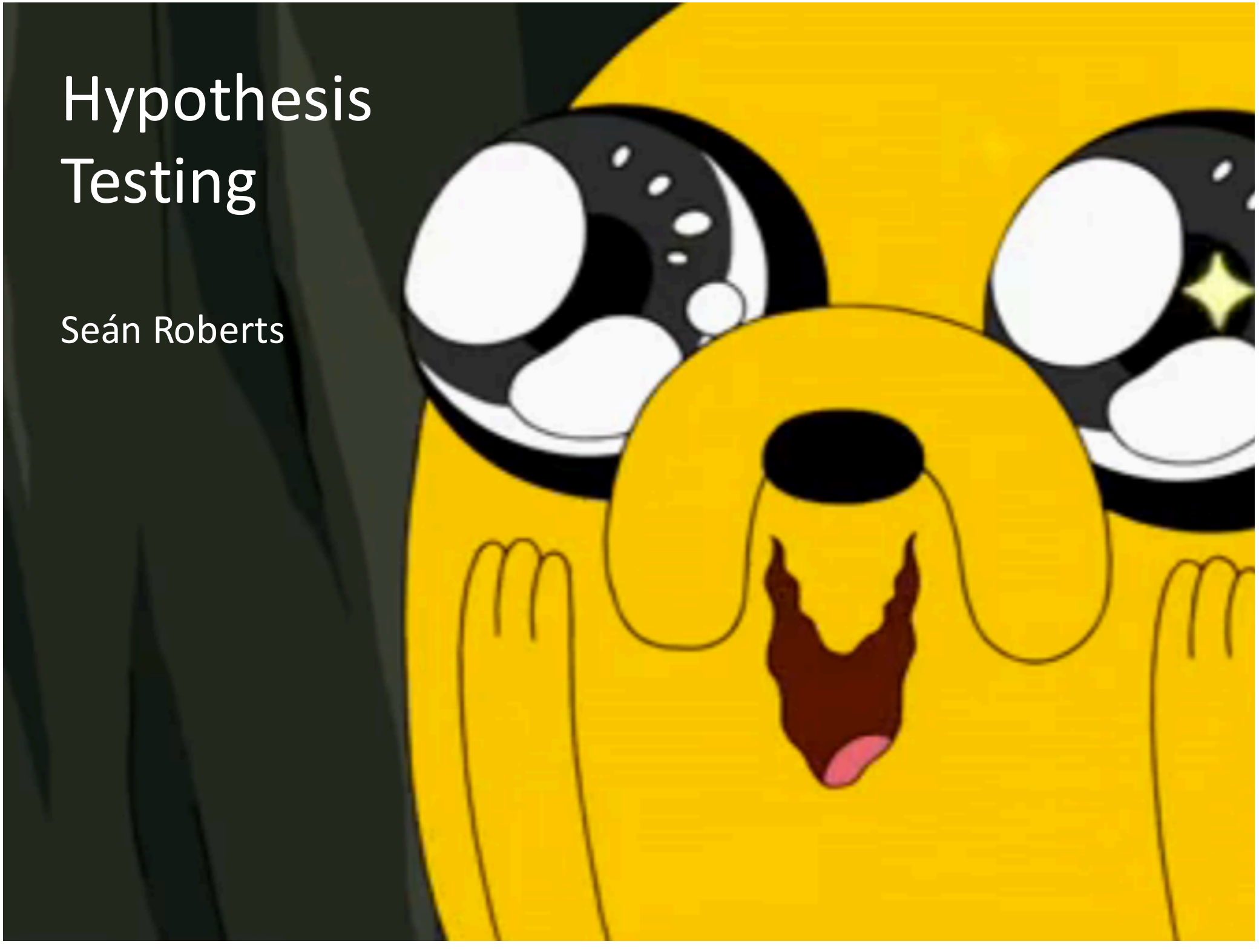
Hypothesis Testing

Seán Roberts



Overview

The strength of your argument relies on the strength of your null hypothesis

- Why use quantitative approaches?
  - Psychological Biases, Galton's problem
- How to formulate hypotheses
- How to test hypotheses:
  - P-values, null hypotheses
- How to create your own statistical tests
  - Permutation, Alternative baselines
- How to argue with data
- How to argue against data

Why Bother?

An extreme 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, which does not depend on our personal belief, and we can show to other people.

Produces results that are:

- Unbiased
- Transparent
- Reproducible
- Informative for theories

Leads to:

- More interesting answers
- More flexible analyses and discussion
- Less stress

Quantitative methods

Using numbers to help make your argument

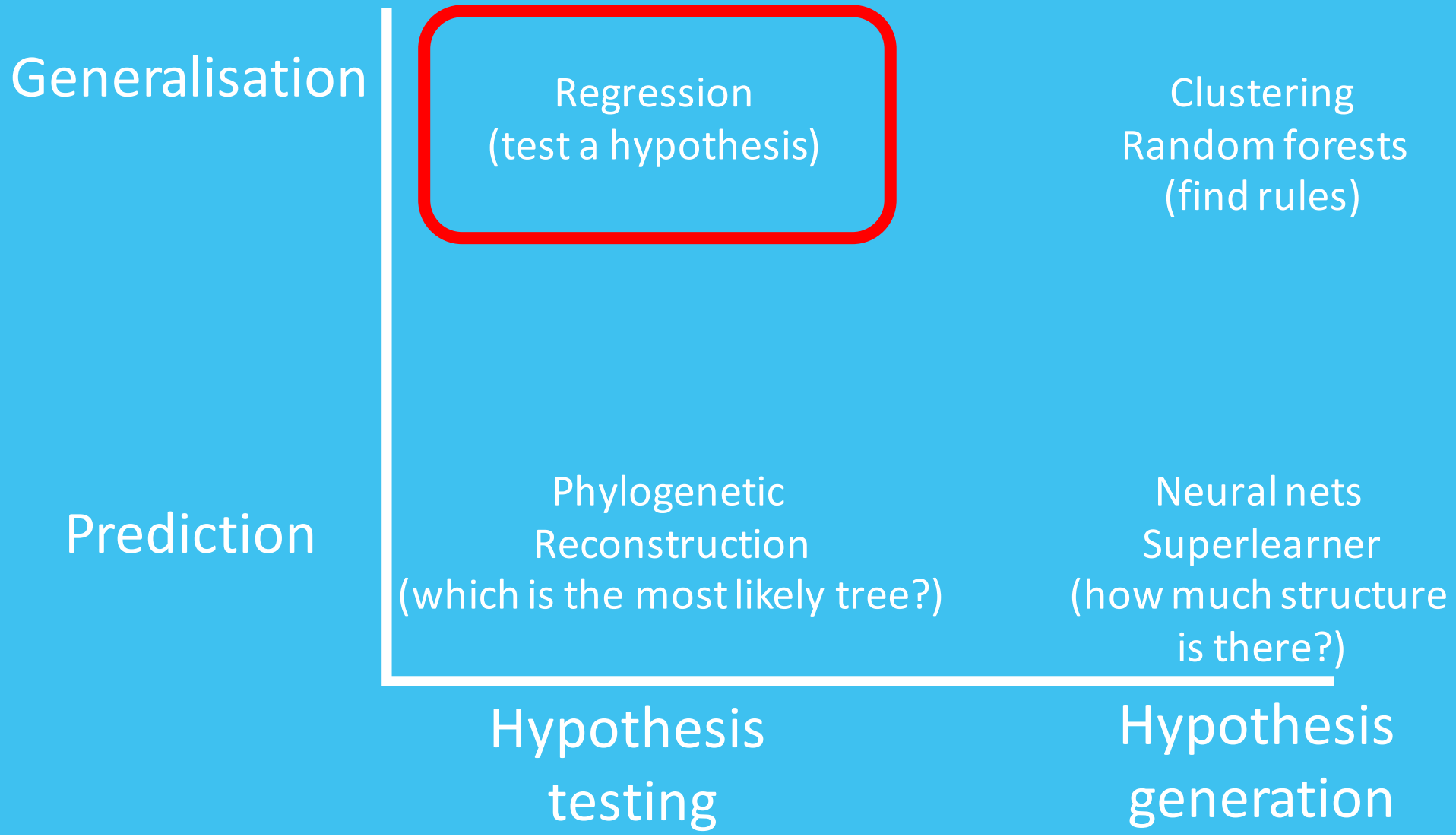
Works together with:

Qualitative analyses

Logical arguments

Theory building

Kinds of quantitative analyses



Why use quantitative methods?

Biases:

Clustering illusion

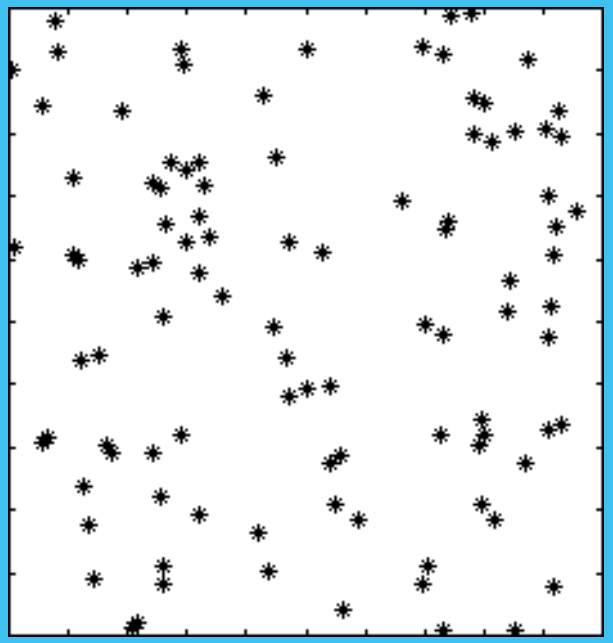
Base rate neglect

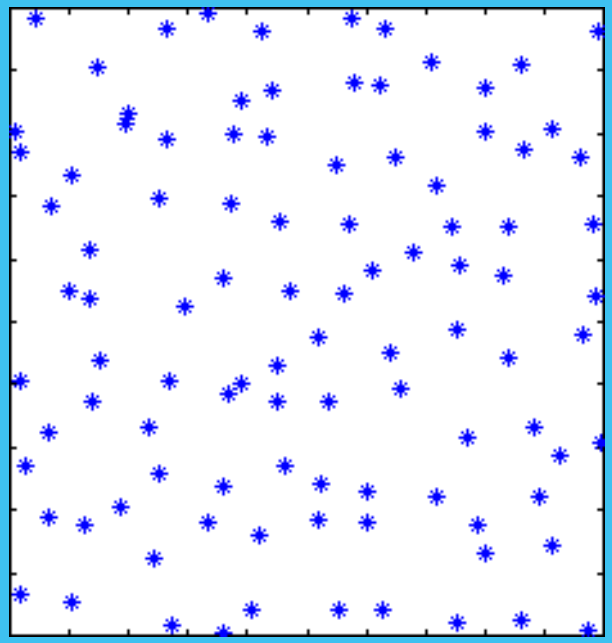
Galton's problem

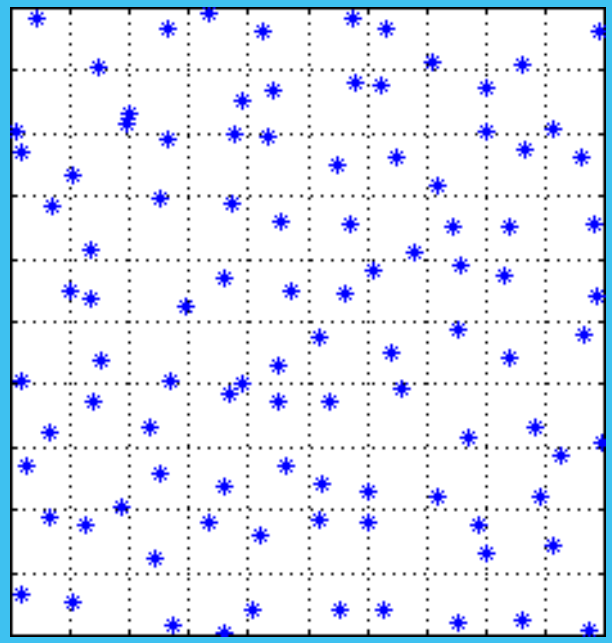
Confirmation bias

Hindsight bias

Illusion of validity

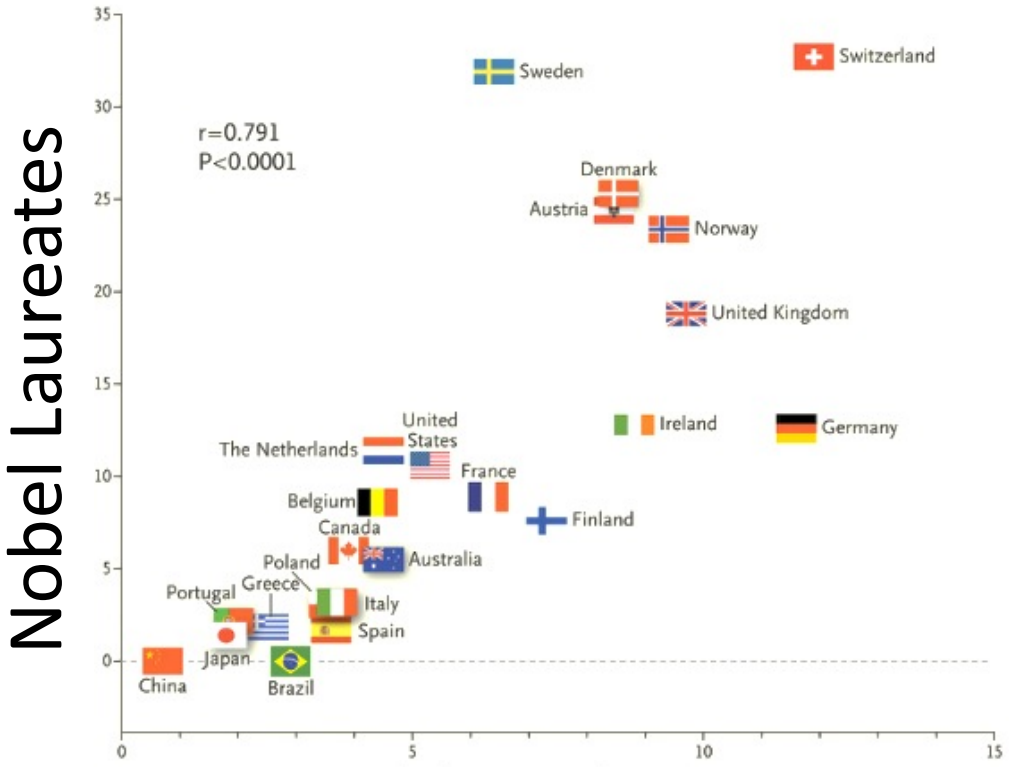




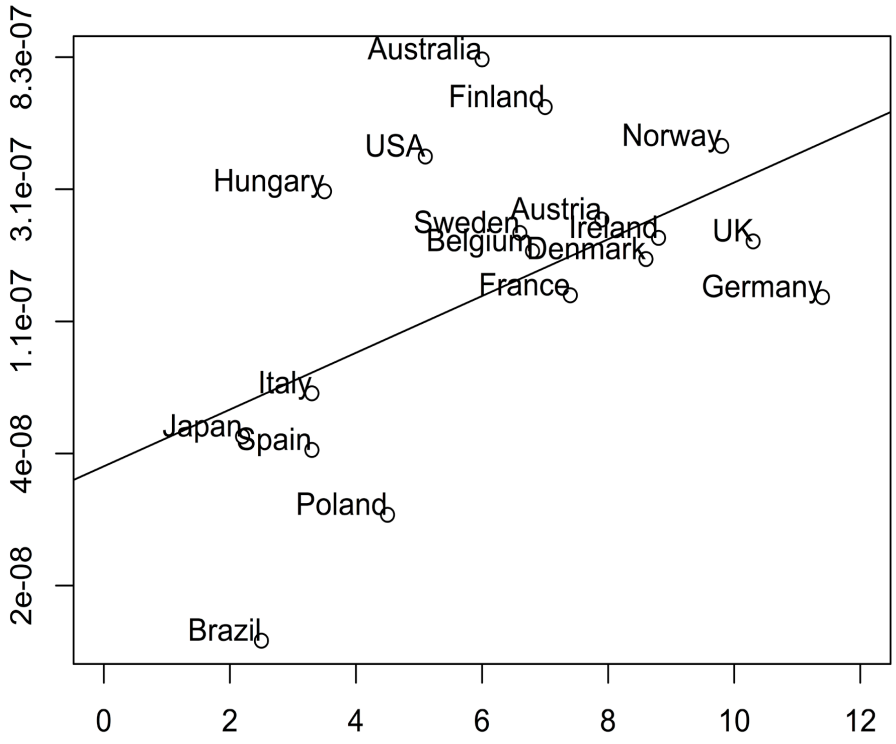


Illusion of validity

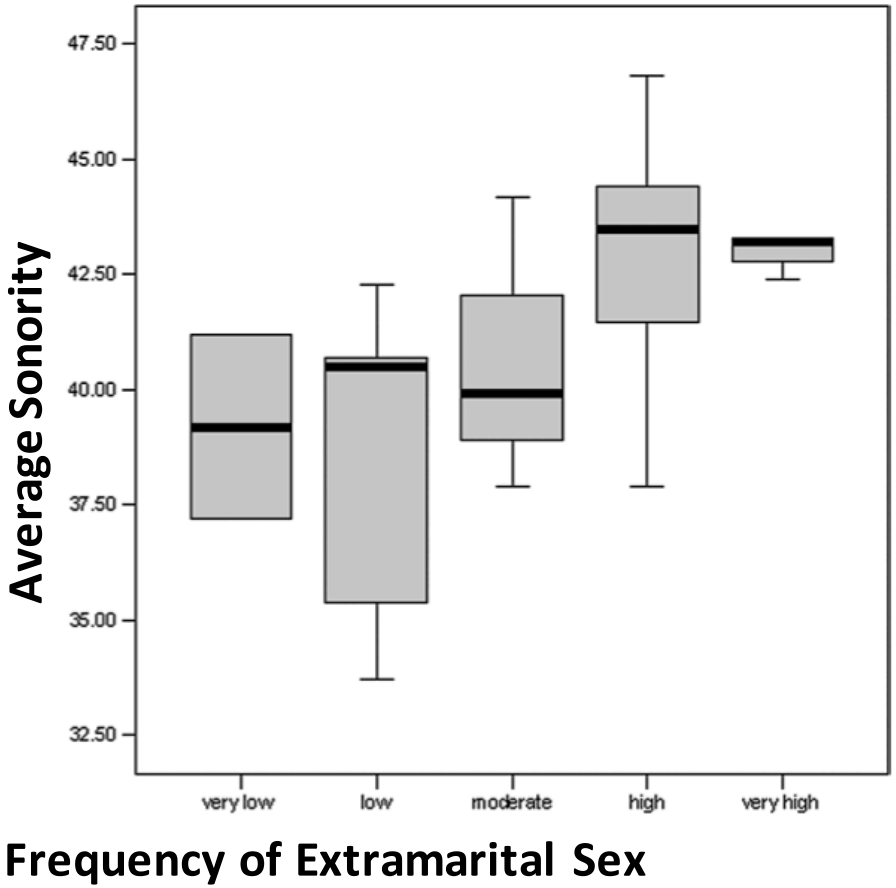
Messerli (2012)



Roberts & Winters (2013)

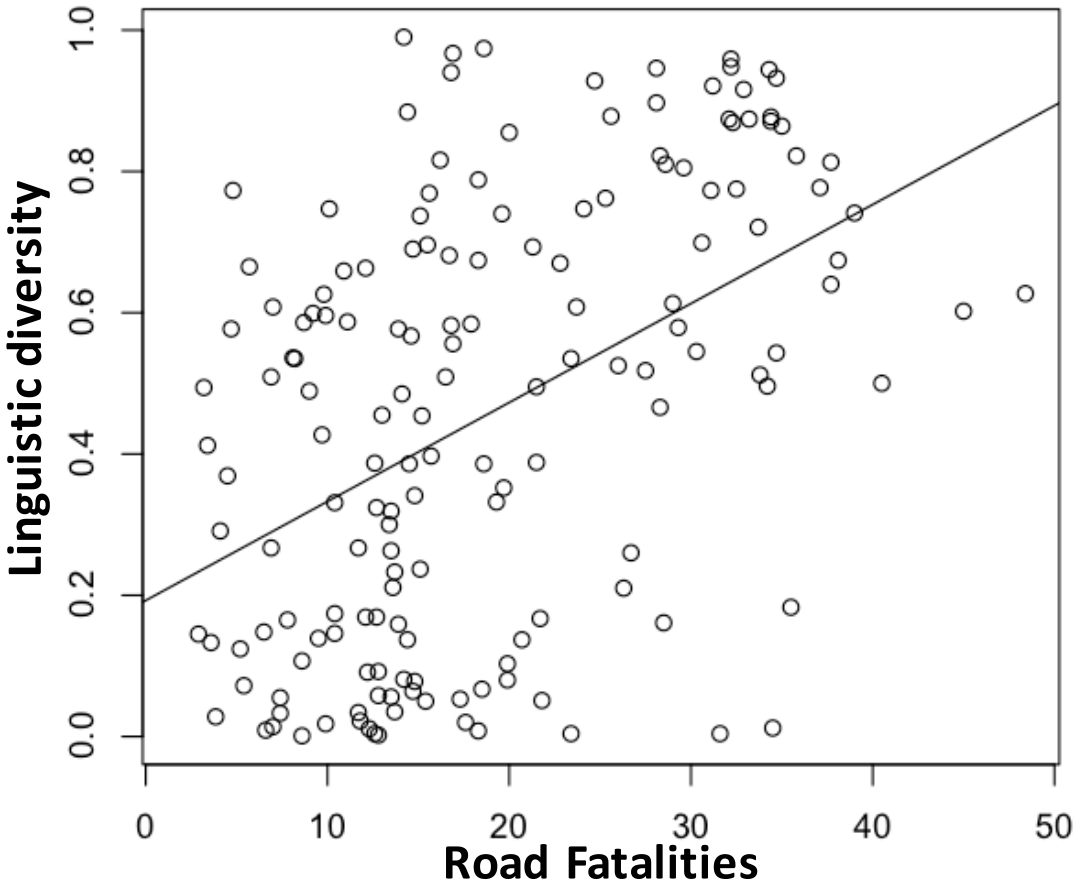


Hindsight bias



r = 0.51, p = 0.01

Ember & Ember (2007)

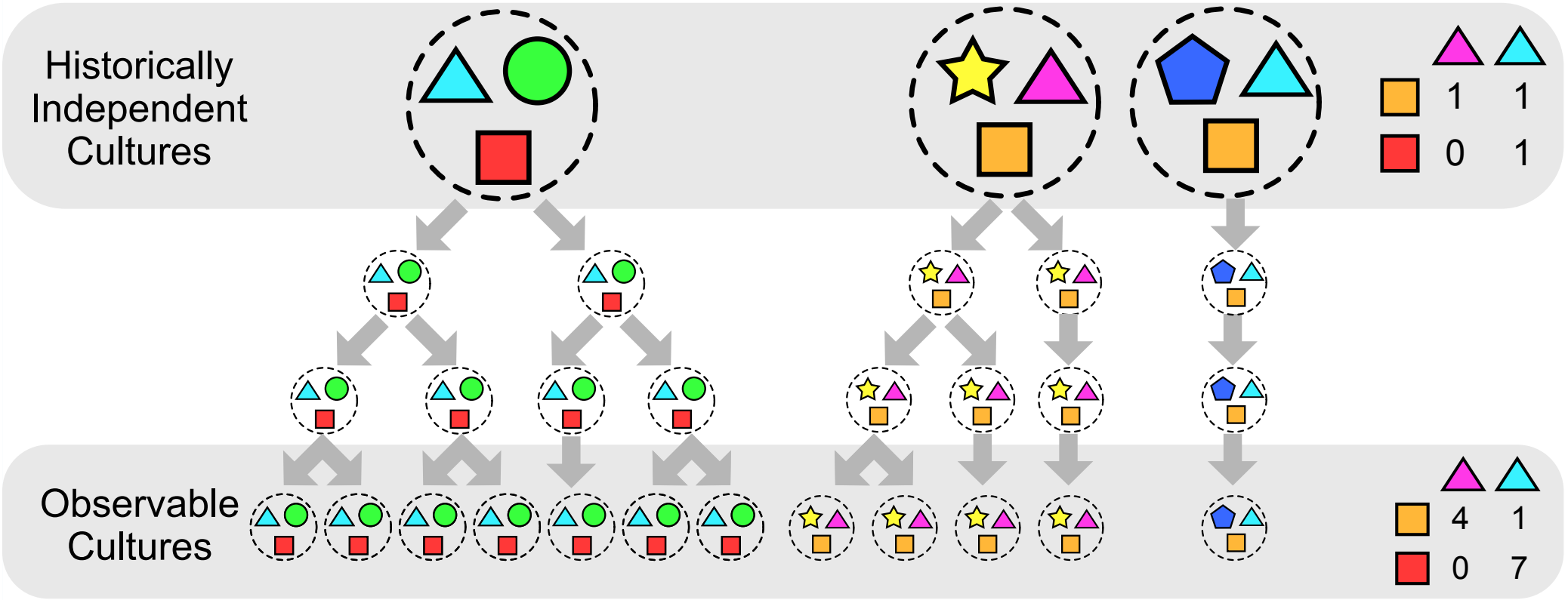


F (97,10) = 4.18, p<0.0001

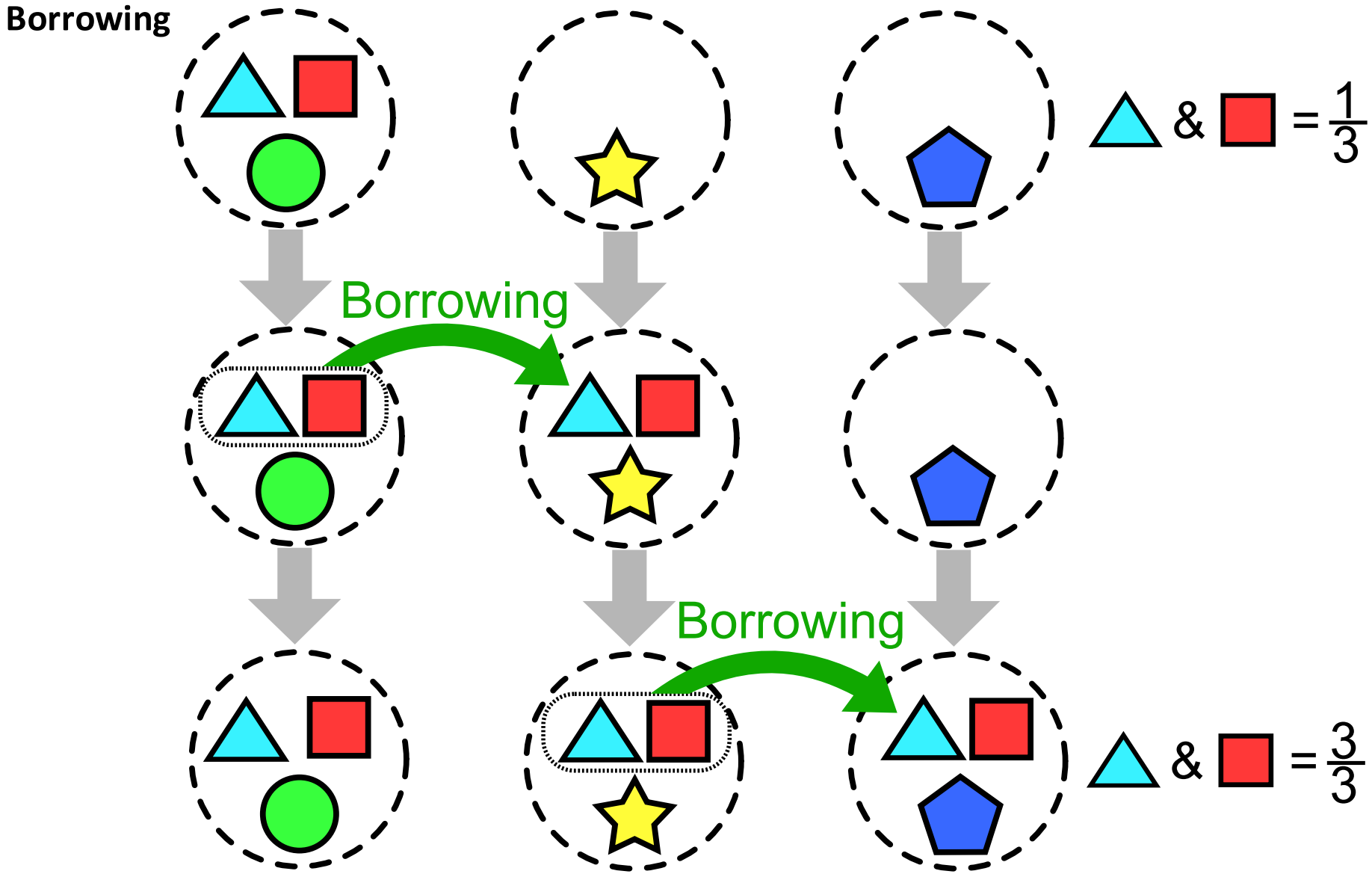
Controlling for:

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

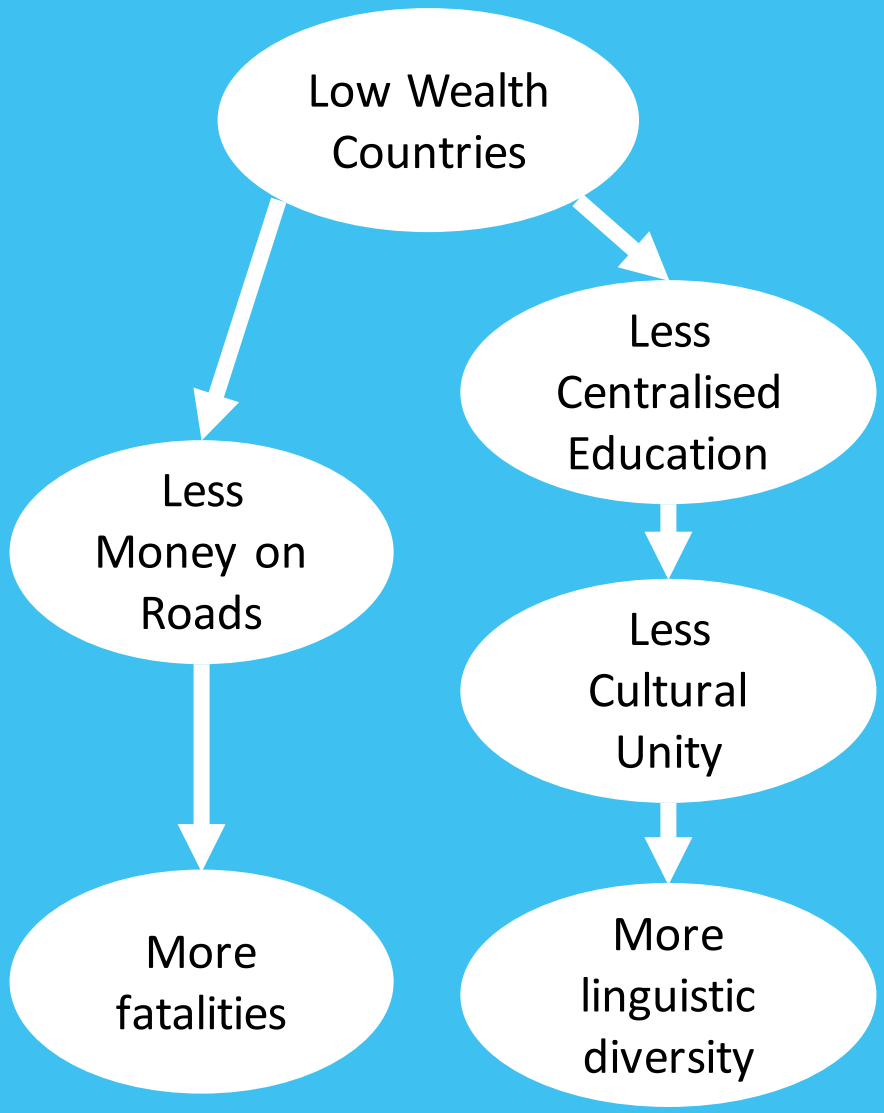
Historical relationships inflate correlations



