savings and future tense

Speakers of languages with no future tense are less likely to save money.

EUROTYP project (Dahl, 1985)

English

*It **will be** mostly cool and windy*

Spanish

Ya desde	la mañana	el viento	será muy	flojito
Already from	the morning	the wind	be+FUT very	weak

From the morning, the wind will be very weak.

Finnish

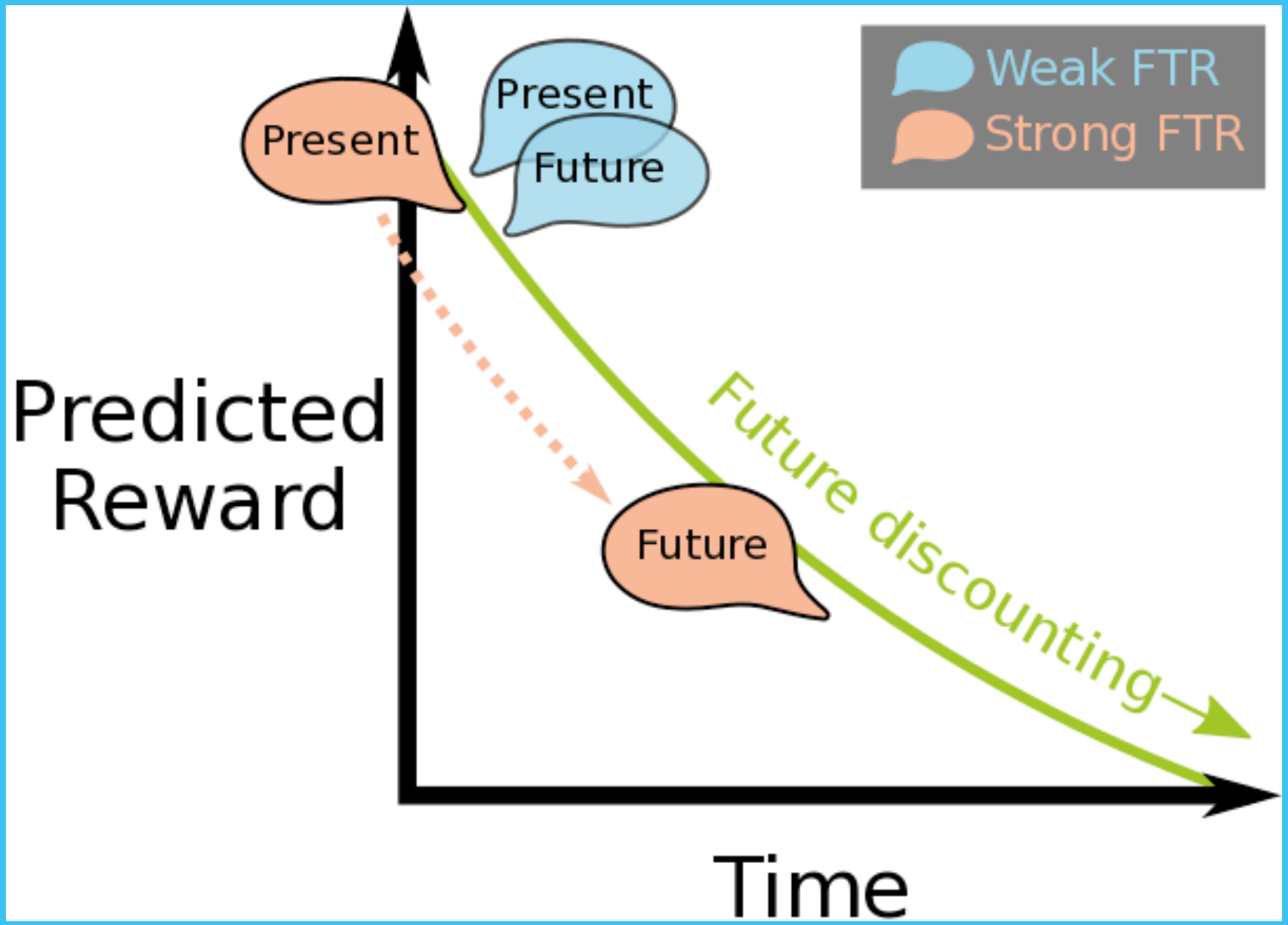
Sää kylmenee,	mutta keskiviikkona	tuulee idästä	ja pyryttää	lunta
Weather grow-cold+PRES	but Wednesday	blow+PRES east	and drifting	snows

The weather becomes cooler, but on Wednesday (the wind) blows from the east and there is drifting snow.

