PHYLOGENY OF THE EARLY GERMANIC LANGUAGES

Dario Papavassiliou & Keith M. Briggs
University of Warwick UWE & BT Research

Contents

Evolutionary linguistics

Charles Darwin offered languages as an illustrative example of evolution

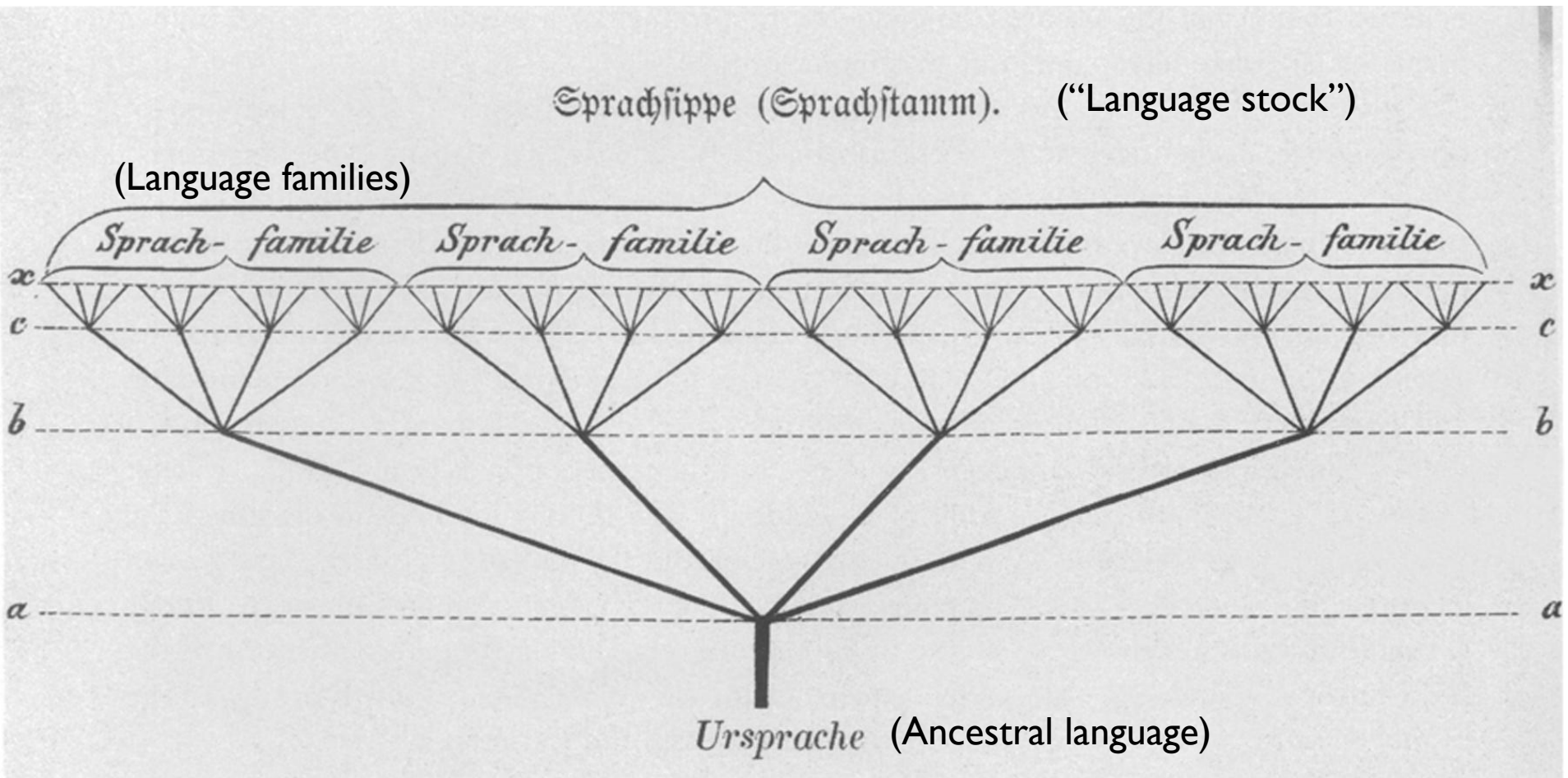
Languages show analogies to genetic features: **mutation** and **inheritance**

The history of languages has a close correspondence to the history of humanity

The origin of language ↔ the origin of modern humanity?

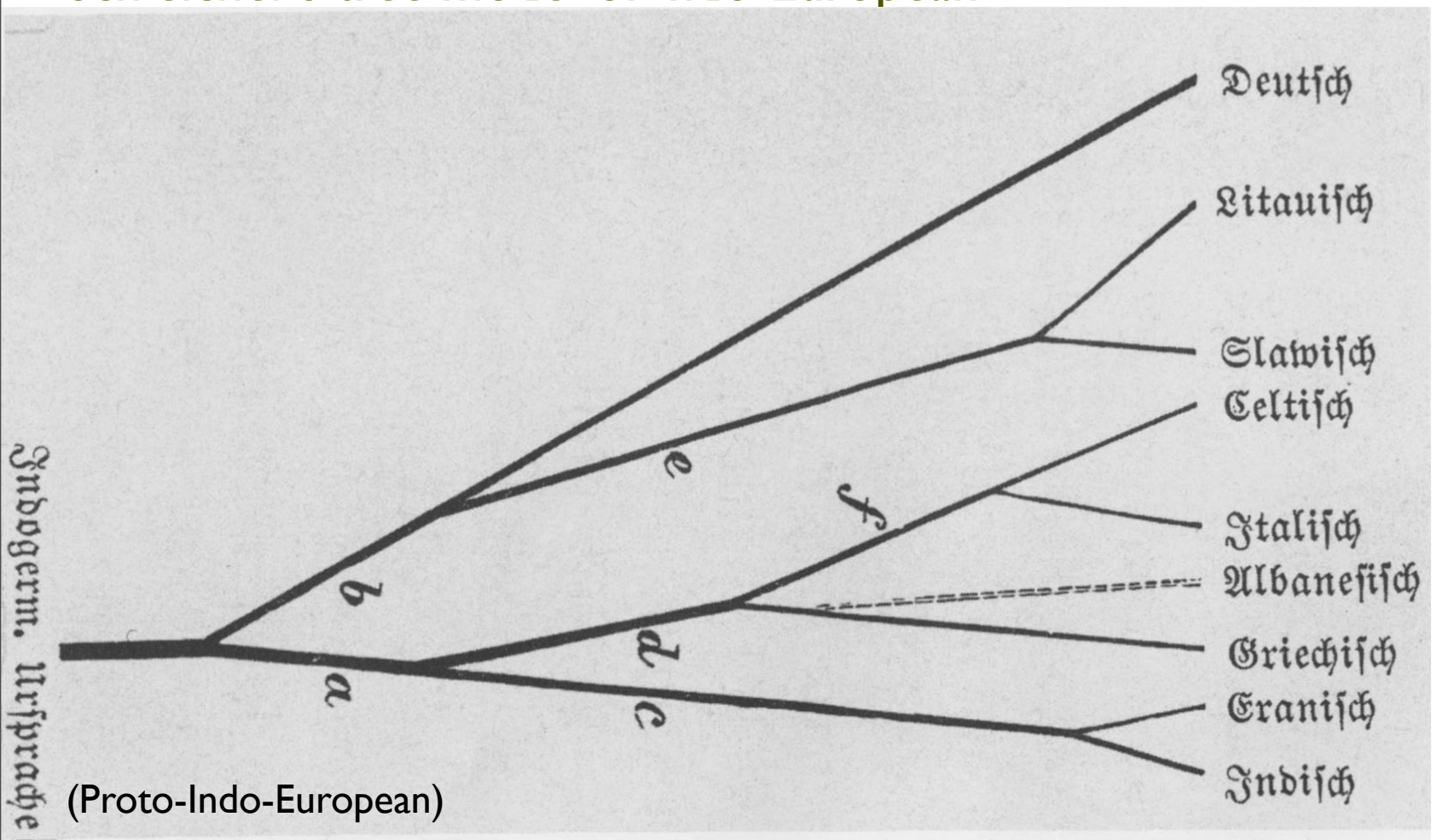
A brief history

Schleicher's tree model



A brief history

Schleicher's tree model of Indo-European



(German)

(Lithuanian)

(Slavic)
(Celtic)

(Italic)
(Albanian)

(Greek)
(Iranian)