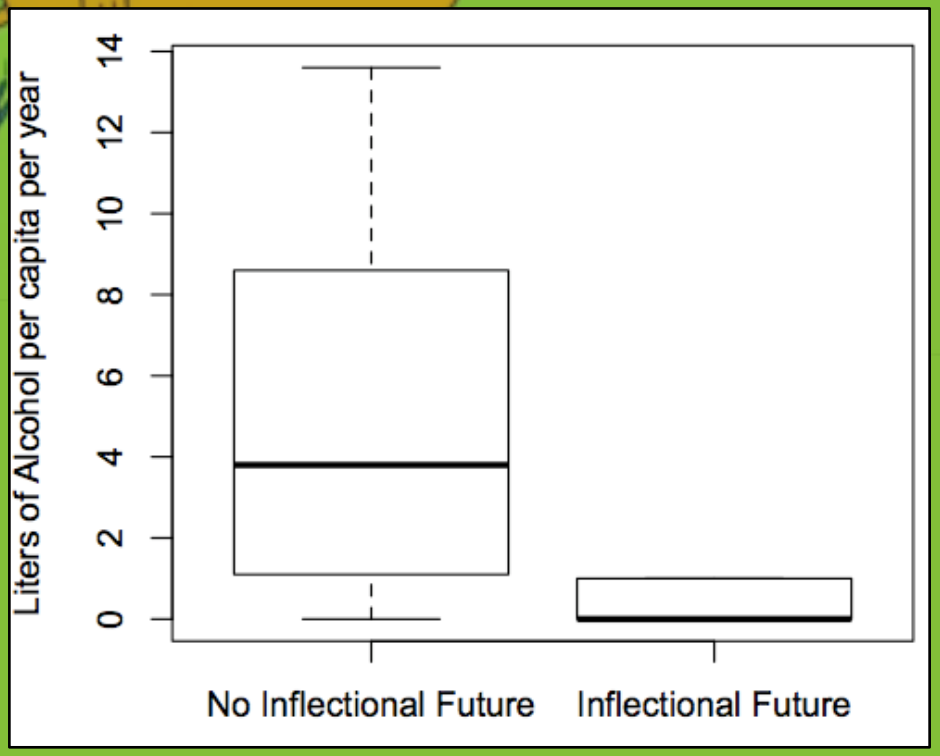
Correlations due to borrowing



Correlations from common causes



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

Solutions

Know your data: visualisation, data fluency

Statistical tests

Data with known relationships:

- Phylogenetic testing

Data with some unknown relationships:

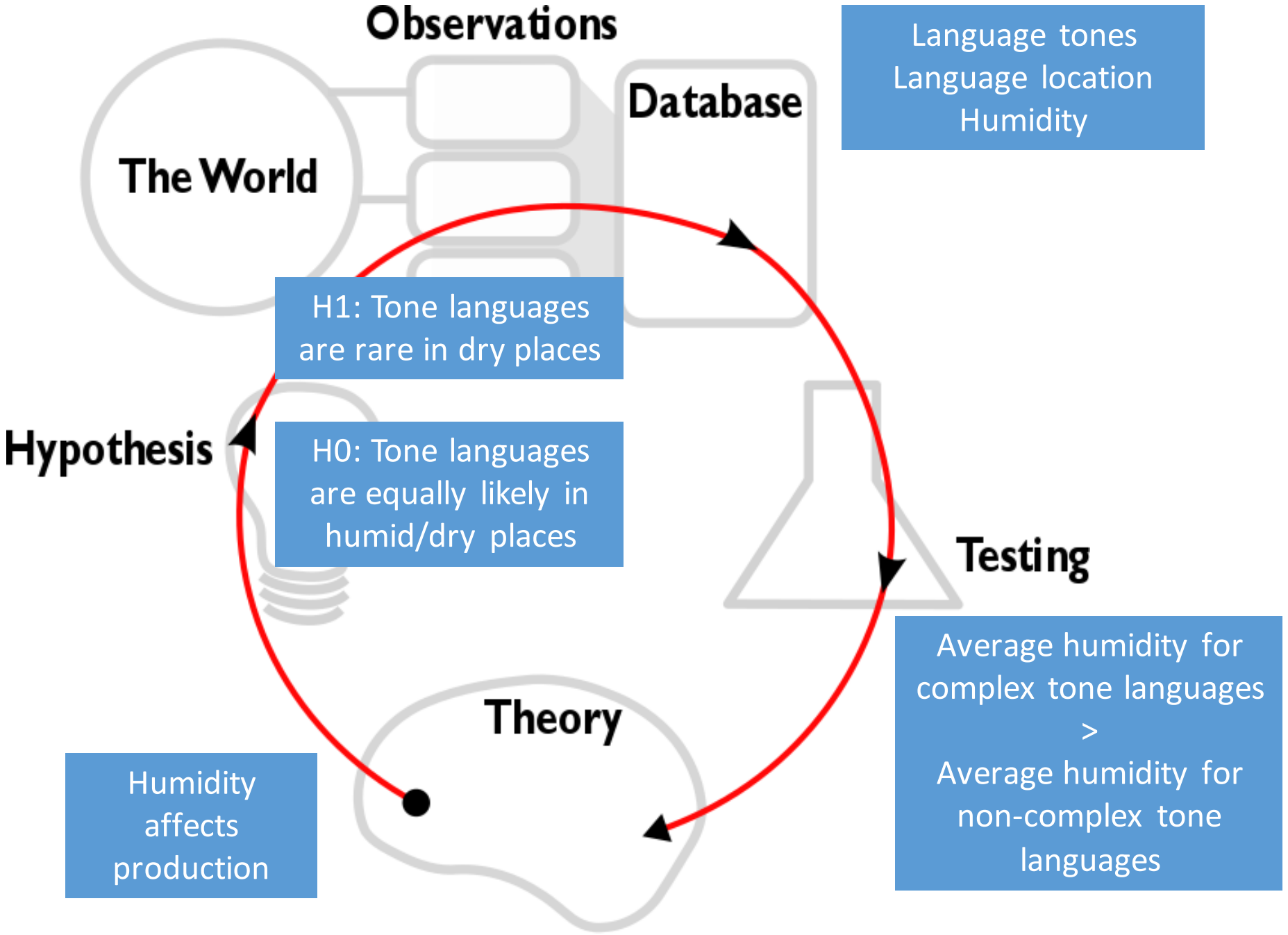
- Multilevel modelling

Dealing with confounds:

- Phylogenetic regression (PGLS)
- Random Forests with random effects

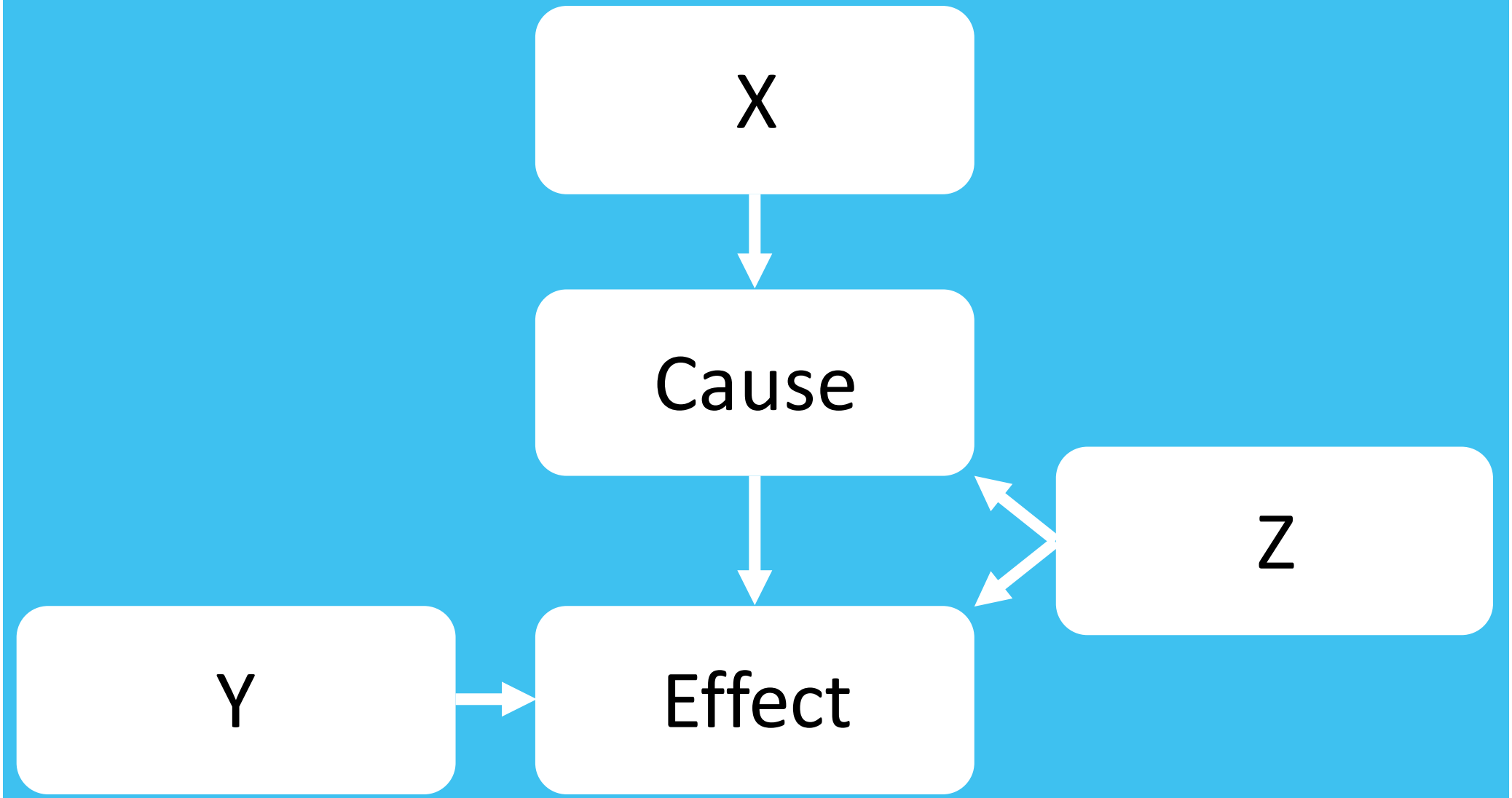
How to formulate hypotheses

Null Hypothesis Testing

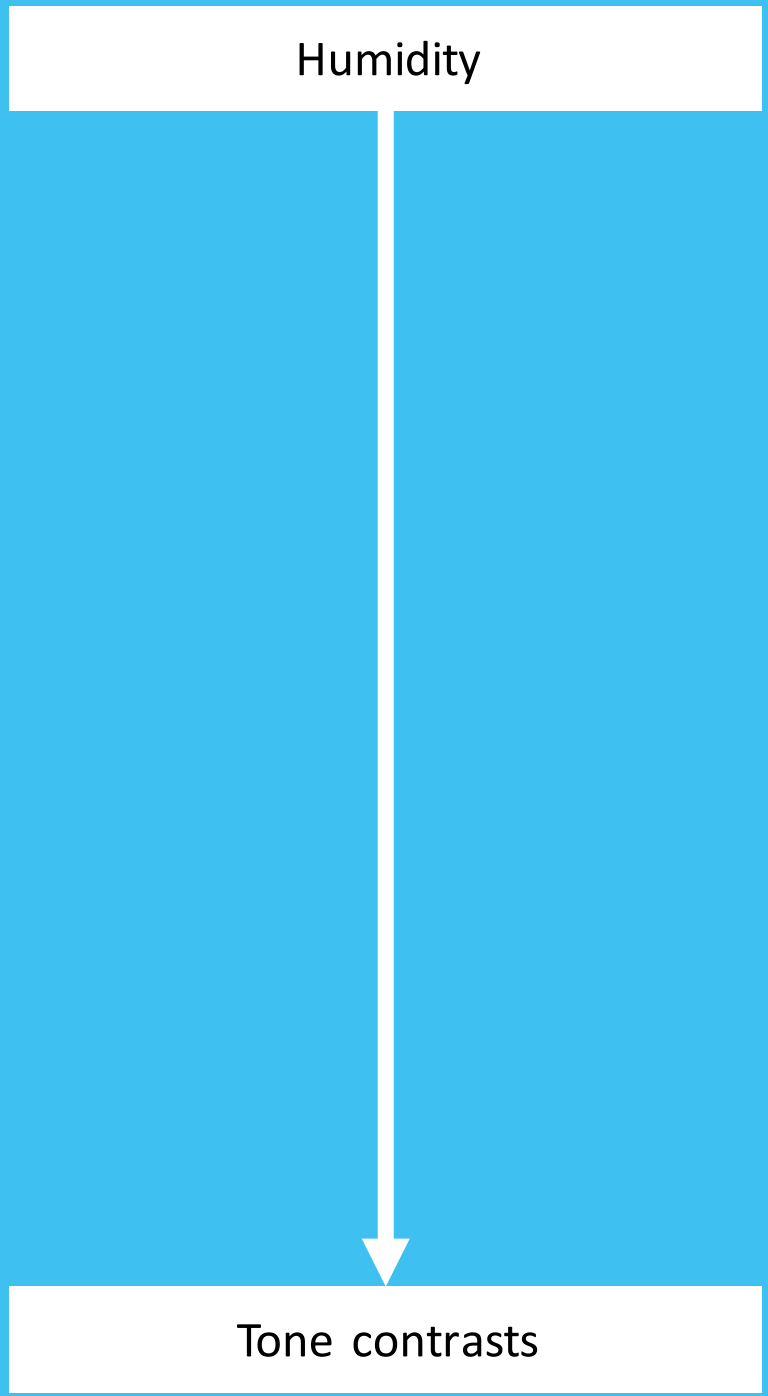


Formulating hypotheses:

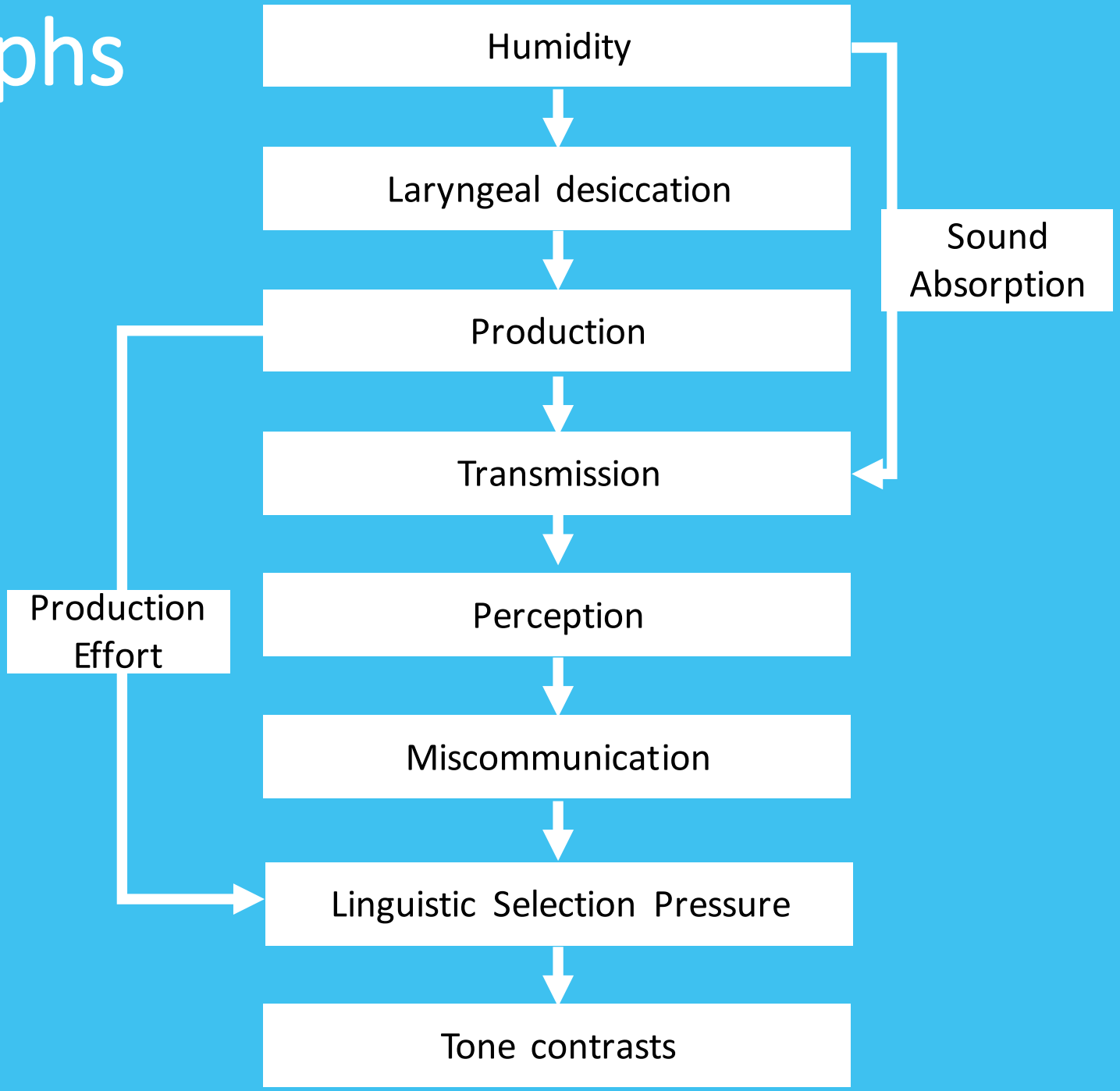
Causal inference approaches



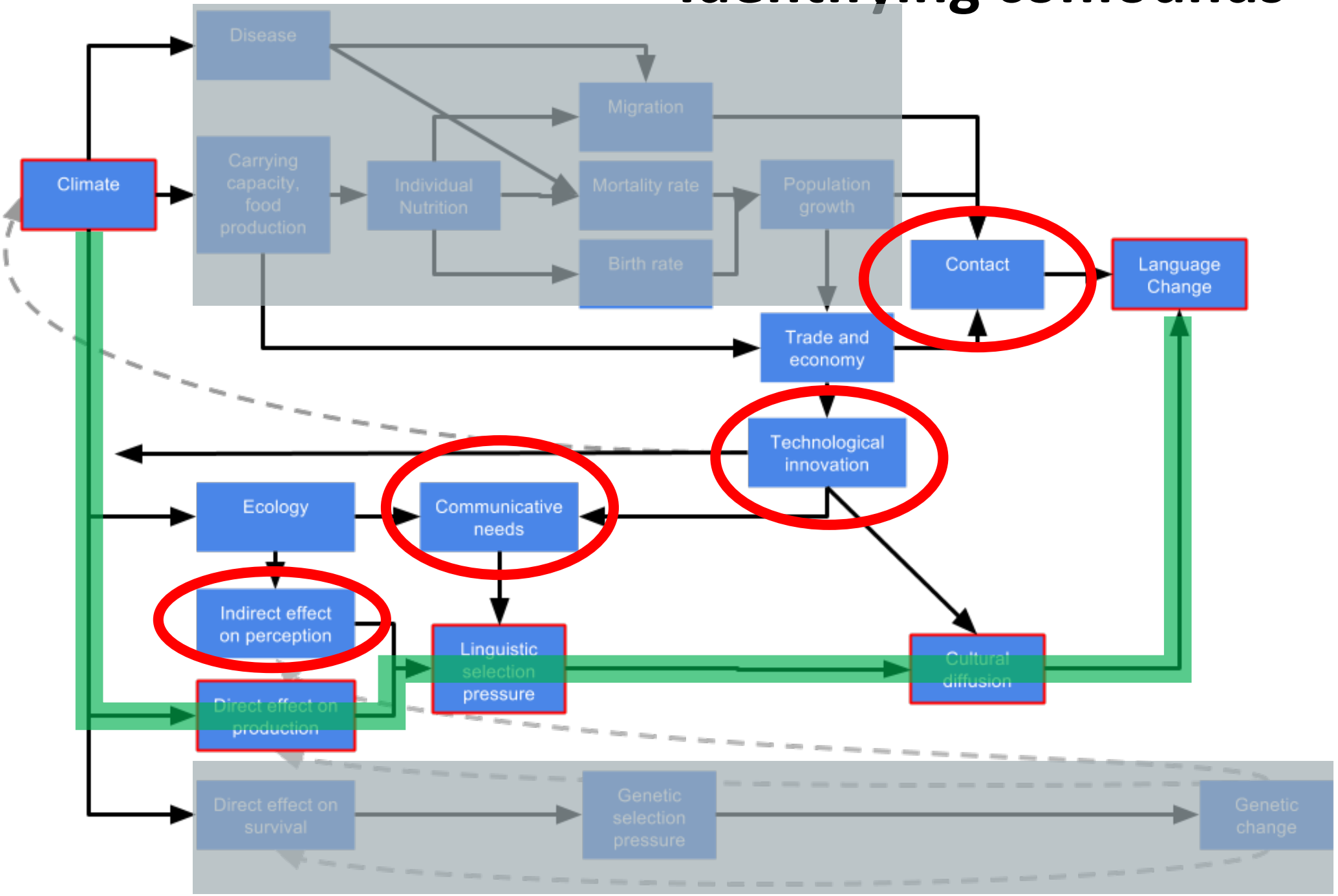
Causal graphs

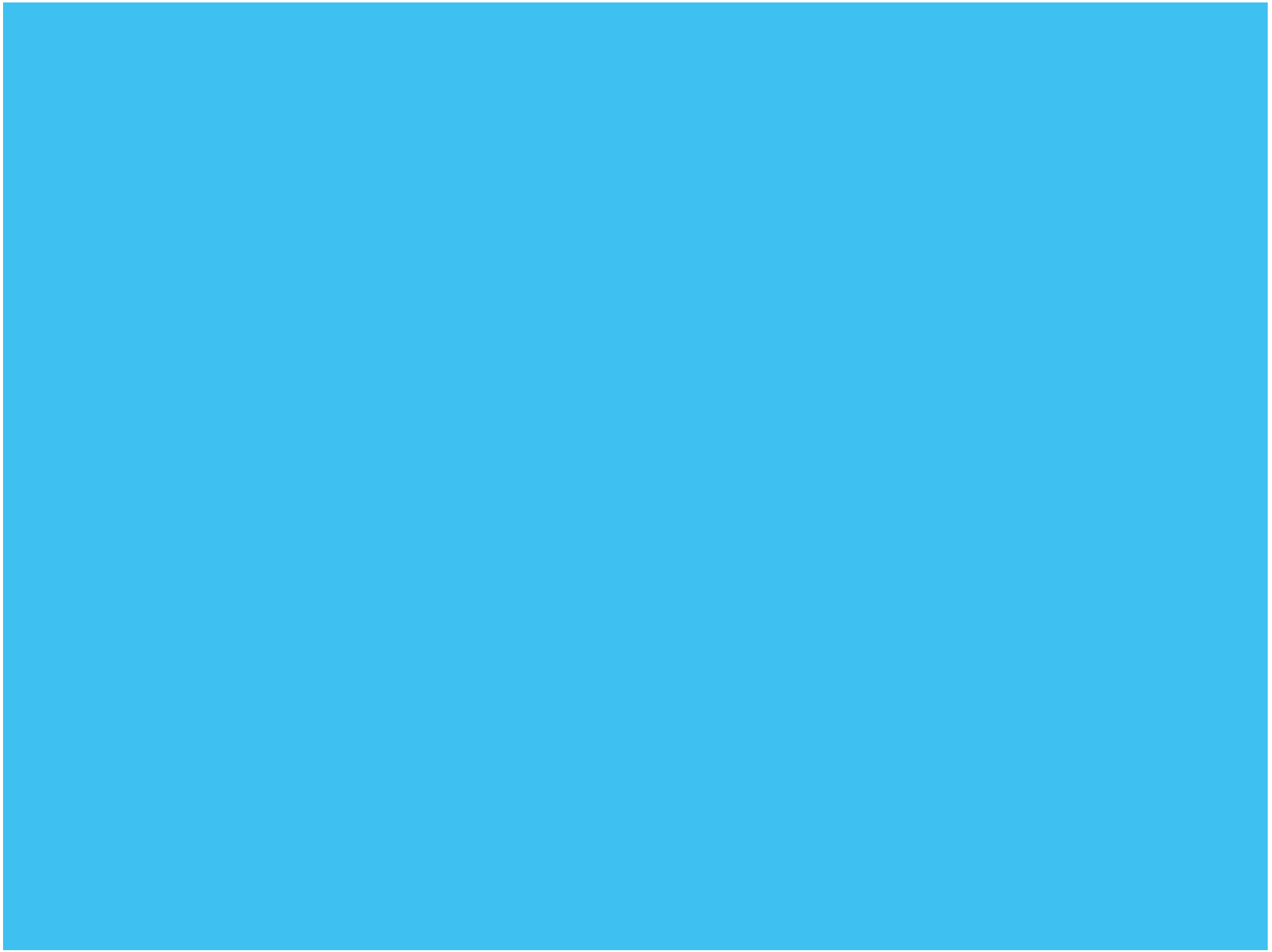


Causal graphs

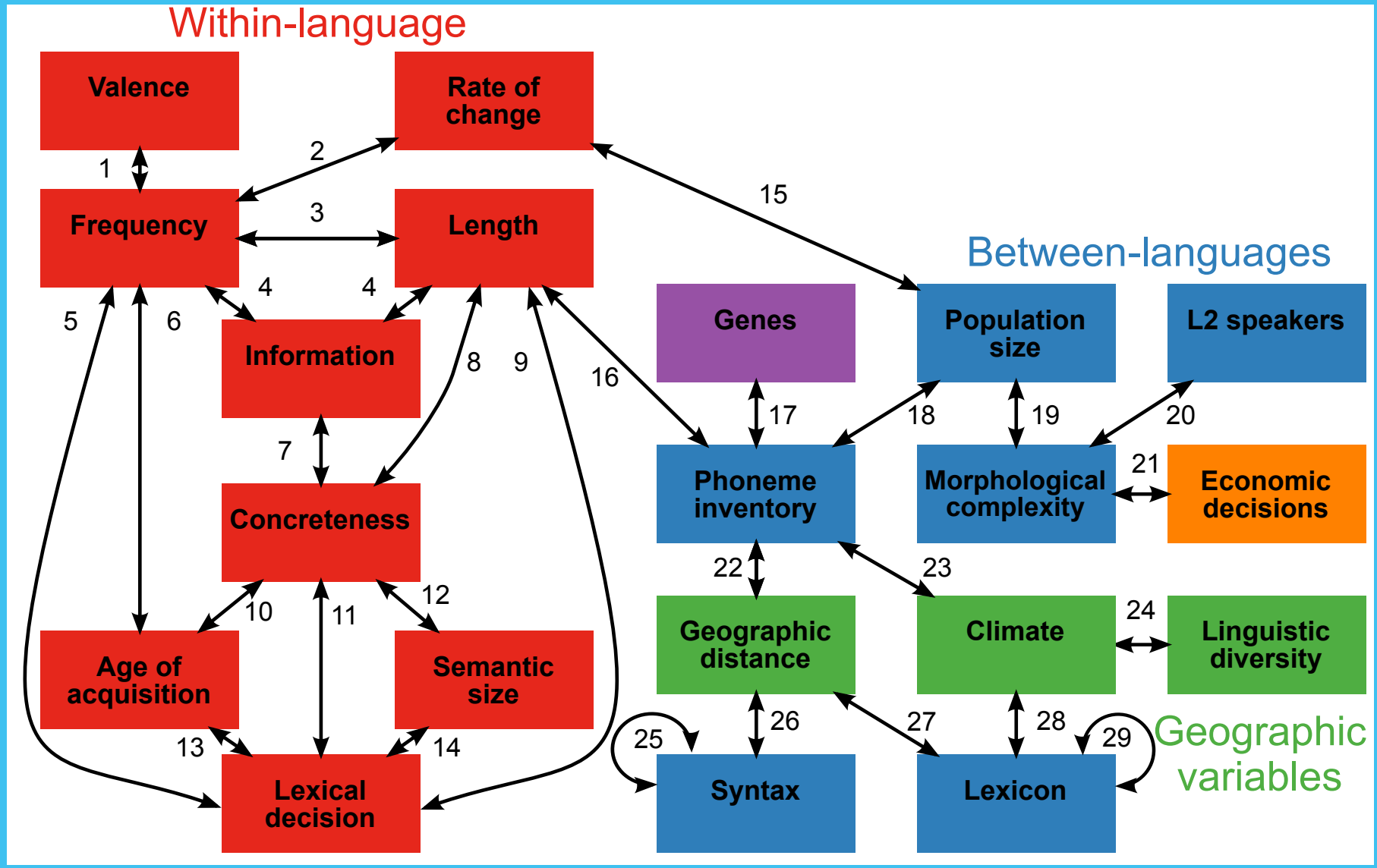


Identifying confounds

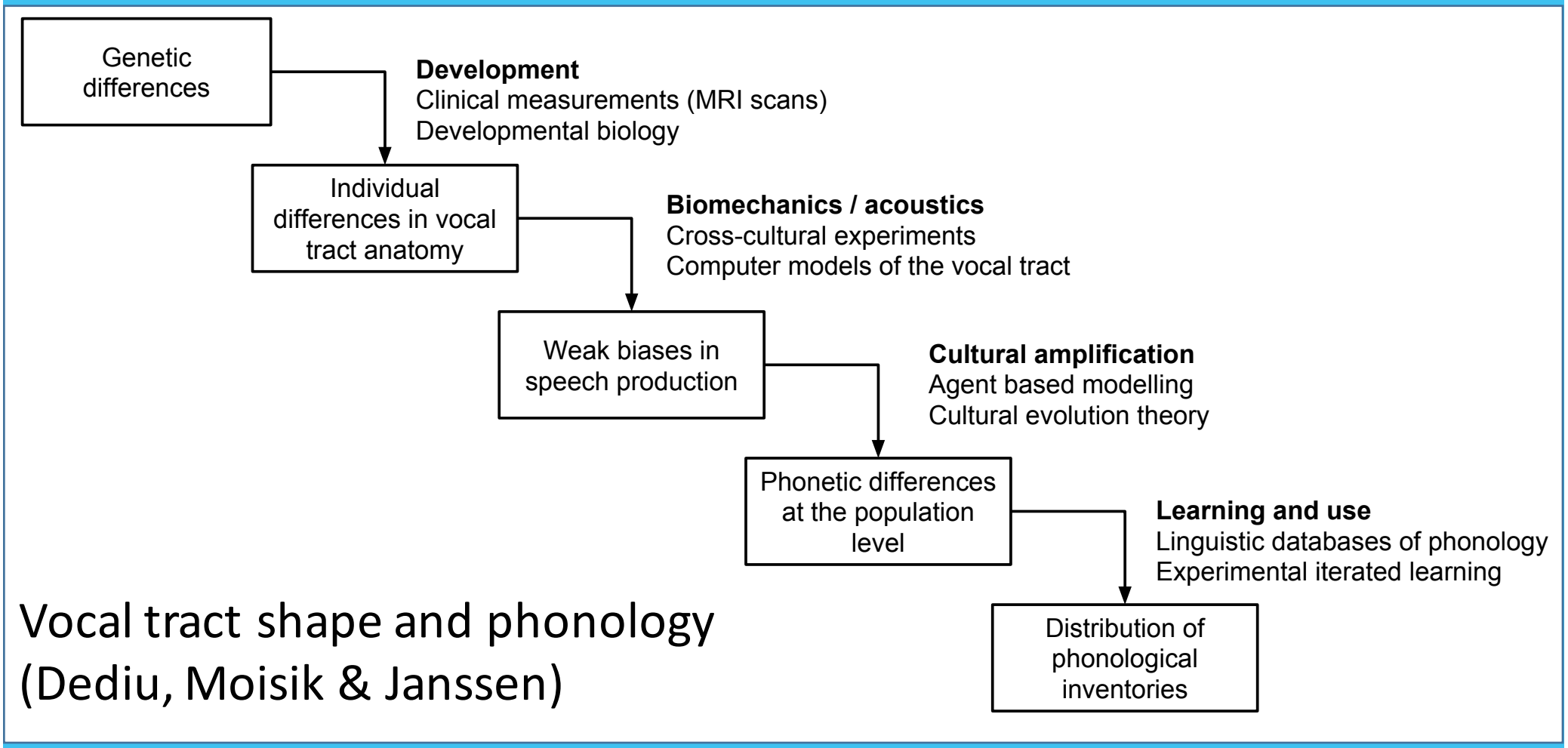




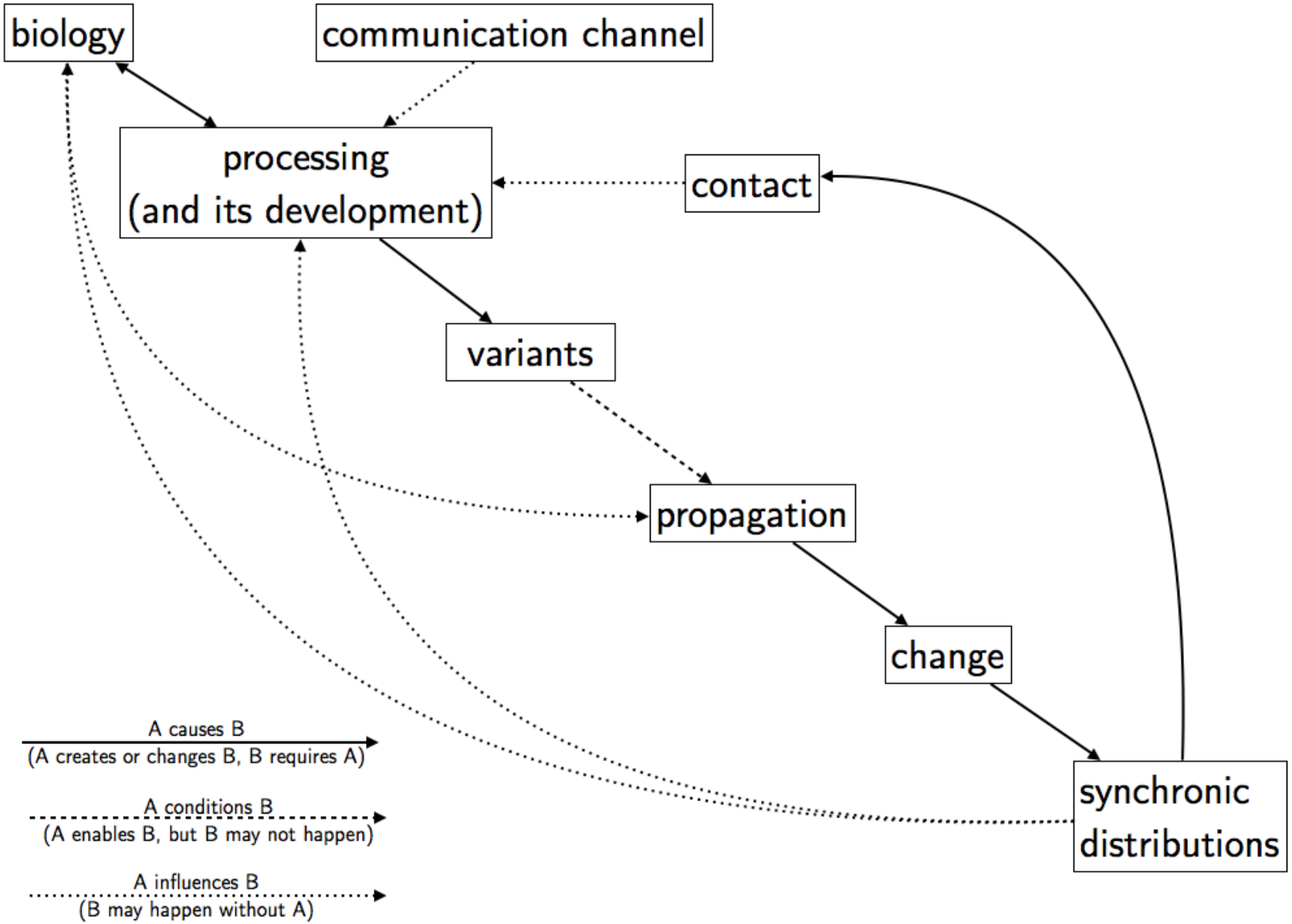
Identifying confounds



Breaking theories down



Thinking with causal chains



Causal thinking

Be explicit about cause and effect

Linguistic factors are part of a causal chain

Break the theory down into individual links

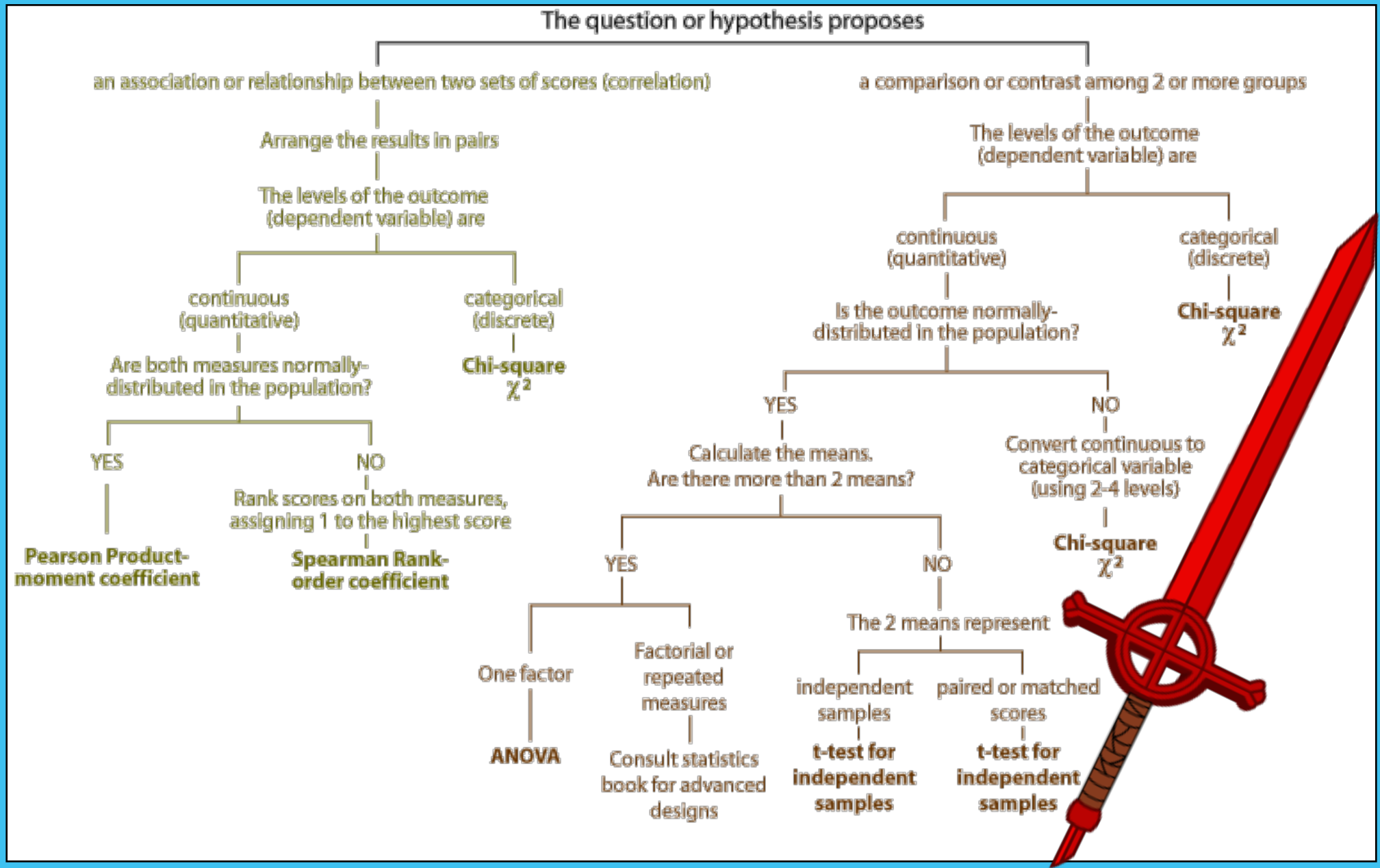
Look upstream to find explanations

Look downstream to find broader implications

Tackle each link separately

How to test hypotheses

Traditional approach



Null hypotheses and p values

The probability of finding an effect as extreme as the observed one if the null hypothesis is true.

92 heads in a row

Null hypothesis:

The coin is fair, each throw is independent

Probability of 92 heads in a fair, independent coin:

0.5^{92} (very small)

Null hypothesis can be rejected: coin is not fair

But there are still many alternative hypotheses:

Coin is weighted or *throws are not independent* or *time has stopped*

Assumptions of the null hypothesis are important

P value: The probability of finding an effect as extreme as the observed one if the null hypothesis is true.

If you want to show that your affect is **not** caused by:

- Borrowing
- Contact
- Population size
- Level of industrialisation

… then your null hypothesis should reflect this.

Null Hypothesis testing with regression

Null: Traffic accidents ~ Population density +
GDP of country +
Spending on transport

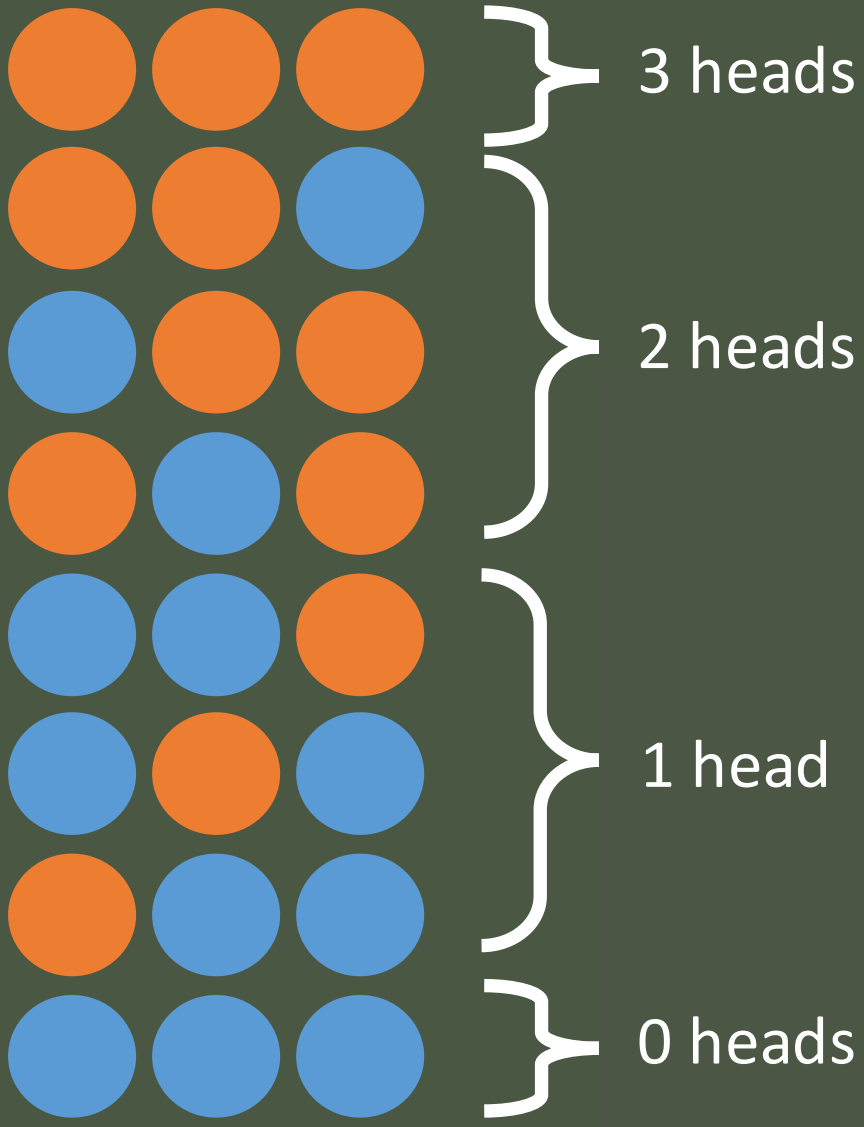
H1: Traffic accidents ~ Population density +
GDP of country +
Spending on transport +
Linguistic diversity

Advanced methods:
Multilevel modelling, Bayesian statistics

3 heads in a row

Assume: the coin is fair, each throw is independent

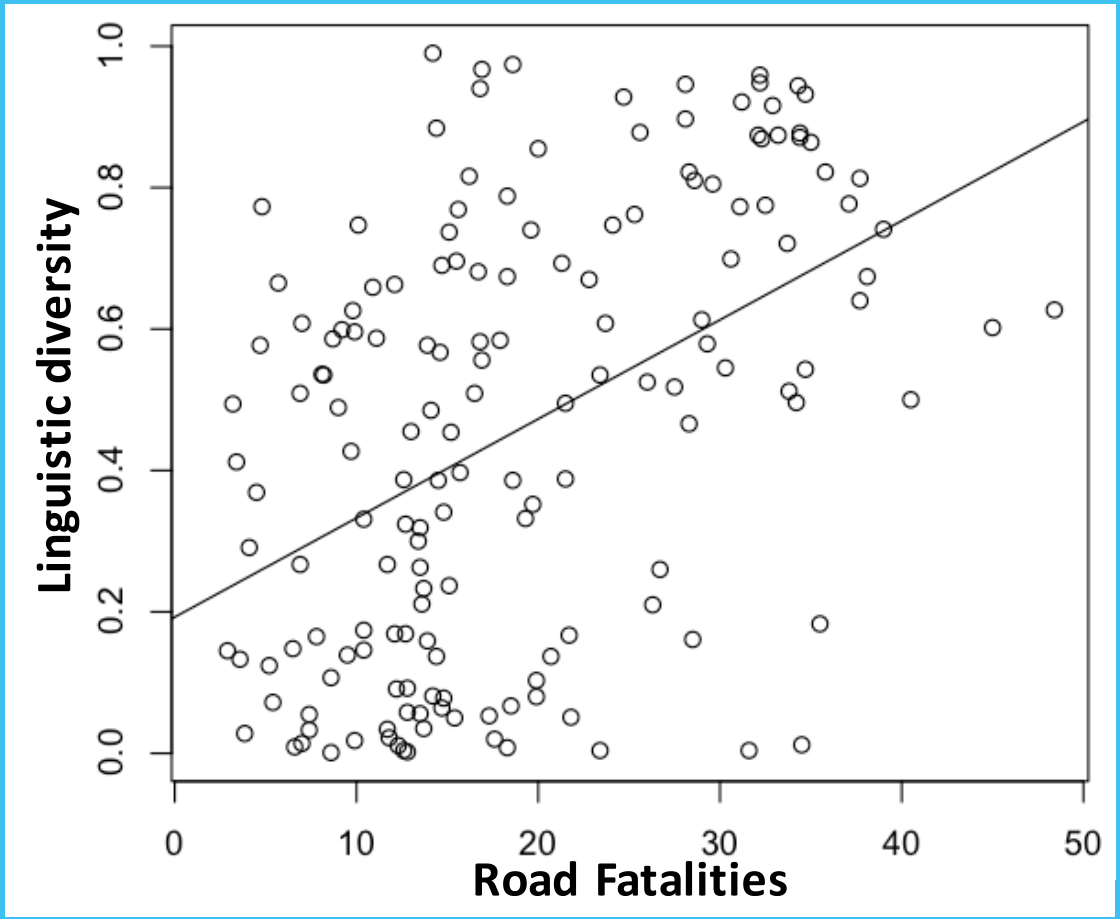
Generate all possibilities:



$$\frac{1}{8}$$

Null hypotheses for the real world

What is the probability of generating this correlation under the null hypothesis?



What are reasonable assumptions?

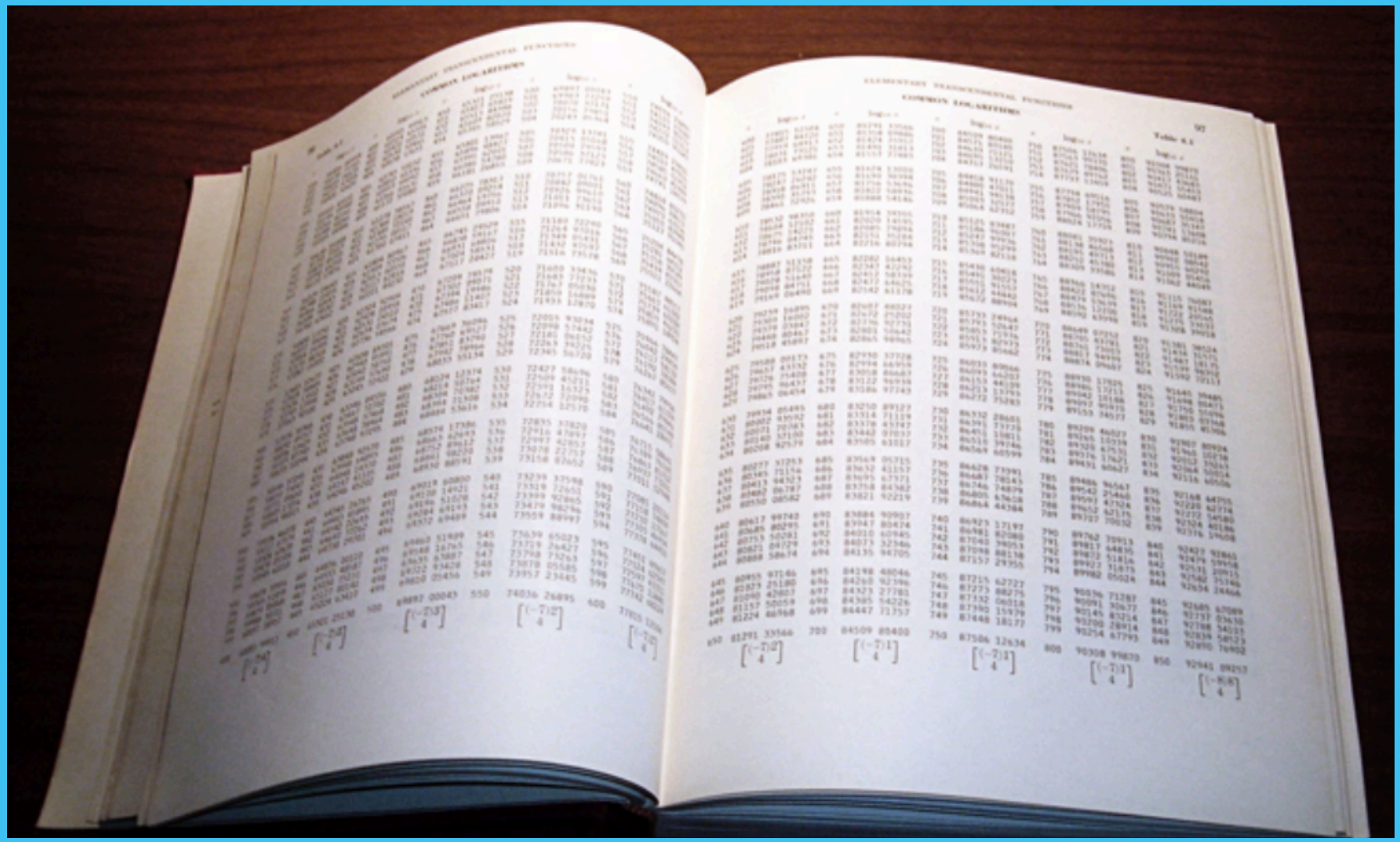
Enumerate all possibilities?
No way!