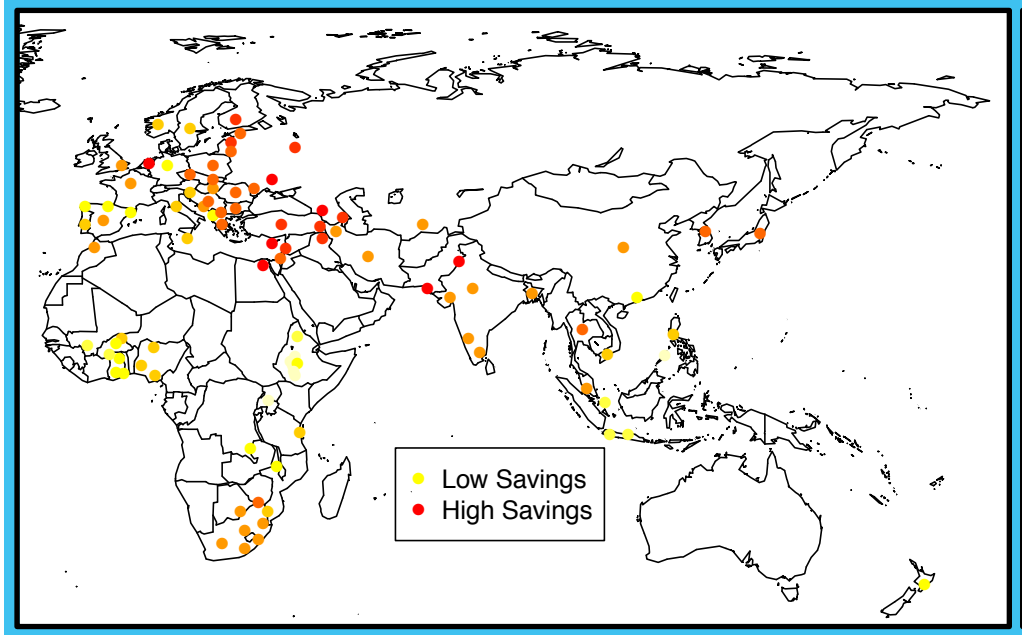
World Values Survey - 200,000 survey participants

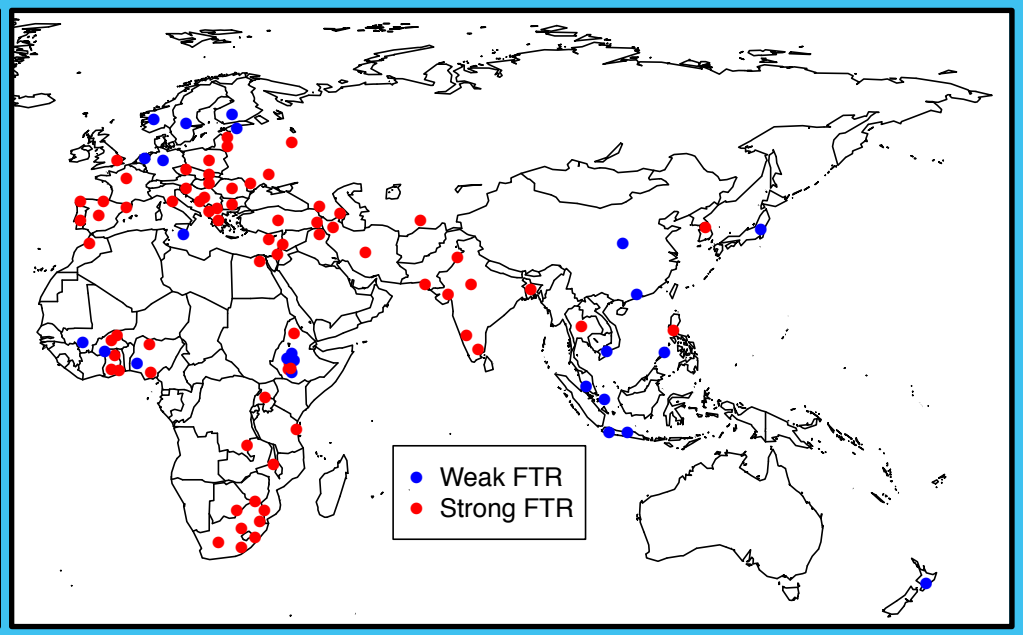
During the past year, did your family save money?

What language do you speak at home?


Weak FTR
Strong FTR
Present
Present
Future
Future
Predicted
Reward
Future discounting
Time

Savings and future tense





Control for:

- GDP
- GDP growth
- Origin of legal system
- Legal rights index
- Interest rate

Compare individuals who are identical in:

- Country of residence
- Year of survey
- Level of Education
- Religion
- Employment

Strong-FTR families saving only 46% as often as weak-FTR

Responses

I don’t believe you

What about this one counterexample?