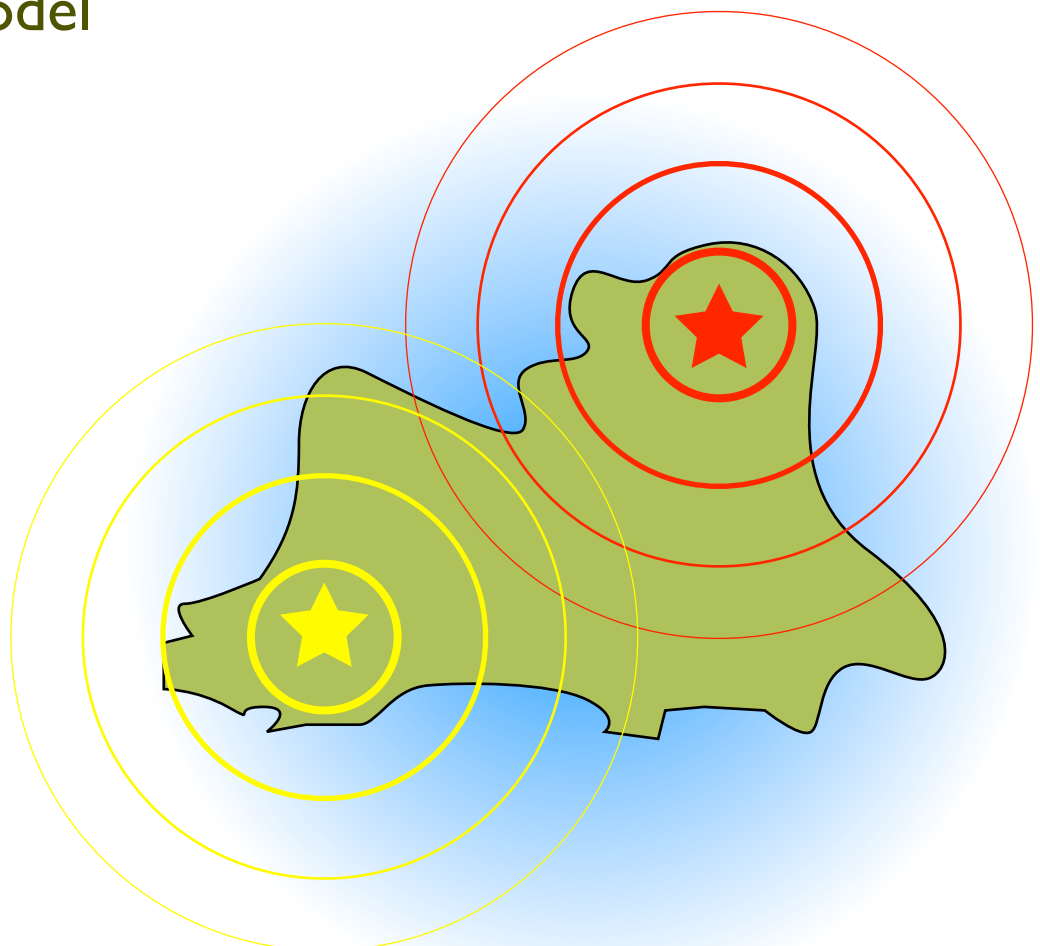
(Indian)

A brief history

Schmidt's wave model

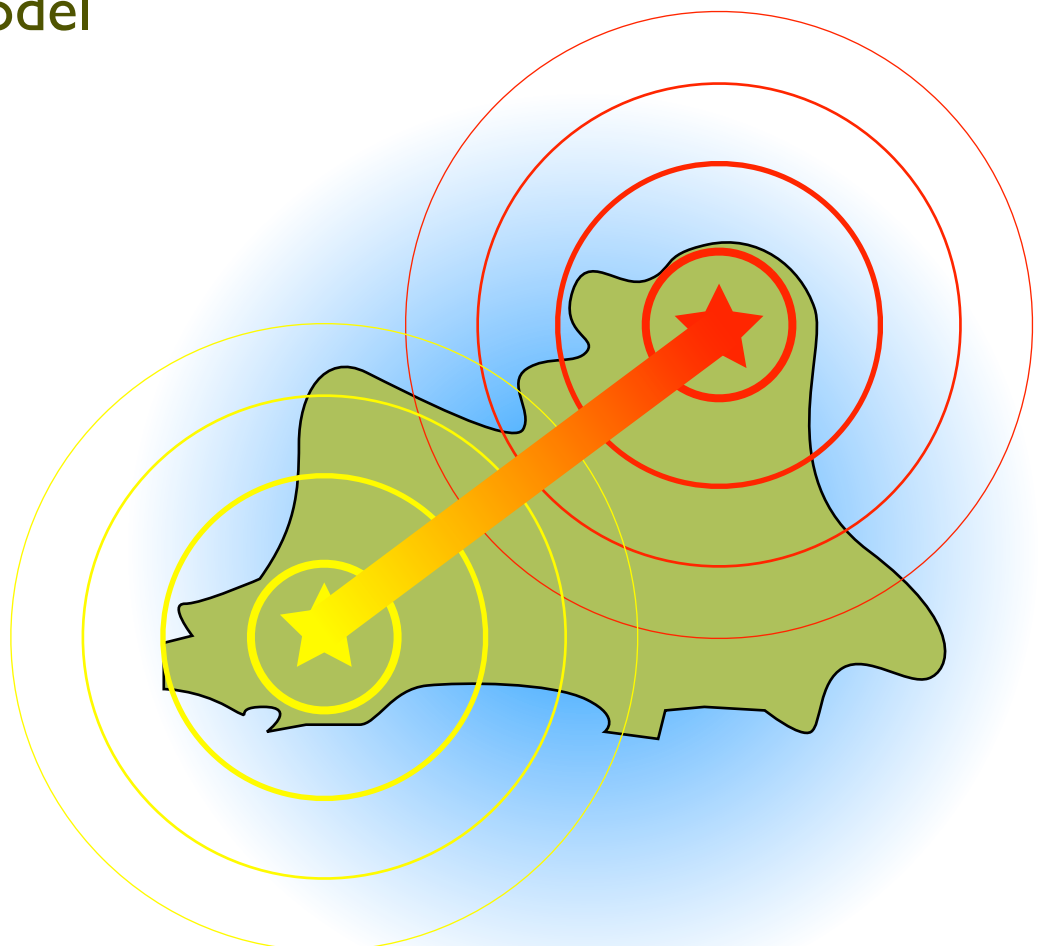
A brief history

Schmidt's wave model



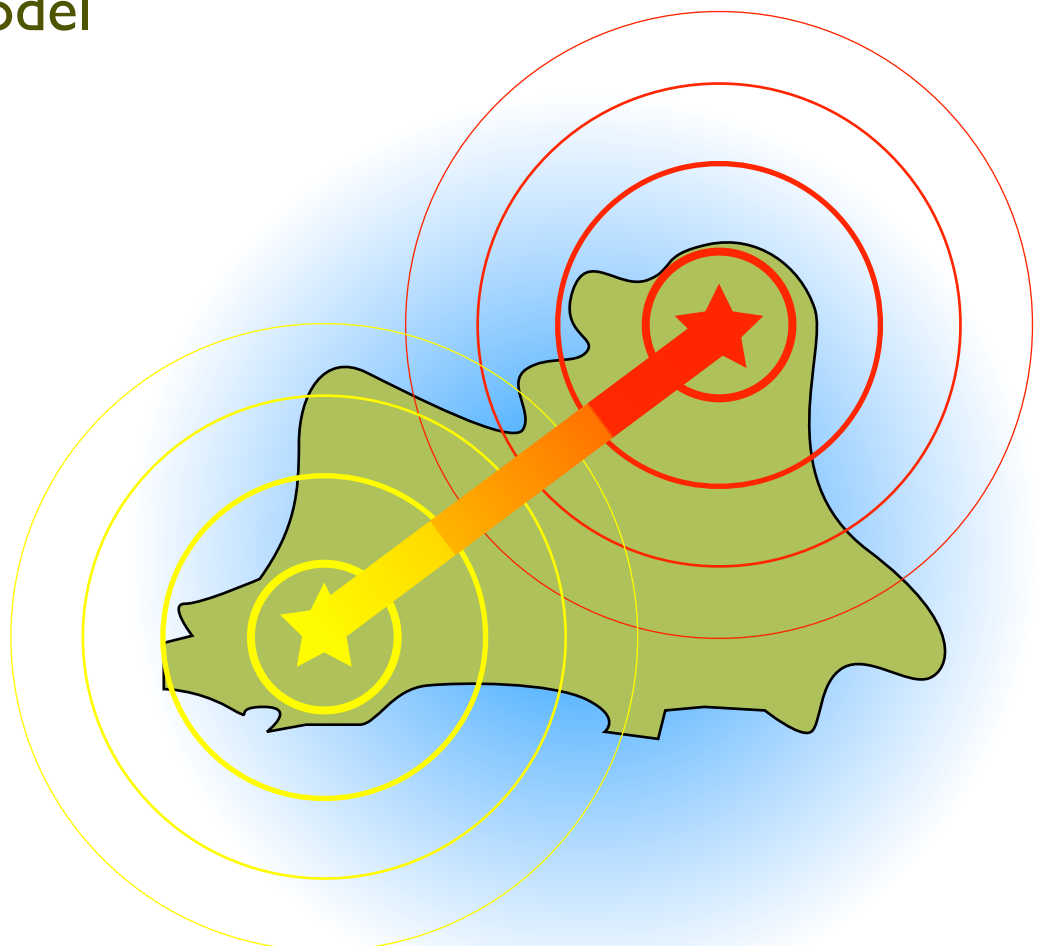
A brief history

Schmidt's wave model



A brief history

Schmidt's wave model



A brief history

Schmidt's wave model - The Balkan Sprachbund

4 Indo-European language families (Greek, Romance, Albanian, Slavic) and the unrelated Turkish

Share many grammatical (and lexical) features not seen elsewhere

A brief history

Real linguistic evolution is driven by a combination of these processes

Analogous to genetic evolution: *inheritance* versus *lateral transfer* (in viruses)

Inheritance is dominant in sparsely populated regions, lateral transfer becomes important when there is much contact between unrelated languages

(Strong influence of technology: writing, printing, internet...)