Traditional approach:
Assume variable is drawn from a random distribution

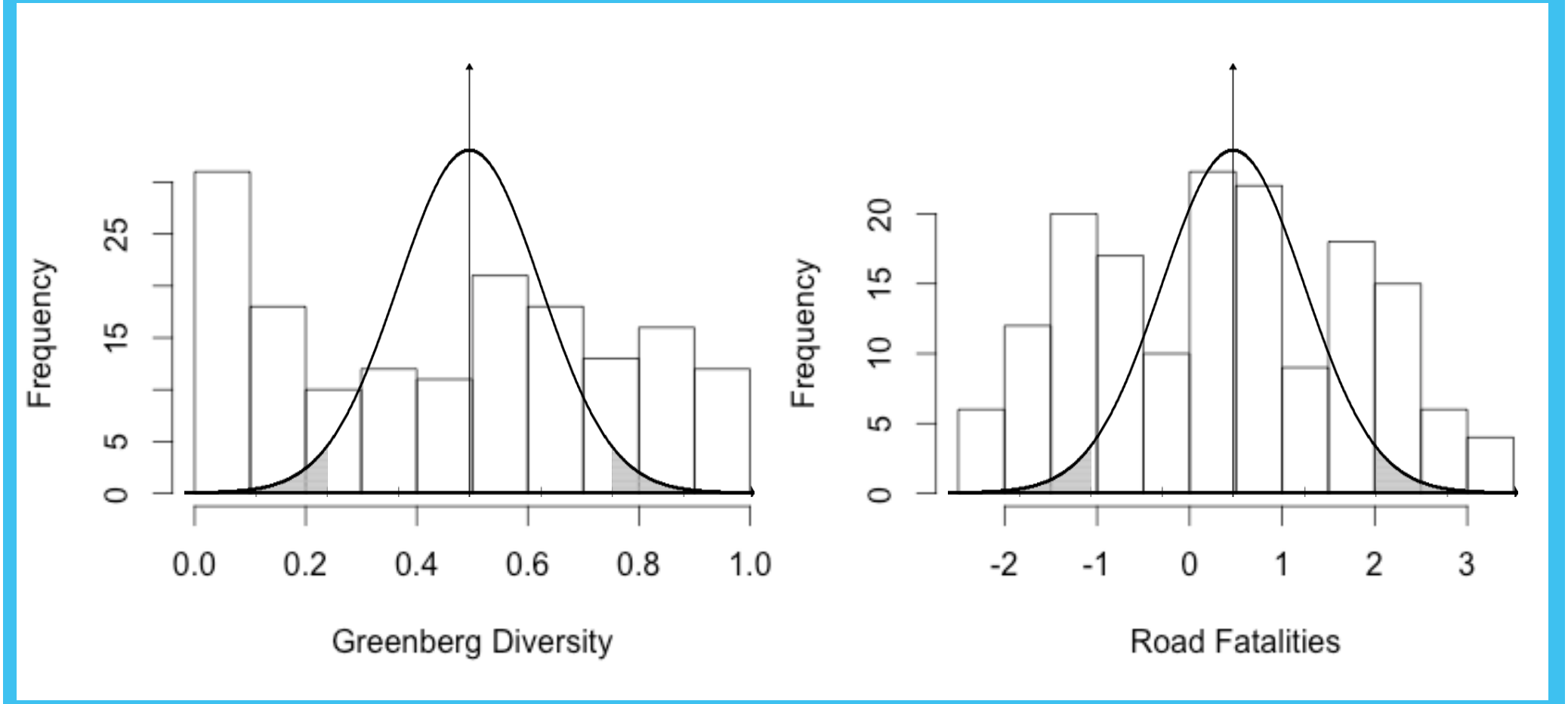
Null hypothesis: variables are independent, drawn from normal distributions.

Assumption: data is normal

Excuse: Necessity



Null hypotheses for the real world



Null hypotheses for the real world

You don’t need to rely on standard tests.

The most important thing:

Produce the best null hypothesis for your data

You can make your own tests:

- Permutation
- Bootstrapping
- Simulation

Permutation

Bootstrap a level of chance from the data

Assumptions: Very few

Advantages

Flexible, adaptable to particular questions

Any measures can be used

More details:

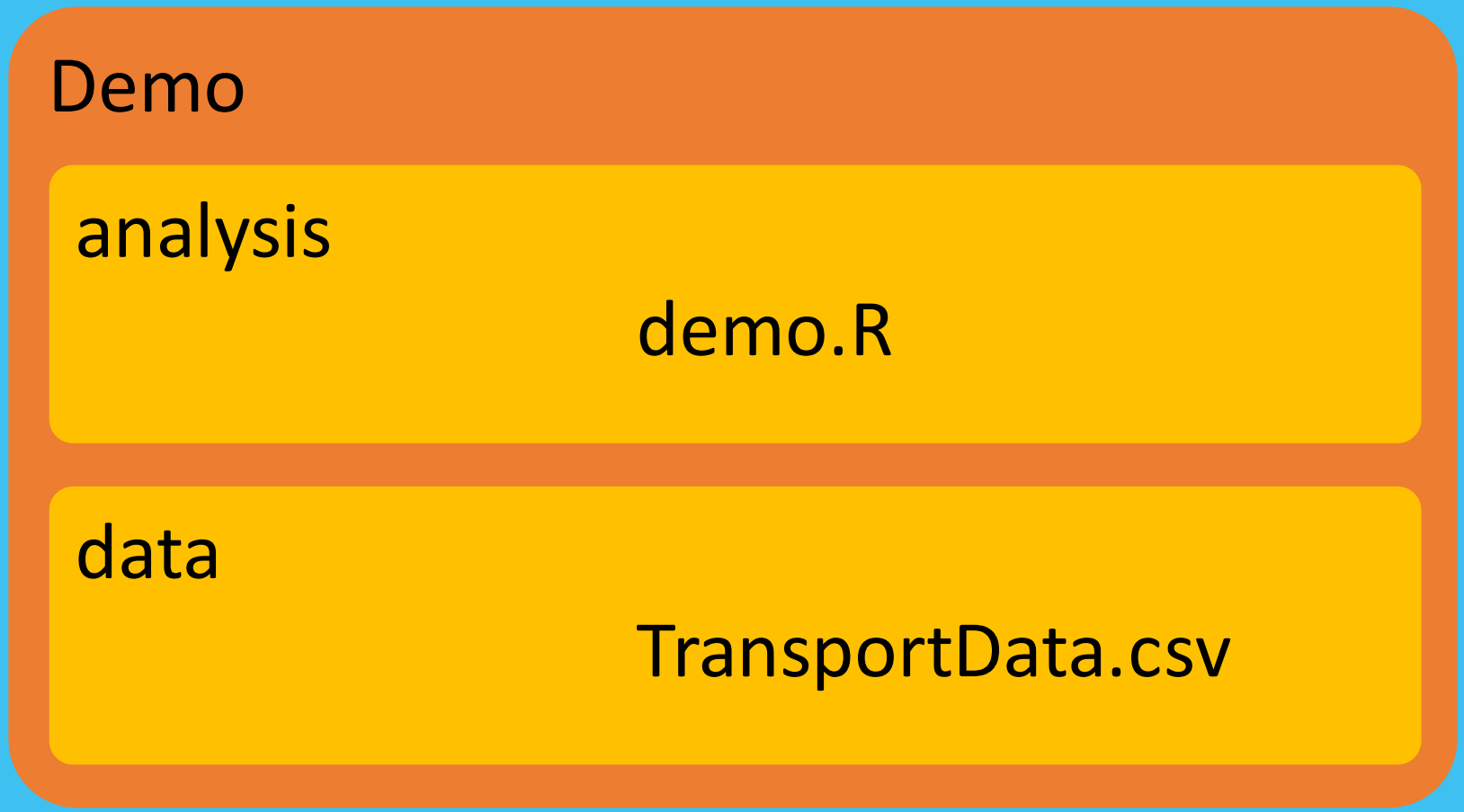
'Permutation Tests' tutorial in the QMSS Intro to R

Demo

Shared Google folder: Data/TransportData.csv

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

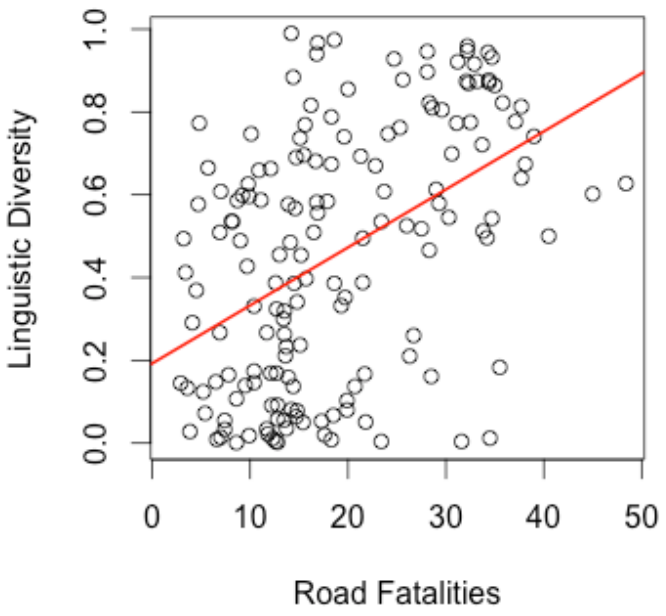
Web: http://tinyurl.com/qmssData1



Permutation

Country	Road Fatalities	Linguistic Diversity	Continent
Bahrain	12.1	0.663	Asia
Bangladesh	12.6	0.387	Asia
Armenia	13.9	0.159	Europe
Austria	8.2	0.535	Europe
Azerbaijan	13	0.455	Europe
Belarus	15.7	0.397	Europe
Belgium	10.1	0.747	Europe
Bahamas	14.5	0.386	North America
Barbados	12.2	0.091	North America
Belize	15.6	0.769	North America
Australia	5.2	0.124	Oceania

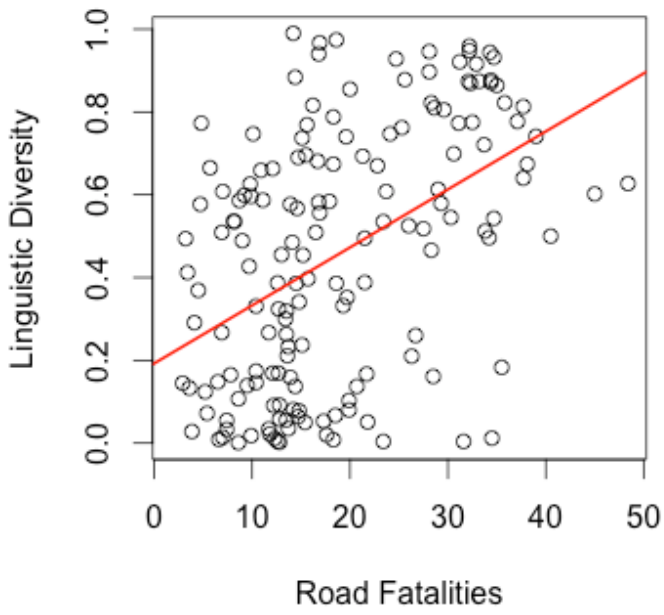
True Data



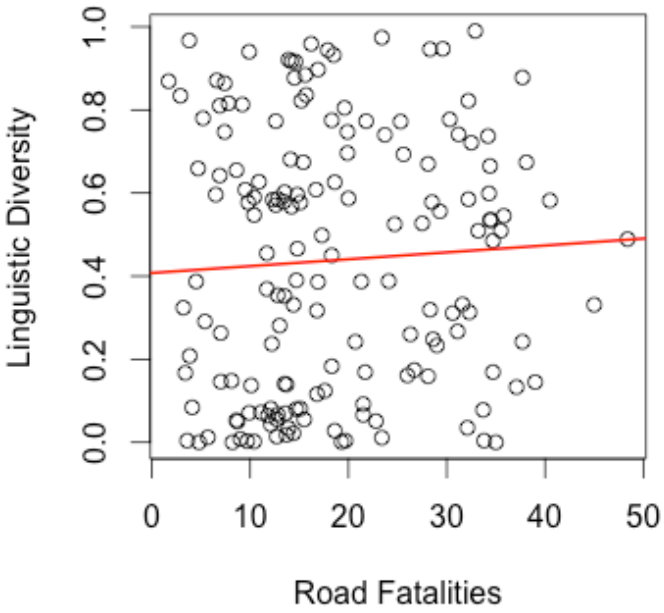
Permutation

Country	Road Fatalities	Linguistic Diversity	Continent
Bahrain	12.1	0.535	Asia
Bangladesh	12.6	0.386	Asia
Armenia	13.9	0.159	Europe
Austria	8.2	0.663	Europe
Azerbaijan	13	0.769	Europe
Belarus	15.7	0.397	Europe
Belgium	10.1	0.747	Europe
Bahamas	14.5	0.387	North America
Barbados	12.2	0.124	North America
Belize	15.6	0.455	North America
Australia	5.2	0.091	Oceania

True Data



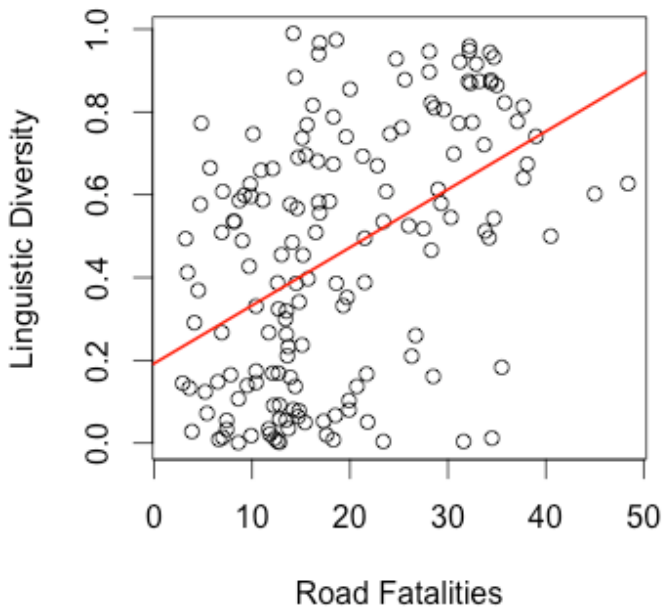
Permuted Data



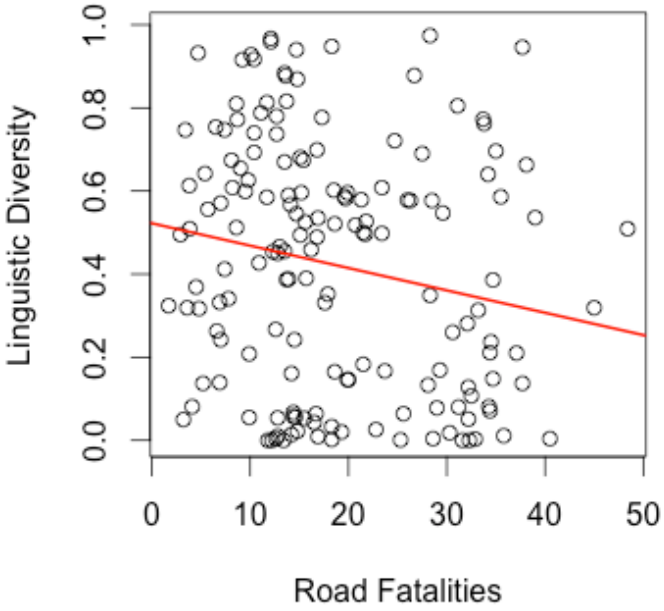
Permutation

Country	Road Fatalities	Linguistic Diversity	Continent
Bahrain	12.1	0.387	Asia
Bangladesh	12.6	0.663	Asia
Armenia	13.9	0.159	Europe
Austria	8.2	0.091	Europe
Azerbaijan	13	0.455	Europe
Belarus	15.7	0.397	Europe
Belgium	10.1	0.386	Europe
Bahamas	14.5	0.747	North America
Barbados	12.2	0.535	North America
Belize	15.6	0.124	North America
Australia	5.2	0.769	Oceania