The typology is wrong/ not detailed enough, so the theory is wrong

Correlation does not imply causality

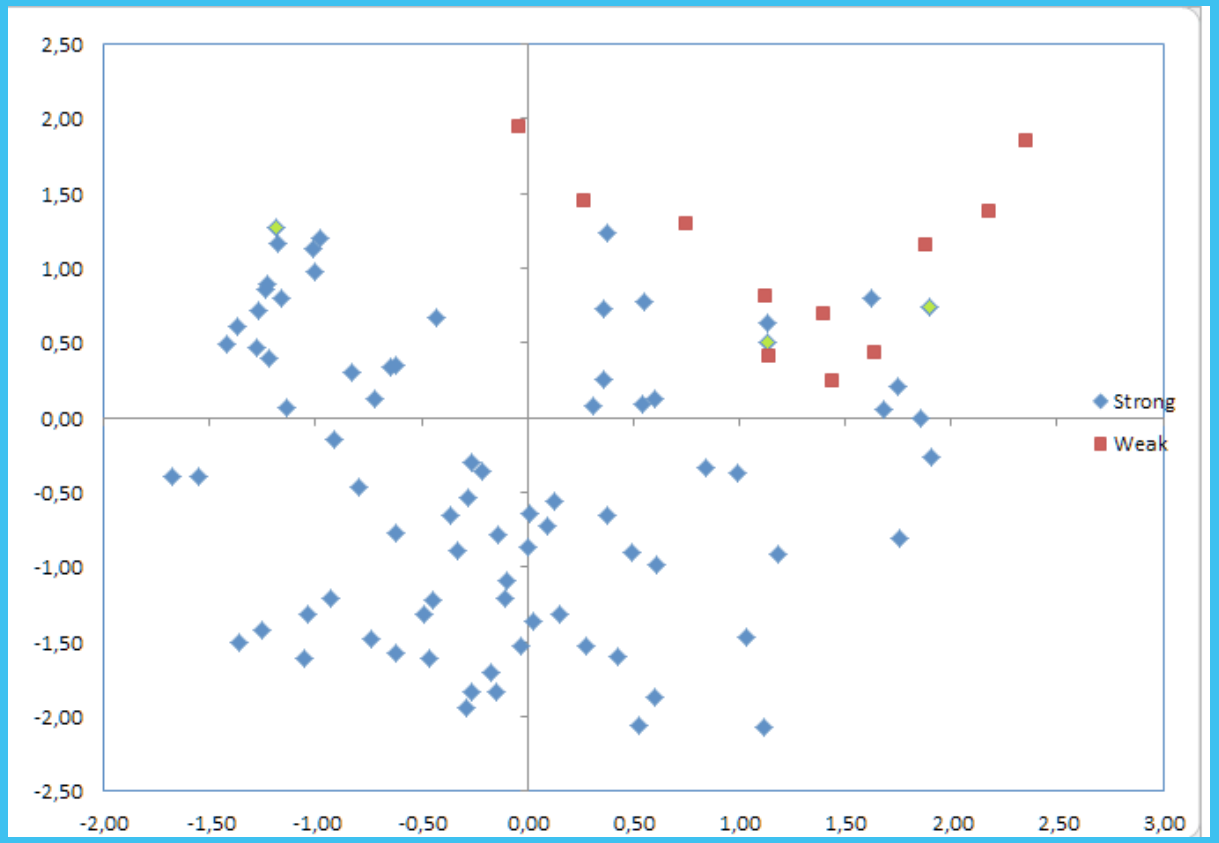
Good arguments

Pointing out a weakness in the study

which provides an alternative explanation of the results

Criticism

Östen Dahl: Confounding variables



Traditional/
Secular-rational
values

Survival/Self-expression

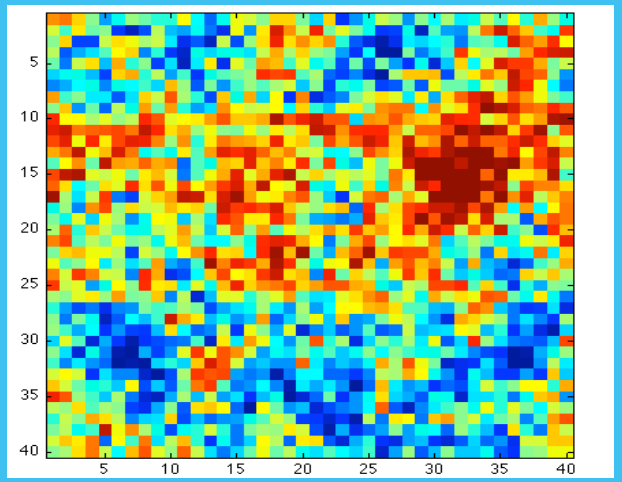
Inglehart & Welzel, see
http://dlc.hypotheses.org/360