Challenges facing evolutionary linguists

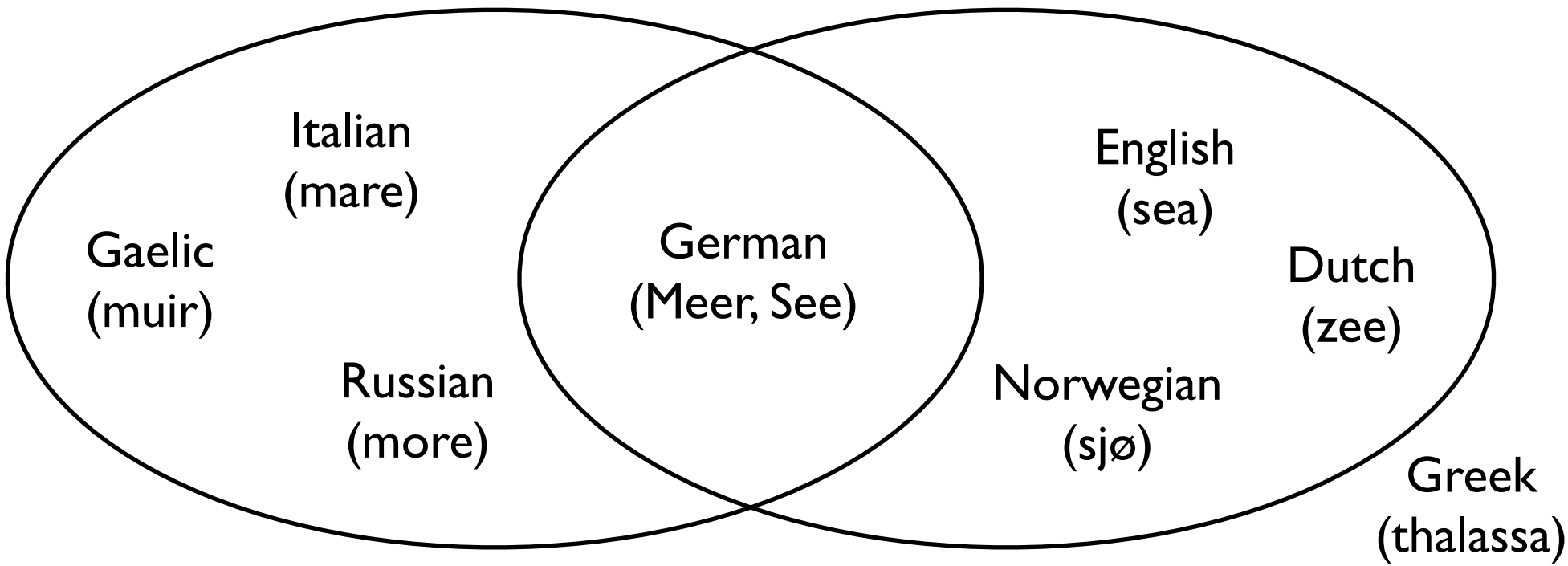
A (nearly) **total** absence of historical data!

Analysis must depend on observation of modern (*i.e.* written) languages, plus (more recently) modelling

Methodology-Swadesh lists

Very common for analyses to be based on lexical data: **Swadesh lists**

List of 100 common words thought to be particularly resistant to replacement by loanwords



Quantitative linguistics

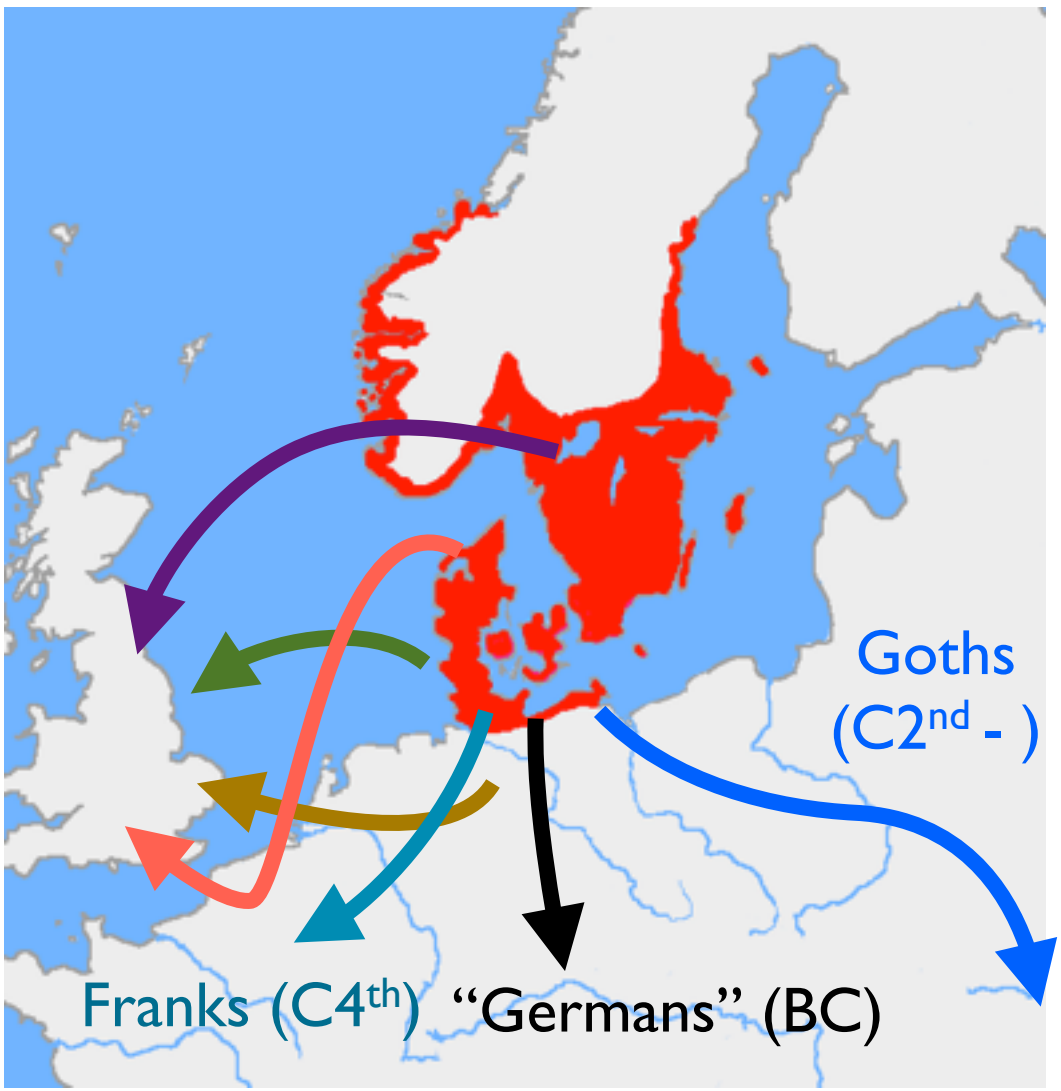
Swadesh lists allow for construction of a **"genome"** for languages

	Dutch	English	Gaelic	German	Italian	Norwegian	Russian	Greek
M*r	0	0	1	1	1	0	1	0
S*	1	1	0	1	0	1	0	0

This is then used with similar machinery as used to compare amino acid or DNA sequences