True Data



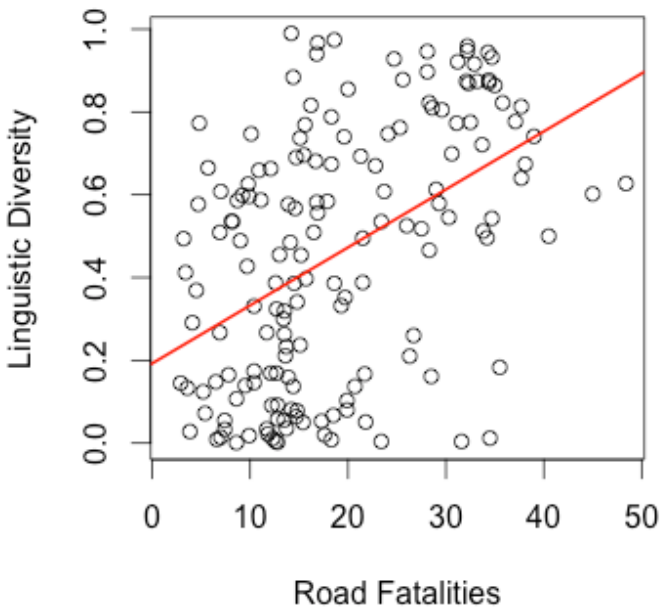
Permuted Data



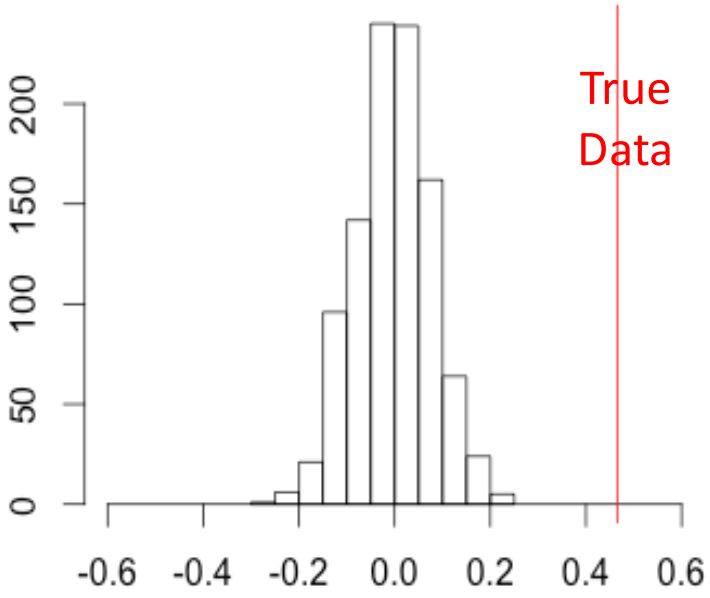
Permutation

Country	Road Fatalities	Linguistic Diversity	Continent
Bahrain	12.1	0.663	Asia
Bangladesh	12.6	0.387	Asia
Armenia	13.9	0.159	Europe
Austria	8.2	0.535	Europe
Azerbaijan	13	0.455	Europe
Belarus	15.7	0.397	Europe
Belgium	10.1	0.747	Europe
Bahamas	14.5	0.386	North America
Barbados	12.2	0.091	North America
Belize	15.6	0.769	North America
Australia	5.2	0.124	Oceania

True Data



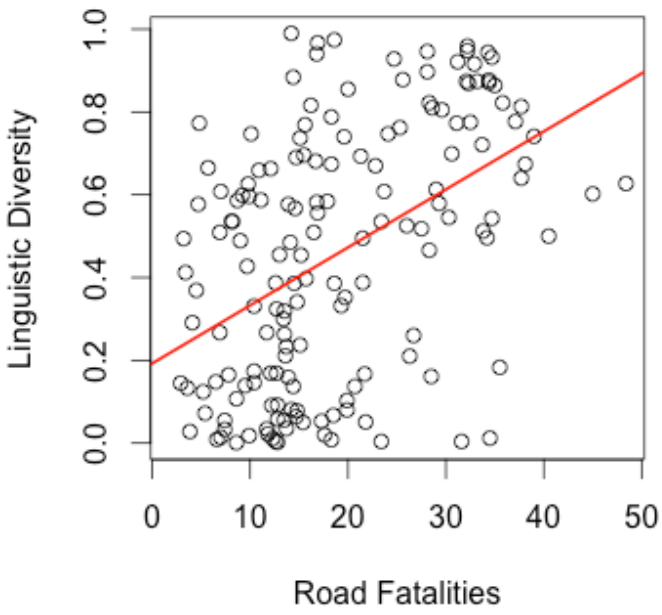
Permuted Data



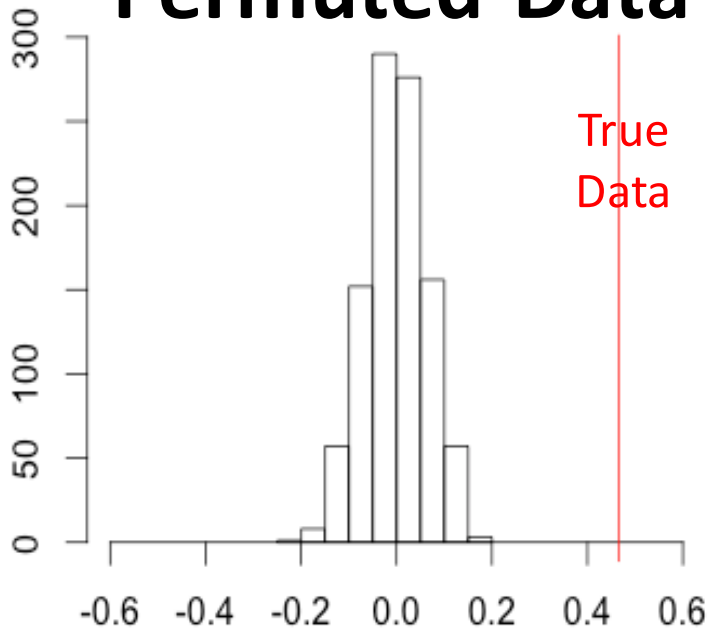
Stratified Permutation

Country	Road Fatalities	Linguistic Diversity	Continent
Bahrain	12.1	0.663	Asia
Bangladesh	12.6	0.387	Asia
Armenia	13.9	0.159	Europe
Austria	8.2	0.535	Europe
Azerbaijan	13	0.455	Europe
Belarus	15.7	0.397	Europe
Belgium	10.1	0.747	Europe
Bahamas	14.5	0.386	North America
Barbados	12.2	0.091	North America
Belize	15.6	0.769	North America
Australia	5.2	0.124	Oceania

True Data



Permuted Data



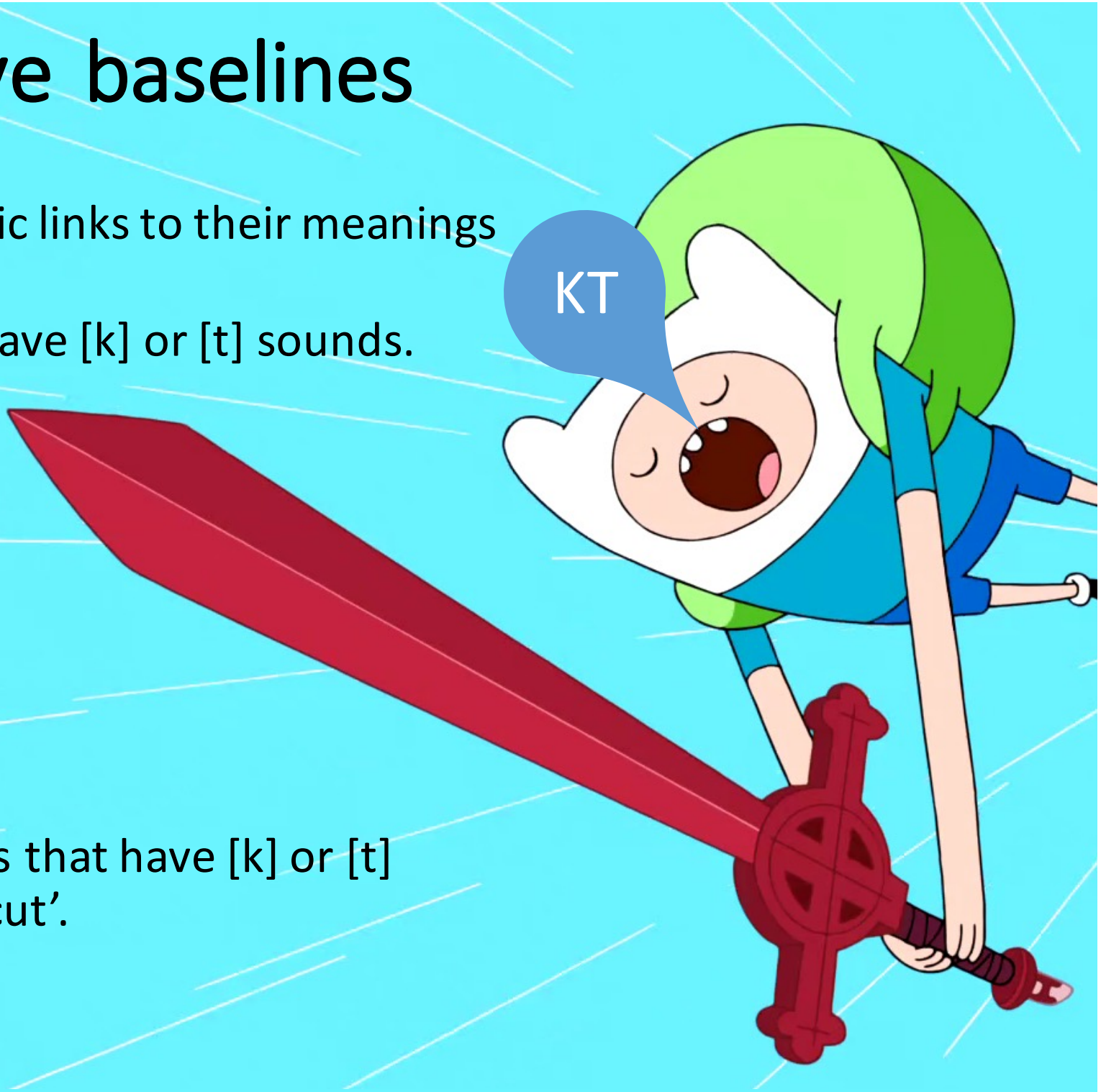
Alternative baselines

Theory:

Words have iconic links to their meanings

Hypothesis:

Words for ‘cut’ have [k] or [t] sounds.



Results:

63% of languages that have [k] or [t] in the word for ‘cut’.

Is this high?

Alternative baselines

Words for ‘cut’ have [g]

Words for ‘cut’ have [b]

Words for ‘push’ have [k] Words for ‘cut’ have [k] Words for ‘dog’ have [k]

Words for ‘cut’ have [a]

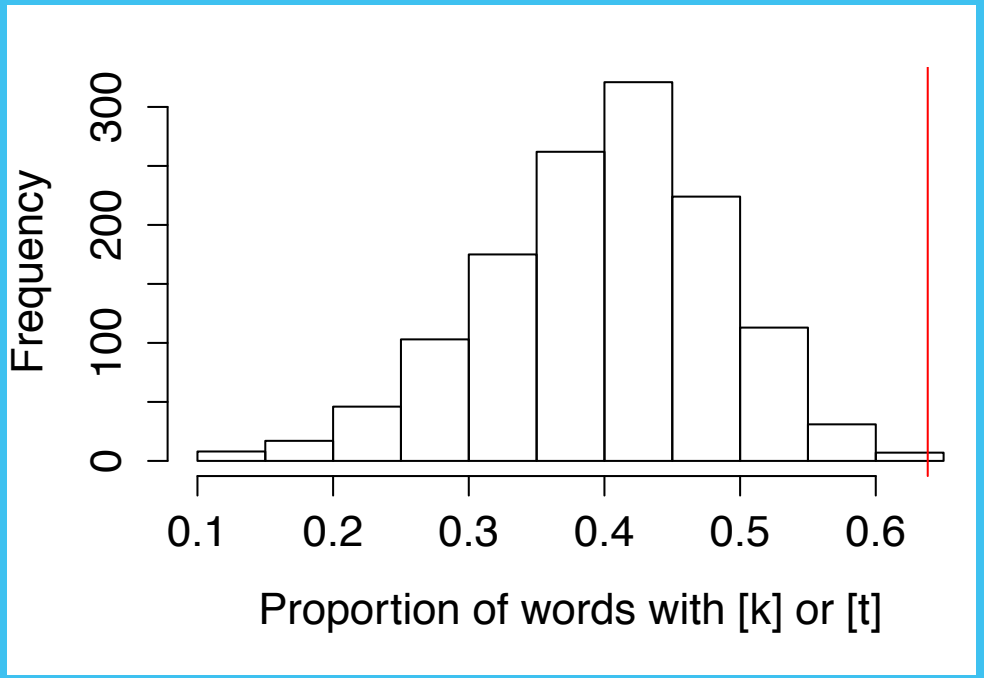
Words for ‘cut’ have [v]

Alternative baselines

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'.

Alternative baselines

Data source	What is sampled	How is it sampled
IDS	All concepts	Randomly
ASJP	Action concepts	Within families
		Within areas
	All sounds	Within areas & families
	Random consonants	

Custom null hypothesis testing

Permutation

Bootstrapping

Alternative baselines

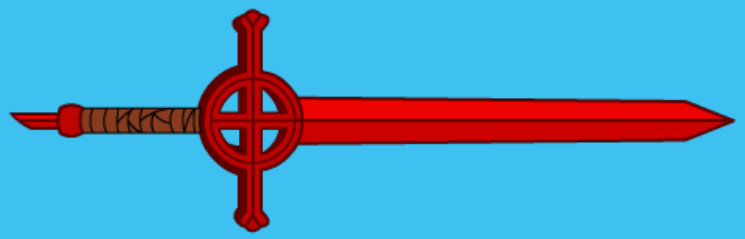
Example: Generating phylogeny for a small family

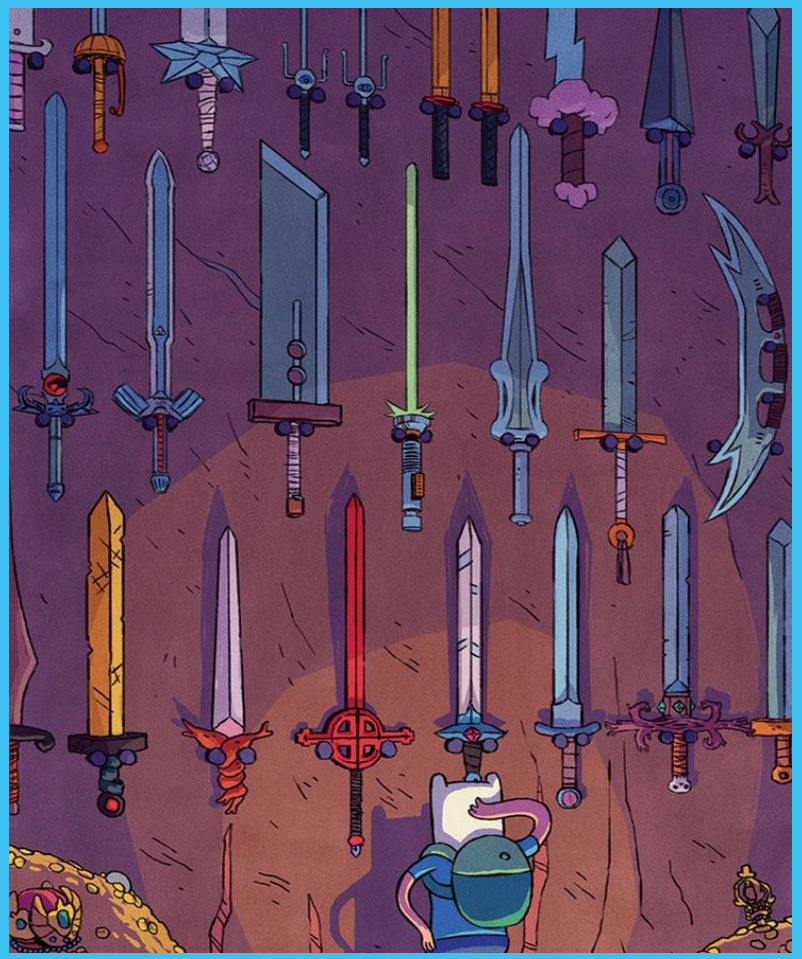
Little cognate overlap -> very general trees?