Criticism

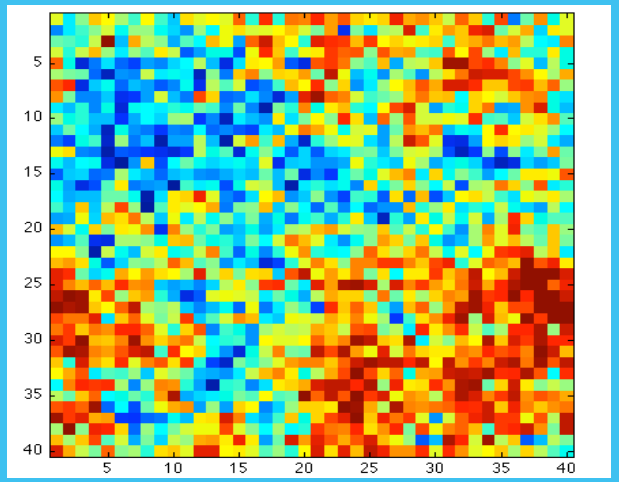
Problems with statistics

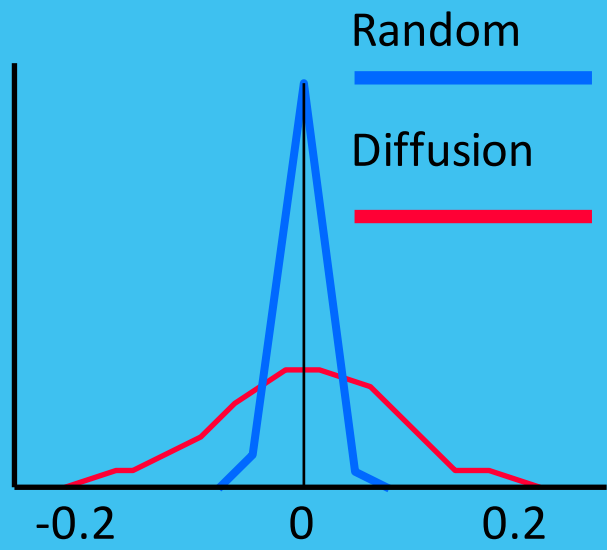
Mark Liberman: No control for geographic diffusion