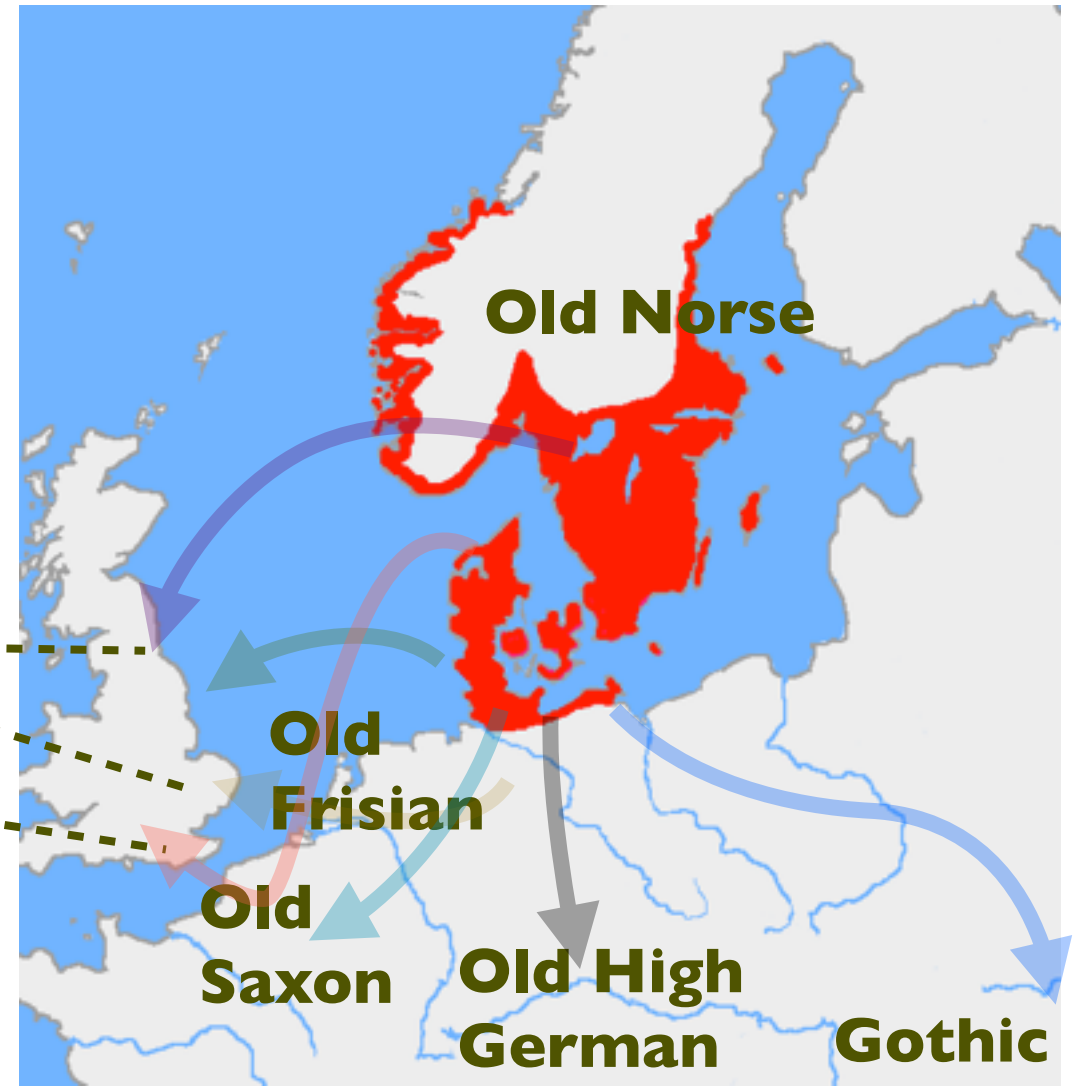
What if non-lexical data are used?

Data set - the early Germanic languages



Data set - the early Germanic languages

Old English dialects
- **Anglian**
- **West Saxon**
- **Kentish**

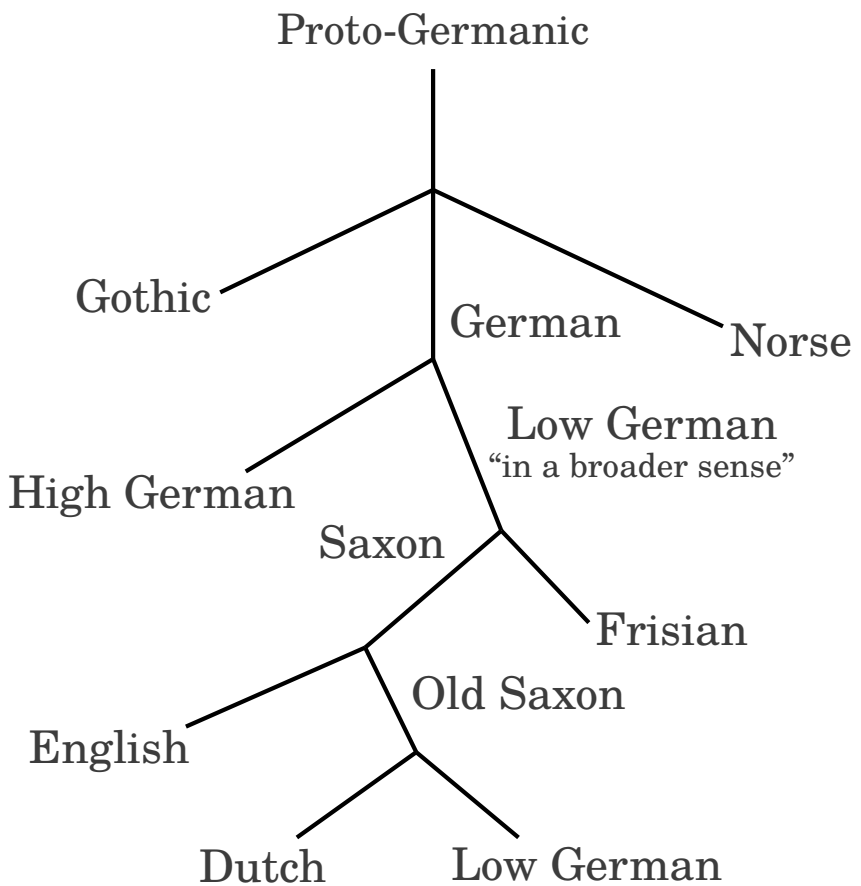


Data set - the early Germanic languages

A classic data set...

Old English dialects
- **Anglian**
- **Kentish**
- **West Saxon**

Old Frisian
Old Norse
Gothic
Old High German
Old Saxon



Schleicher's classification

Data set - source

Old English and the Continental Germanic Languages: A Survey of Morphological and Phonological Interrelations

Hans Frede Nielsen

Data set - source

Sample entry:

"The [Indo-European genitive singular] *ō*-stem ending *-ãs* is reflected in Gothic *gibōs*, ON *skarar*, OS *geƀa* and OHG *geba*, but not in OE *giefe* and OFris. *ieve*, where the original suffix has been analogically replaced by the [dative singular] ending ([reflecting Indo- European] *-ãi*)..."

Data set - interpretation as binary genome

Sample entry:

"The [Indo-European genitive singular] *ō*-stem ending *-ãs* is reflected in Gothic *gibōs*, ON *skarar*, OS *geƀa* and OHG *geba*, but not in OE *giefe* and OFris. *ieve*, where the original suffix has been analogically replaced by the [dative singular] ending ([reflecting Indo- European] *-ãi*)..."

	Reflects IE	
	Gen	Dat
OE Anglian	**0**	**1**
OE Kentish	**0**	**1**
OE W Saxon	**0**	**1**
O Frisian	**0**	**1**
O Saxon	**1**	**0**
O H German	**1**	**0**
O Norse	**1**	**0**
Gothic	**1**	**0**