Is a small subset of the data having a big effect on the result?

Permute the data, see if it changes the result

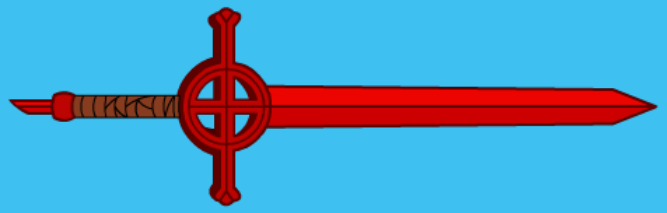
How to argue with data





Validity

Robustness



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

Result is the best answer given the assumptions

Easy to interpret
Susceptible to argument from authority

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

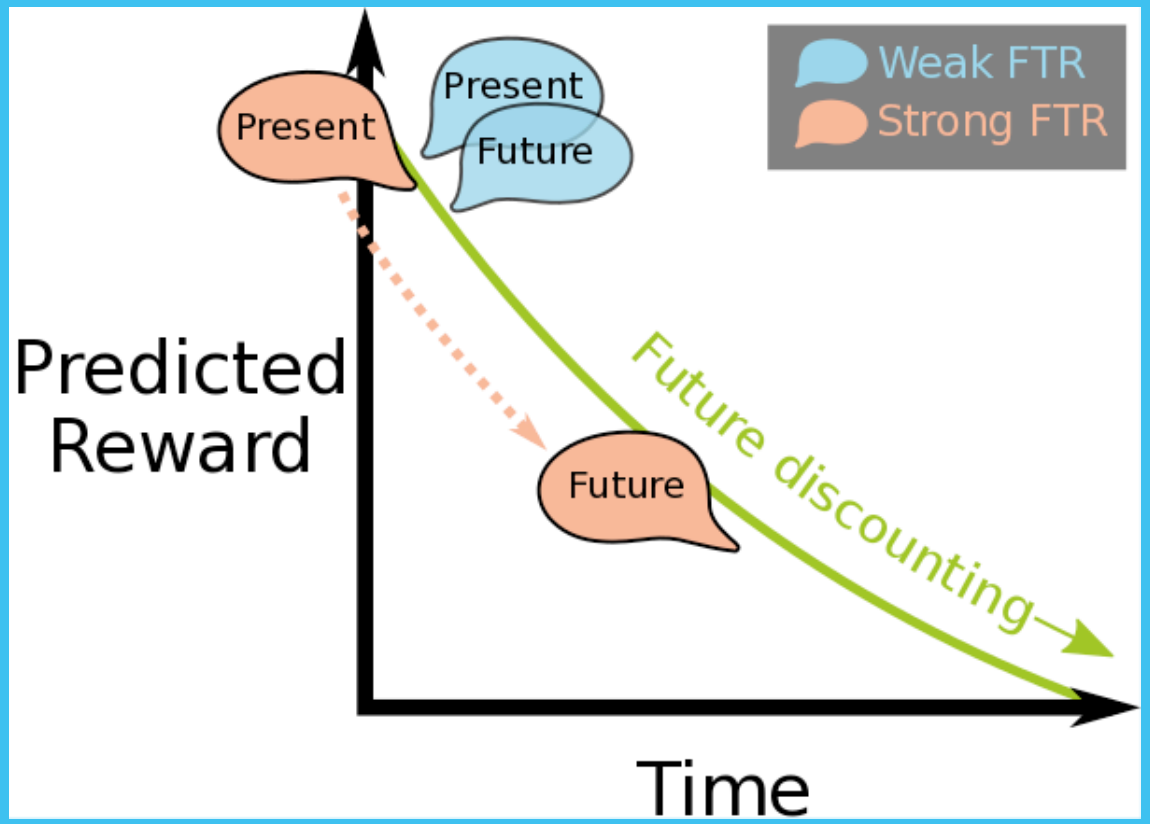
Result is a space where researchers can argue about data

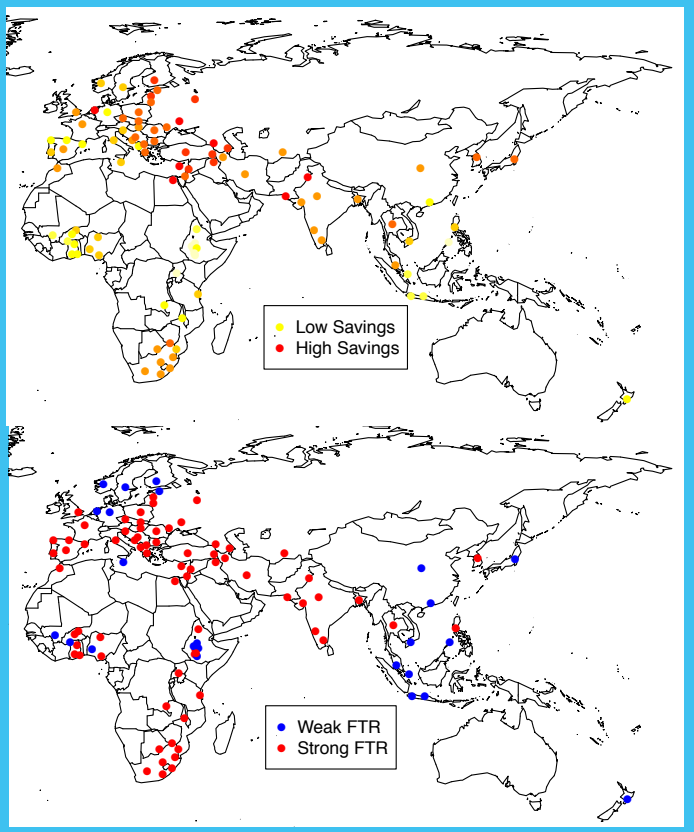
Engages more researchers
Susceptible to fishing / circular research

Keith Chen

Savings and future tense

Speakers of languages with no future tense are less likely to save money (Chen, 2013)



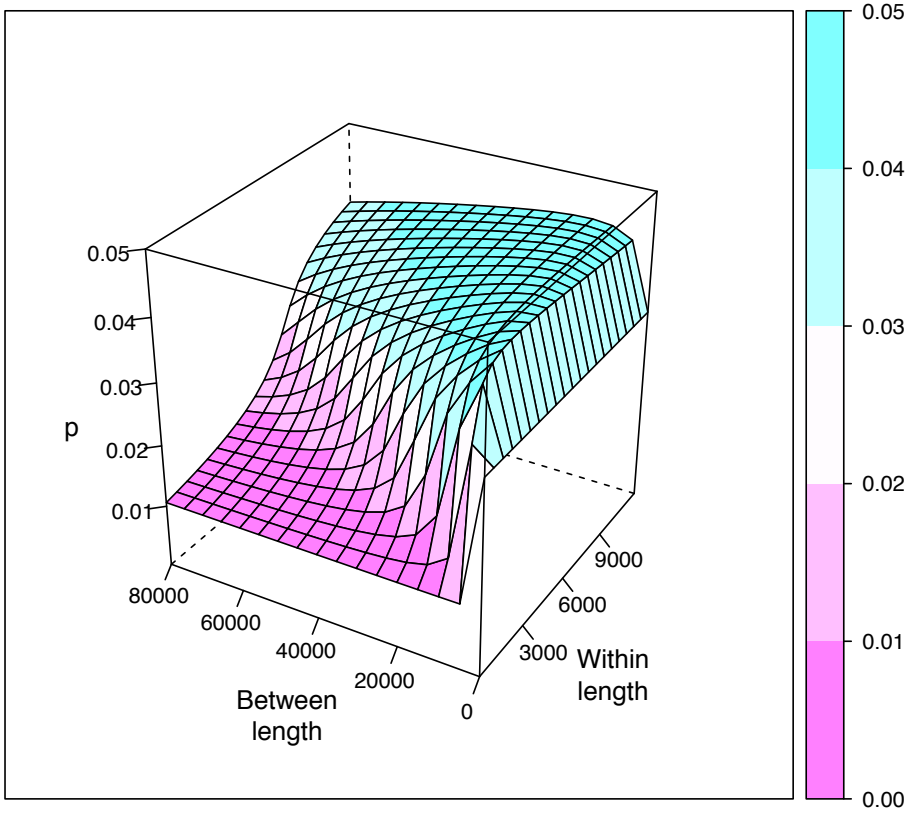


A space of results

Roberts, Winters & Chen (2015)

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

How to argue against data

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

Identify an assumption

which is problematic or missing

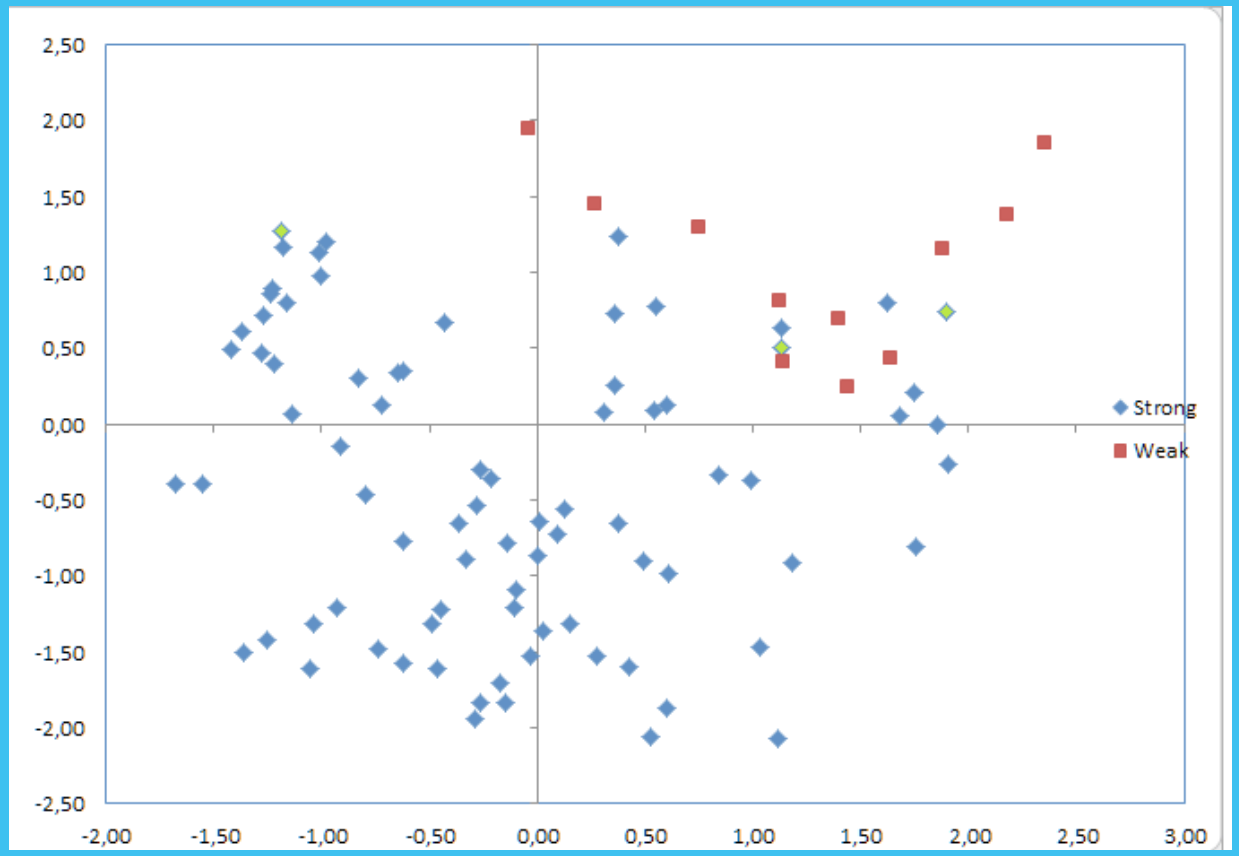
and

which provides an alternative explanation of the results

Criticism

Östen Dahl: Confounding variables

Traditional/
Secular-rational
values



Survival/Self-expression

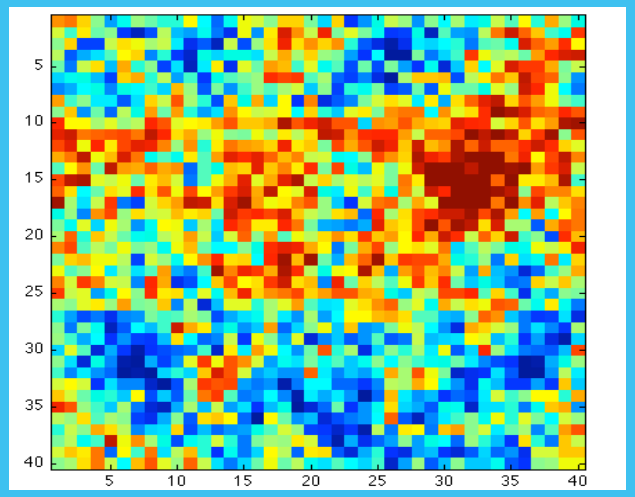
Inglehart & Welzel, see
http://dlc.hypotheses.org/360

Criticism

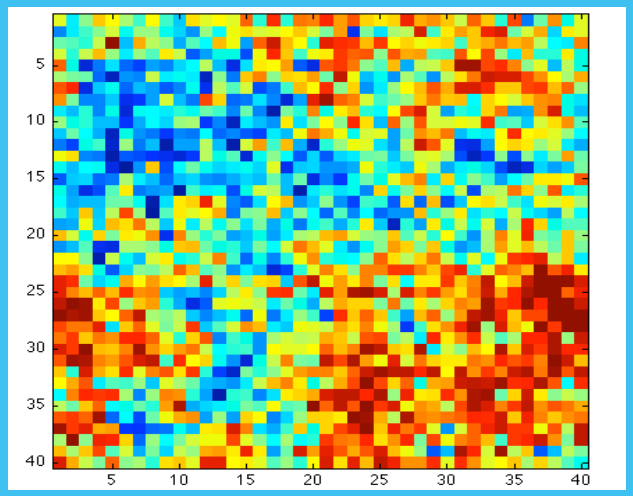
Problems with statistics

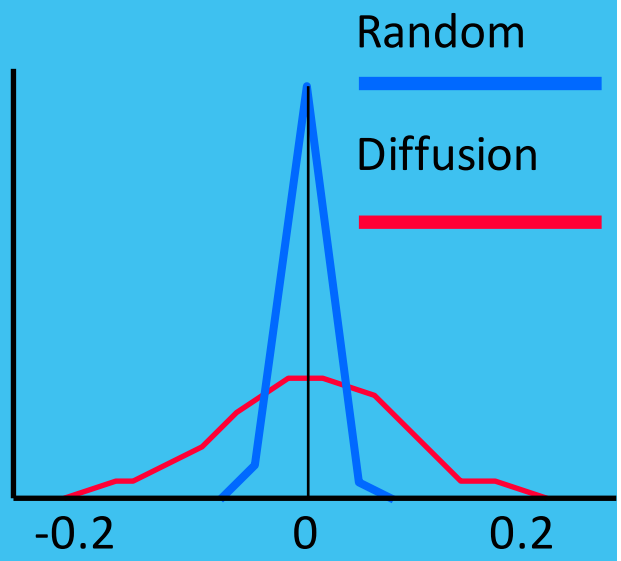
Mark Liberman: No control for geographic diffusion

Trait 1



Trait 2





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

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

Criticism

James Winters

Keith Chen

Data has historical relations

(Roberts, Winters & Chen, 2015)

Random effects for:

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

Result: Correlation between language and saving is not robust ($\chi^2 = 1.15$, $p = 0.28$)

Responses

I don’t believe you

But look …

What about this one counterexample?

Change it and see how the results change

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

Collaborate to use the right typology

Correlation does not imply causality

Use causally informed methods, experiments,

Robust approach

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.

Conclusion: Open Science

If you transparently communicate your:

- Data
- Assumptions
- Methods
- Conclusions

Then argument can be directed towards those,
rather than focusing on reputation/authority/orientation

As long as everyone has a common language
Be open! Learn your stats! Collaborate!

Questions?