Trait 1



Trait 2



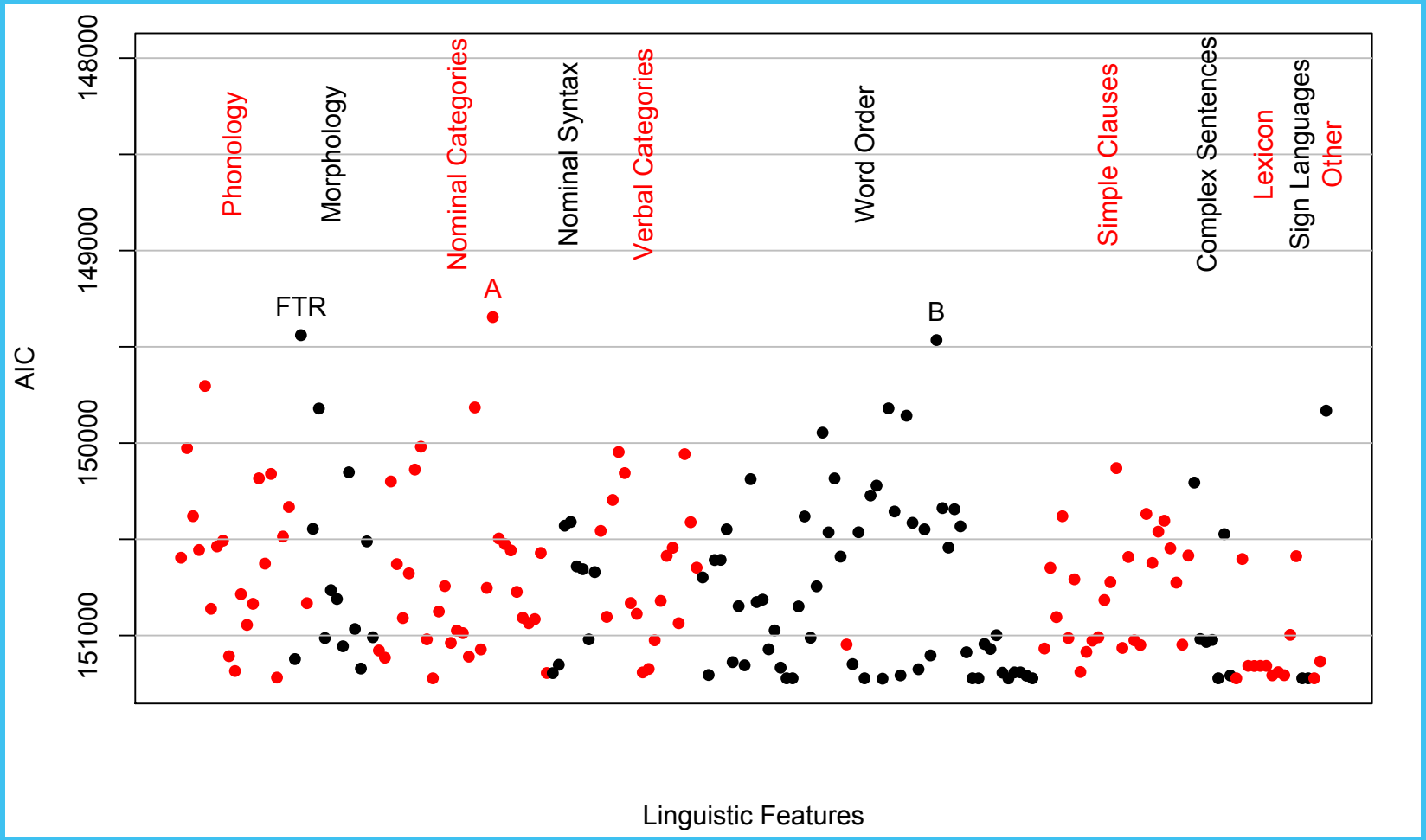


Geographic correlations are much more likely if features spread through local diffusion

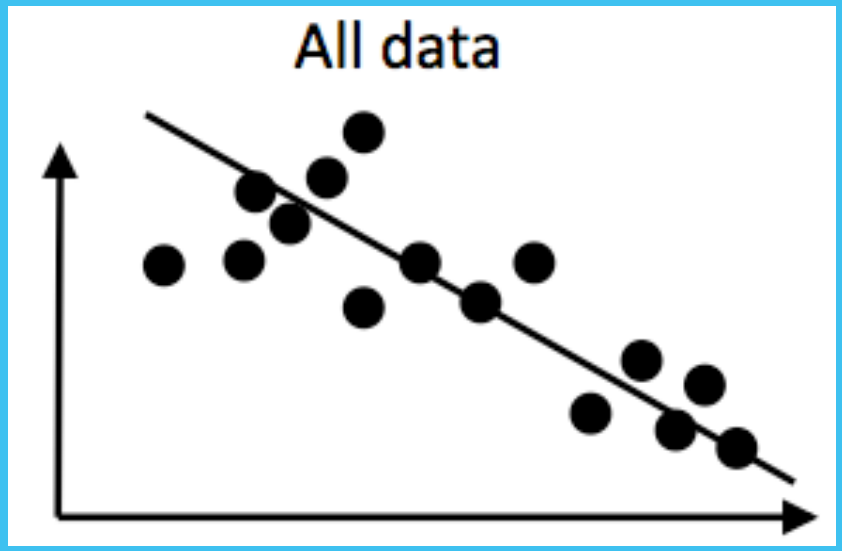
http://languagelog.ldc.upenn.edu/nll/?p=3764

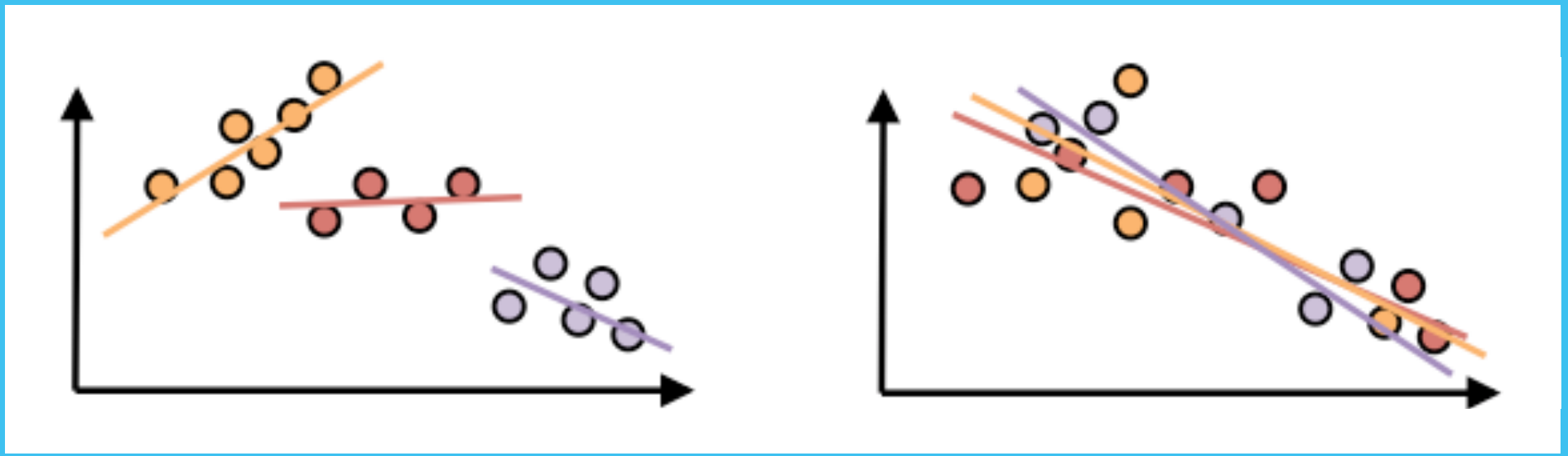
Serendipity test

Run the original regression with different 193 linguistic variables from WALS



Mixed effects modelling





Results

Roberts, Winters & Chen (2015)