Data set - interpretation as binary genome

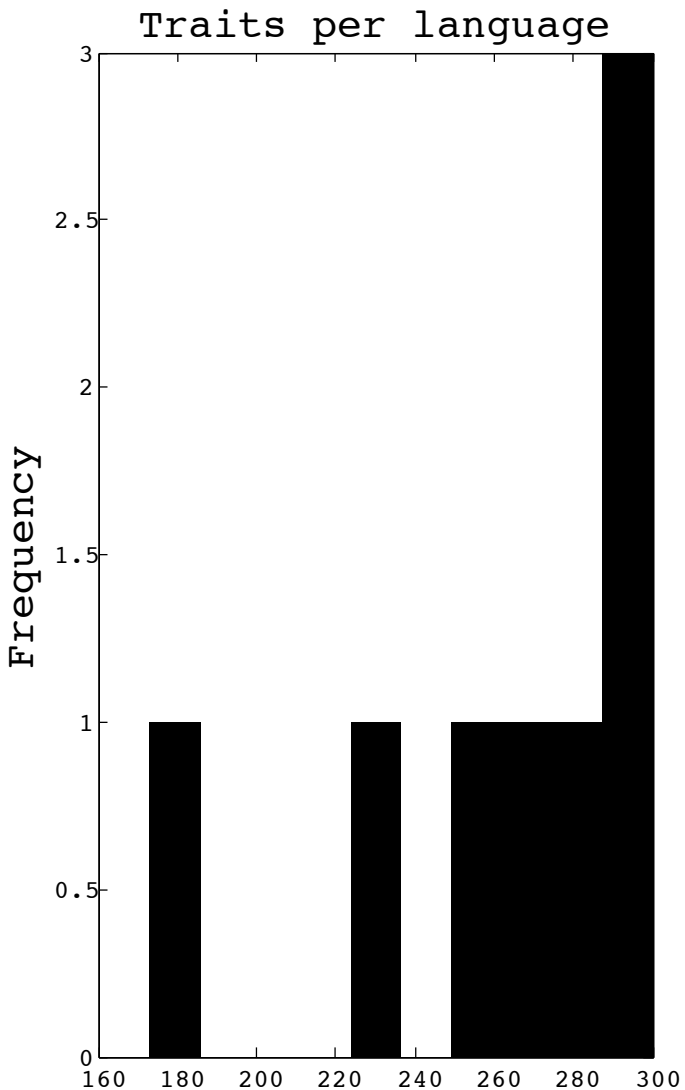
Missing data marked with **?**
Omitted data (*duplicate entries, "insignificant/late", too subtle*) marked with **-** and disregarded

Results in a 'genome' of 531 characters for each language

Can be **filtered** into sub-genomes for different linguistic categories (*nouns, verbs, numerals..., vowels, consonants*) and (in principle) **weighted**

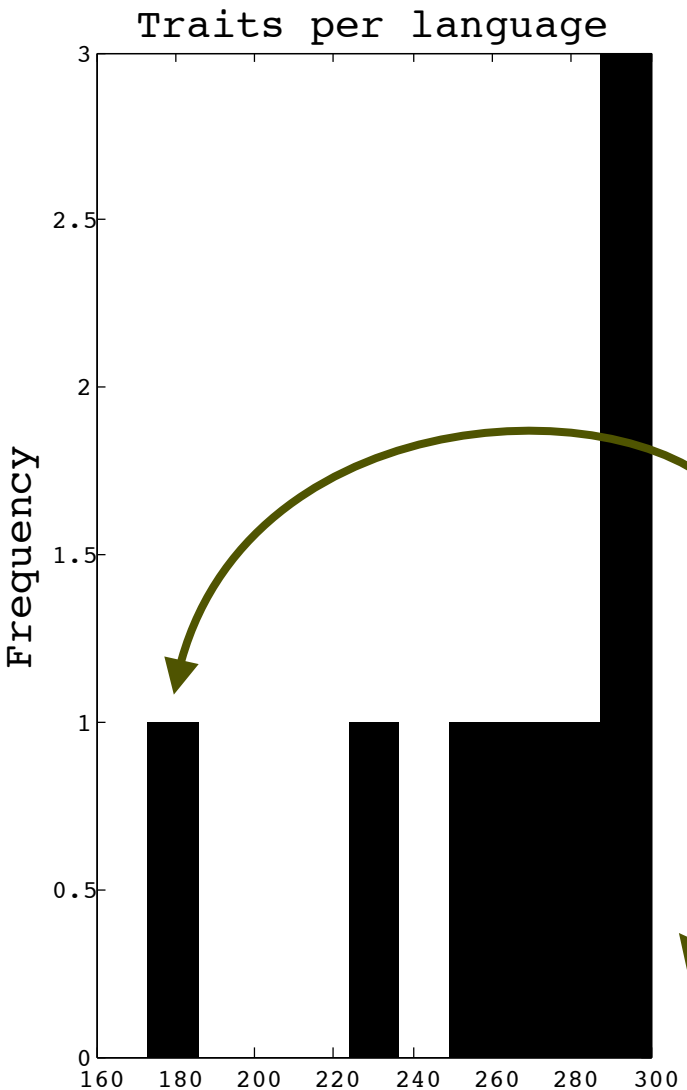
	Reflects IE	
	Gen	Dat
OE Anglian	**0**	**1**
OE Kentish	**0**	**1**
OE W Saxon	**0**	**1**
O Frisian	**0**	**1**
O Saxon	**1**	**0**
O H German	**1**	**0**
O Norse	**1**	**0**
Gothic	**1**	**0**

Statistics - traits per language



A very basic indication of the completeness of the data

Statistics - traits per language

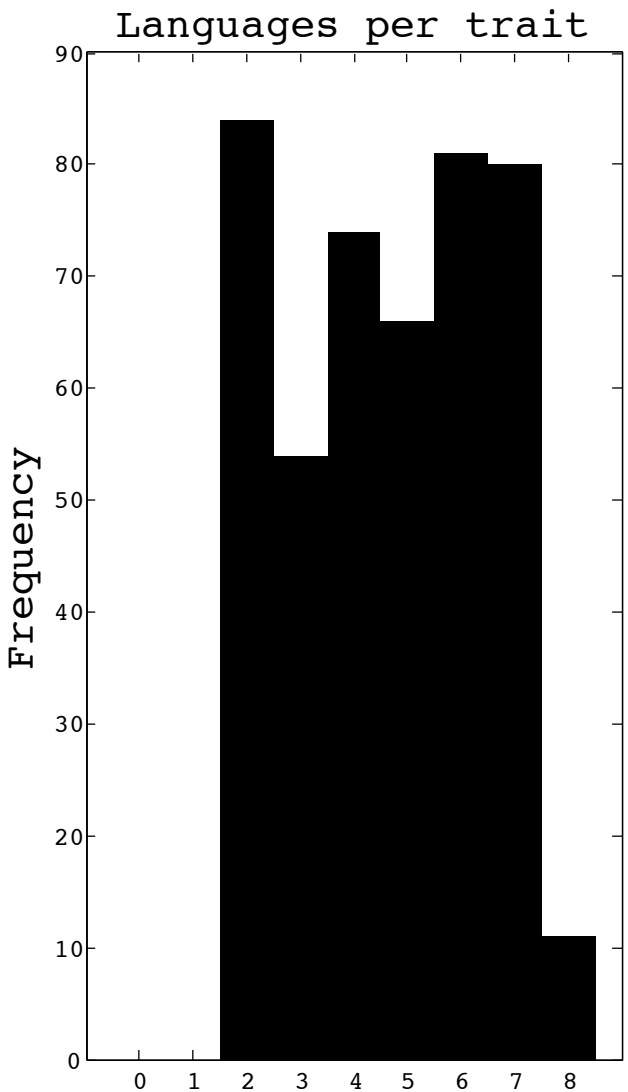


A very basic indication of the completeness of the data

Gothic under-represented (due to a lack of Gothic sources)

Old English dialects over-represented (due to subject of book)