James Winters

Keith Chen

Random effects for:

- **Inheritance** **(language family)**
- **Borrowing** **(linguistic area)**
- **Economic policies** **(state)**

Wald-z test ($z = 1.58$, $p = 0.11$) , Likelihood ratio test ($\chi^2 = 1.15$, $p = 0.28$)

Correlation is driven by relationships in the data

The same model for employment status, trust and sex are significant

Responses

I don’t believe you

Stitch said,
August 15, 2015 @ 11:03 am

I'm so old…I can remember when English speakers were considered the rock-ribbed Calvinist work-ethical inventors of capitalism. Now it's apparently necessary to come up with an explanation of why we're so shiftless and feckless compared to everybody else. Has the language (or the behavior) changed a lot without me noticing it?

What about this one counterexample?

Xmun said,
August 14, 2015 @ 2:54 pm

Doesn't everybody know that Jews are good at saving money although Hebrew has no future tense? (Allow me to share this thought although I know it's pretty silly.)

The typology is wrong/ not detailed enough, so the theory is wrong

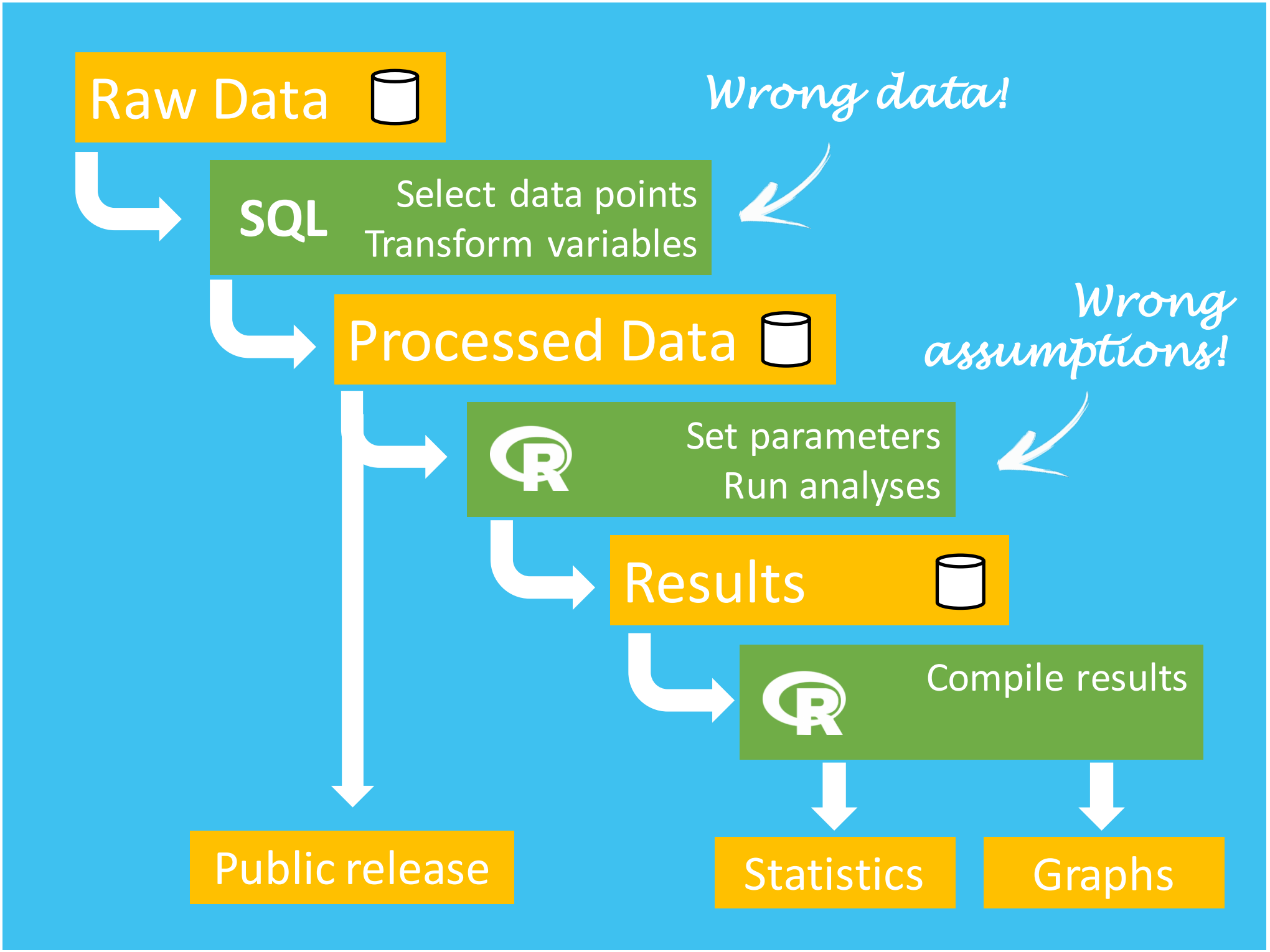
@myl
I didn't notice the comment by Östen Dahl in the thread by Chen before, it would seem to confirm my suspicion that Chen engaged in some degree of synthesis in converting the "raw data" into a stron/weak FTR classification. In that post, Chen seems to indicate that his criteria is the obligatory marking (it sounds like only inflectional and grammaticized periphrastic marking is considered to "count") of future time reference under conditions

Correlation does not imply causality

Daniel de França said,
August 14, 2015 @ 6:19 pm

It might be that things are correlated, but the order of causation is inverted. The lack of future mark is due not saving, not the other way around.

Arguing with data


Raw Data
Wrong data!
SQL
Select data points
Transform variables
Processed Data
Wrong assumptions!
Set parameters
Run analyses
Results
Compile results
Public release
Statistics
Graphs

Parameter Robustness

Try different parameters for the same analysis

➔ Results are not simple artifacts from a model.

Structural Robustness

Try different datasets or conceptual frameworks

➔ Common results across different approaches makes it possible to identify the core causal structure that generates these results.

Representational Robustness

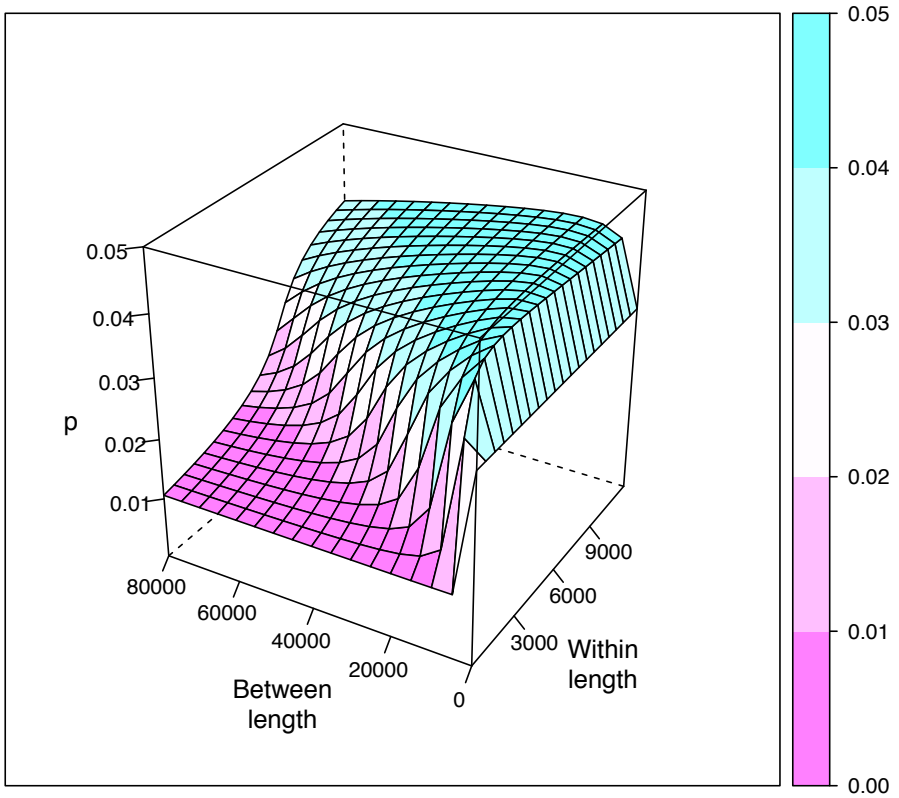
Try different frameworks (R, python etc.)

➔ Results are not related to specific features of a computational frameworks.

A space of results

Test	Is the correlation robust?
Mixed effects model	**No**
Regression on matched samples	**Yes**
Serendipity test	**Yes**
Independent samples	**Yes**
Partial Mantel test	**Yes**
Partial Stratified Mantel test	**Yes**
Geographic autocorrelation	**Yes**
Phylogenetic Generalised Least Squares	**Yes**
PGLS within families	**No**

Parameter robustness

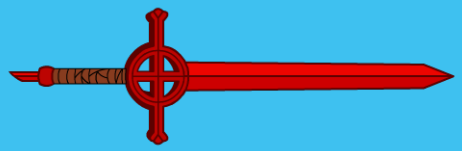


Structural robustness:
WALS and Glottolog language families
Representational robustness:
Mixed effects done in *lme4* and *blme*

A space of results

Test	Is the correlation robust?	Individual data	Control for language family	Control for geographic area	Control for country
Mixed effects model	**No**	Yes	Yes	Yes	Yes
Regression on matched samples	**Yes**	Yes	Yes	No	Yes
Serendipity test	**Yes**	Yes	Yes	No	Yes
Independent samples	**Yes**	No	Yes	No	No
Partial Mantel test	**Yes**	No	Yes	Yes	No
Partial Stratified Mantel test	**Yes**	No	Yes	Yes	No
Geographic autocorrelation	**Yes**	No	No	Yes	No
Phylogenetic Generalised Least Squares	**Yes**	No	Yes	No	No
PGLS within families	**No**	No	Yes	No	No