Statistics - languages per trait

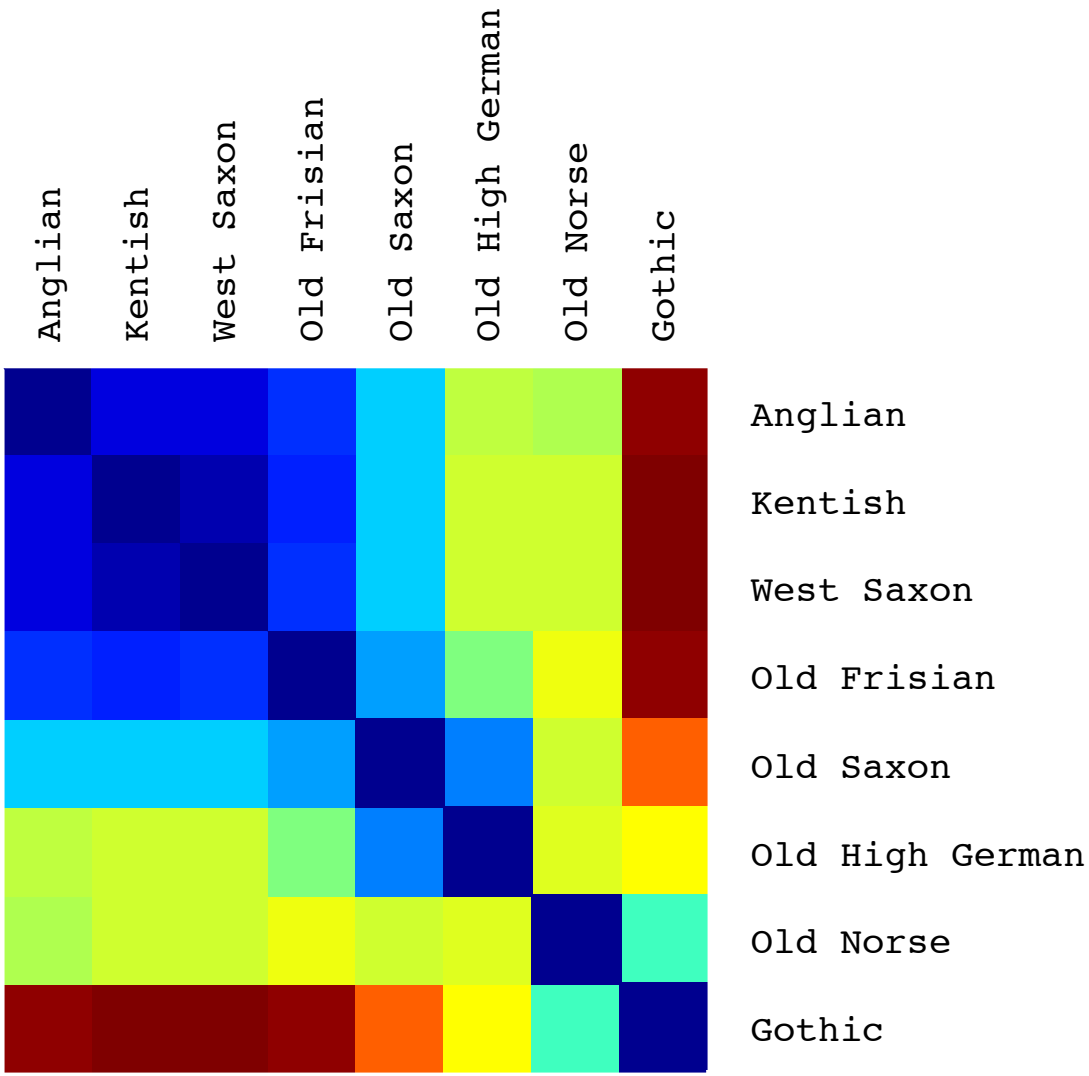


Since book focuses on *relationships* between languages it does not discuss traits seen in only one language