Arguing with data



Maximum Validity method:
Set out assumptions
Code data according to assumptions
Run the most relevant test

Maximum Robustness method:
Run tests with as many assumptions and sources of data as possible
Demonstrate all tests give qualitatively the same answer
OR
Identify similarities in approaches which lead to negative results

Result is the best answer given the assumptions

Result is a space where researchers can argue about data

Easy to interpret
Susceptible to argument from authority

Engages more researchers
Susceptible to fishing / circular research

The Future

Collaboration

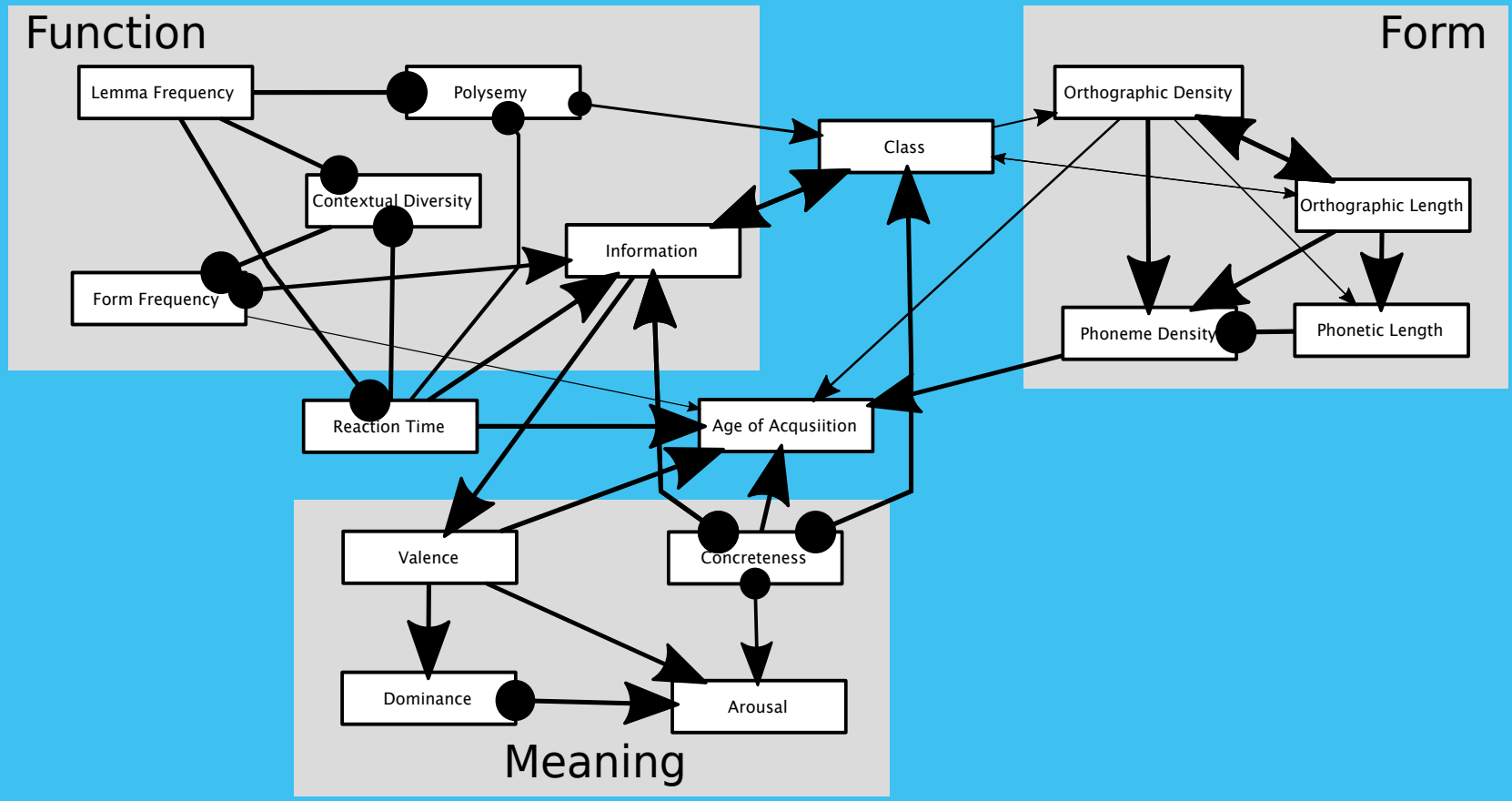
Statistical knowledge required to engage with debate

Using fast analyses as feasibility studies

Dependent variables -> Networks

Correlations between tone and:

Genes, Intonation, Climate



Questions?