Traits seen in all, or none, of the species are uninformative

Flat distribution → timescale of evolution is *long*

Statistics - distance matrix



Form distance matrix by counting differences in genome

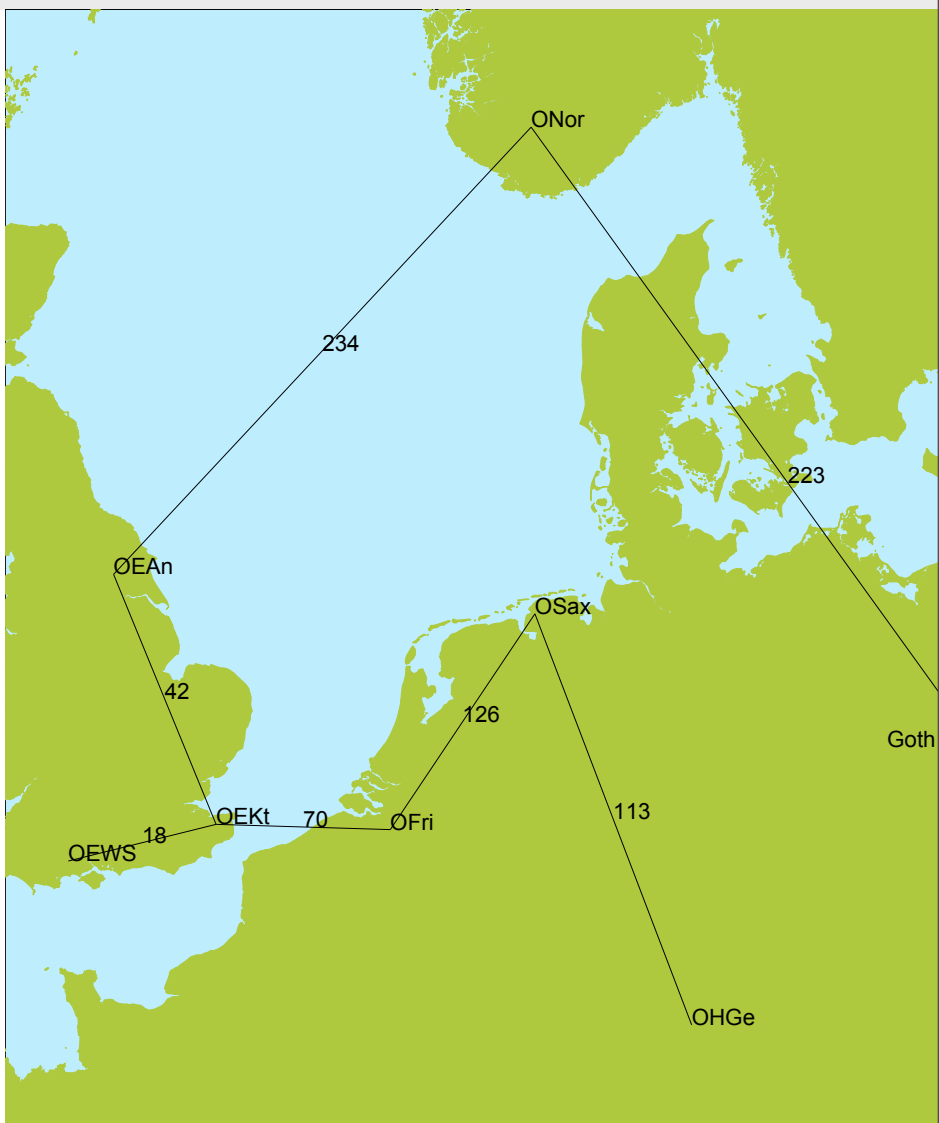
Some relationships immediately apparent

Minimal spanning tree

A very crude quantification of distances between languages

Construct a full graph with edge weights defined as distance

Delete edges with large weight to give *minimal spanning tree*

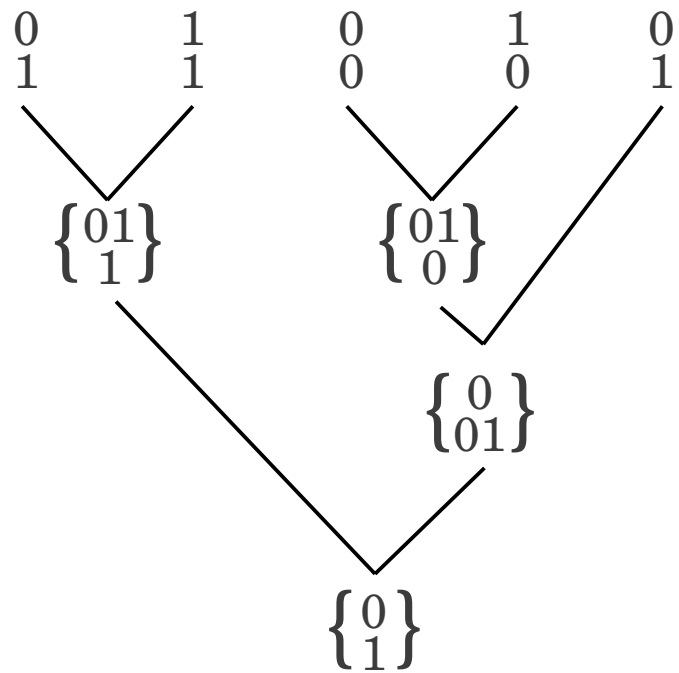


Maximum parsimony

Minimises number of changes over tree to obtain observed genomes

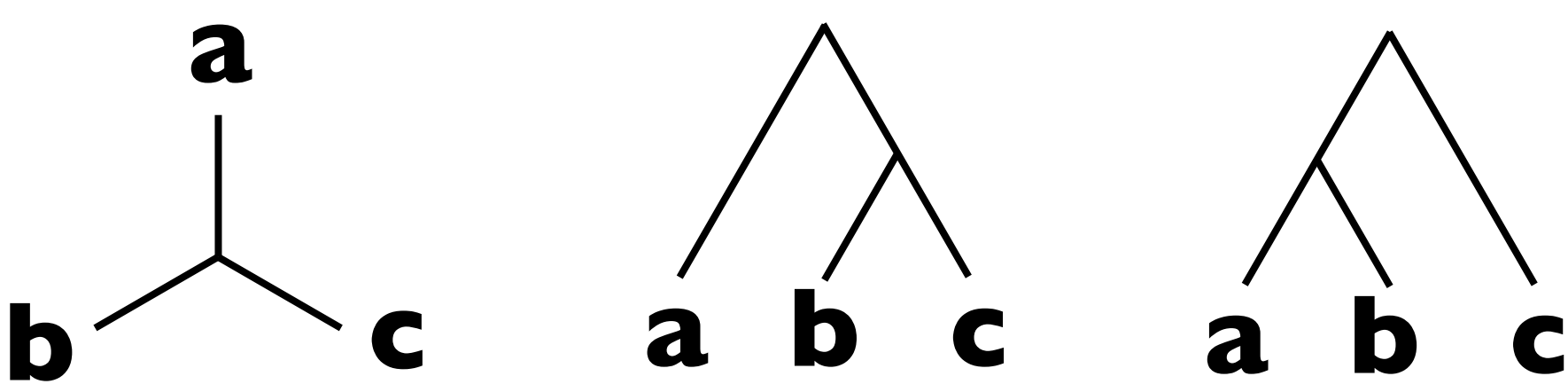
Implemented using the *Fitch algorithm*

Repeated for each character in genome, then for each possible tree topology



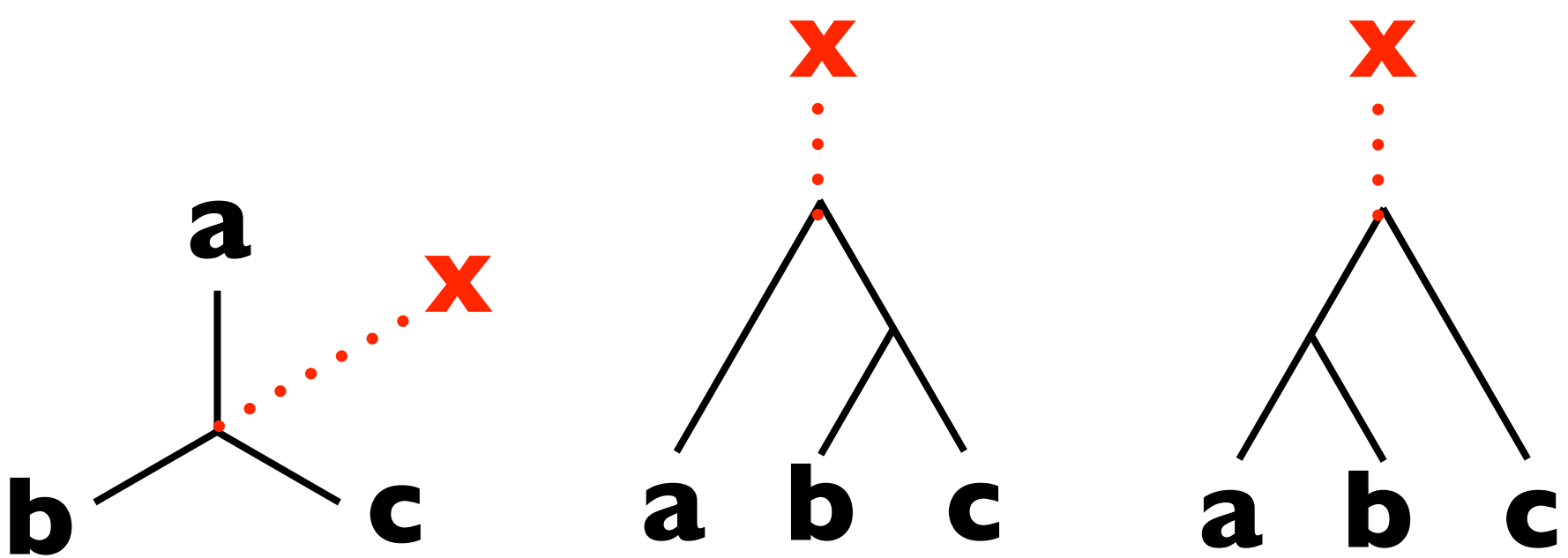
Maximum parsimony

Unless ancestral state is a leaf, the tree is *unrooted*



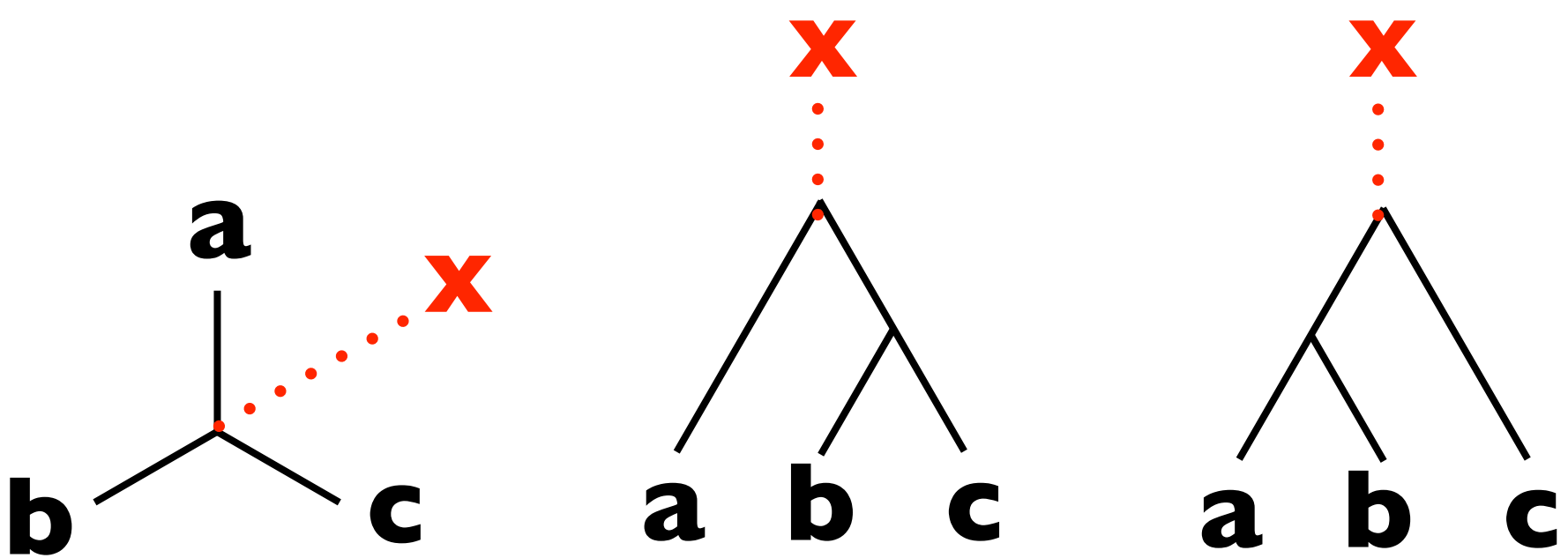
Maximum parsimony

Unless ancestral state is a leaf, the tree is *unrooted*



Maximum parsimony

Unless ancestral state is a leaf, the tree is *unrooted*



Gothic chosen as outgroup due to distance from other languages

Maximum parsimony

Gives a sensible tree topology, but unrooted tree → cannot resolve EG/WG/NG split!

Gives only information on topology, not chronology

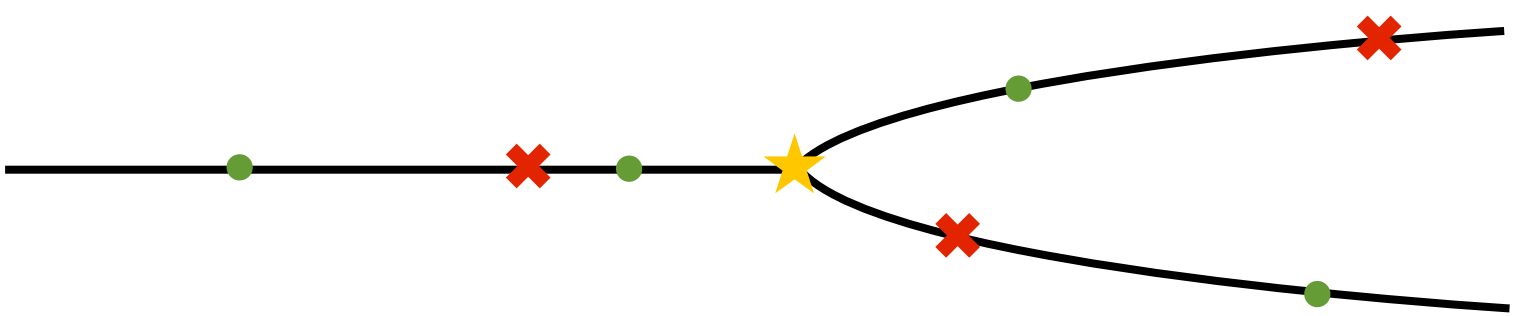
Markov chain Monte Carlo - Dollo model

Evolution modelled as a collection of Poisson processes:

Trait **born** with rate λ

Trait **dies** with rate μ

Lineage **splits** with rate θ



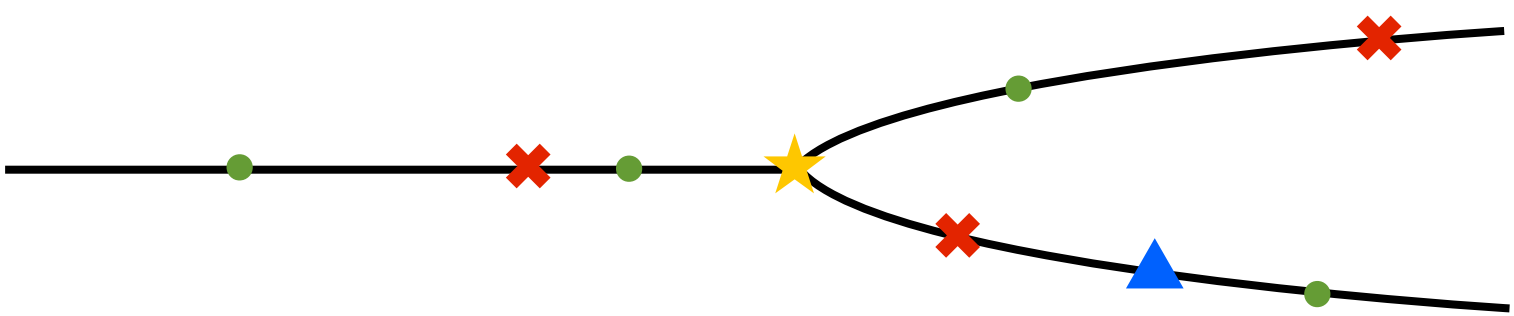
Markov chain Monte Carlo - Dollo model

Evolution modelled as a collection of Poisson processes:

Trait **born** with rate λ

Trait **dies** with rate μ

Lineage **splits** with rate θ



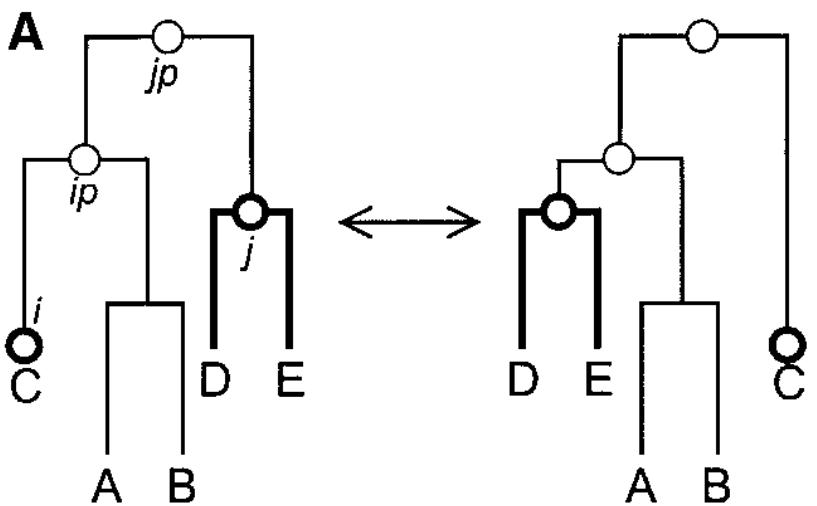
Catastrophe occurs with rate ρ: each trait dies with P(κ), Poisson(κλ/μ) new traits born

Equivalent to an edge lengthening

Markov chain Monte Carlo - Implementation

Implemented using the *TraitLab* package*

MCMC scheme example moves



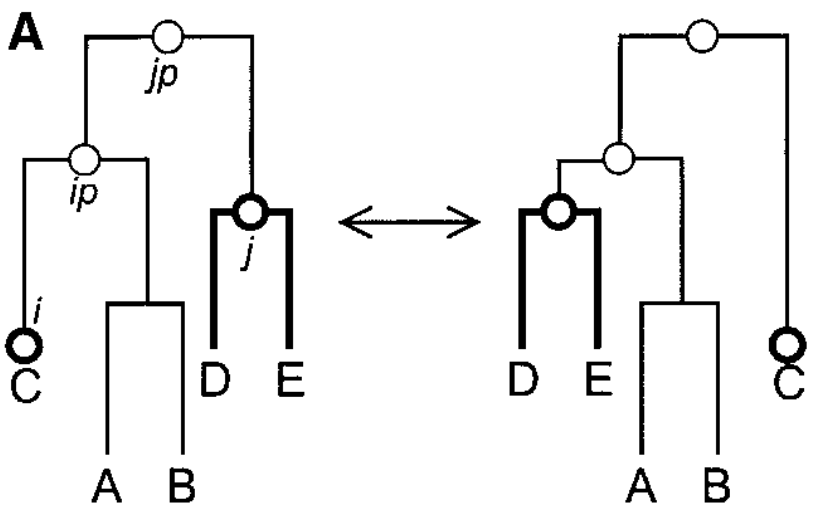
Change tree topology

*Geoff Nicholls, Oxford

Markov chain Monte Carlo - Implementation

Implemented using the *TraitLab* package*

MCMC scheme example moves



Change tree topology

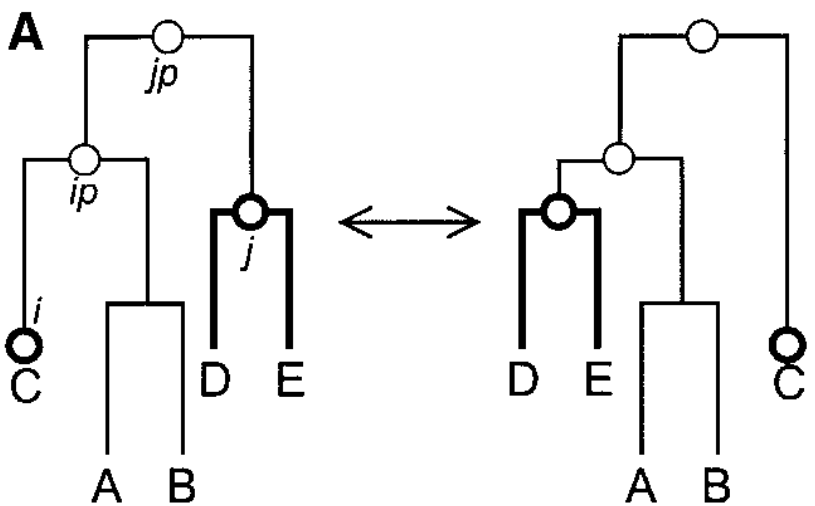
Vary model parameters

*Geoff Nicholls, Oxford

Markov chain Monte Carlo - Implementation

Implemented using the *TraitLab* package*

MCMC scheme example moves



Change tree topology

Vary model parameters

Vary locations of catastrophes

*Geoff Nicholls, Oxford

Markov chain Monte Carlo - Implementation

1,000,000 steps performed

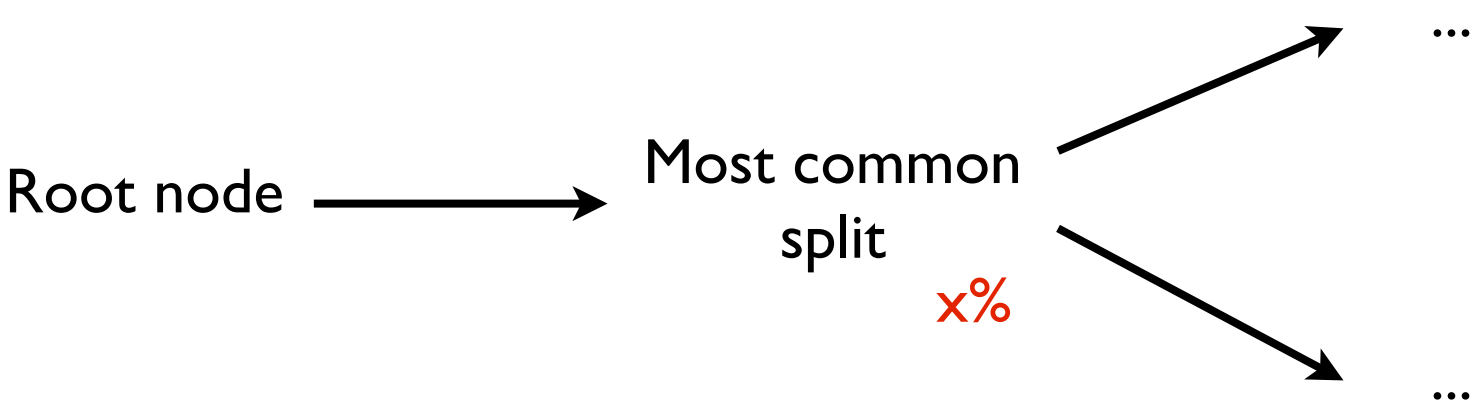
First 100,000 discarded (equilibration)

Remaining sampled every 100 steps

Samples averaged to give a *consensus tree*

Consensus tree

Given a set of *N* trees, a *consensus tree* representing an 'average' topology is constructed:



Results

Obtain same tree (topologically) as from parsimony

Chronological resolution groups NG with WG

Very good consensus between samples

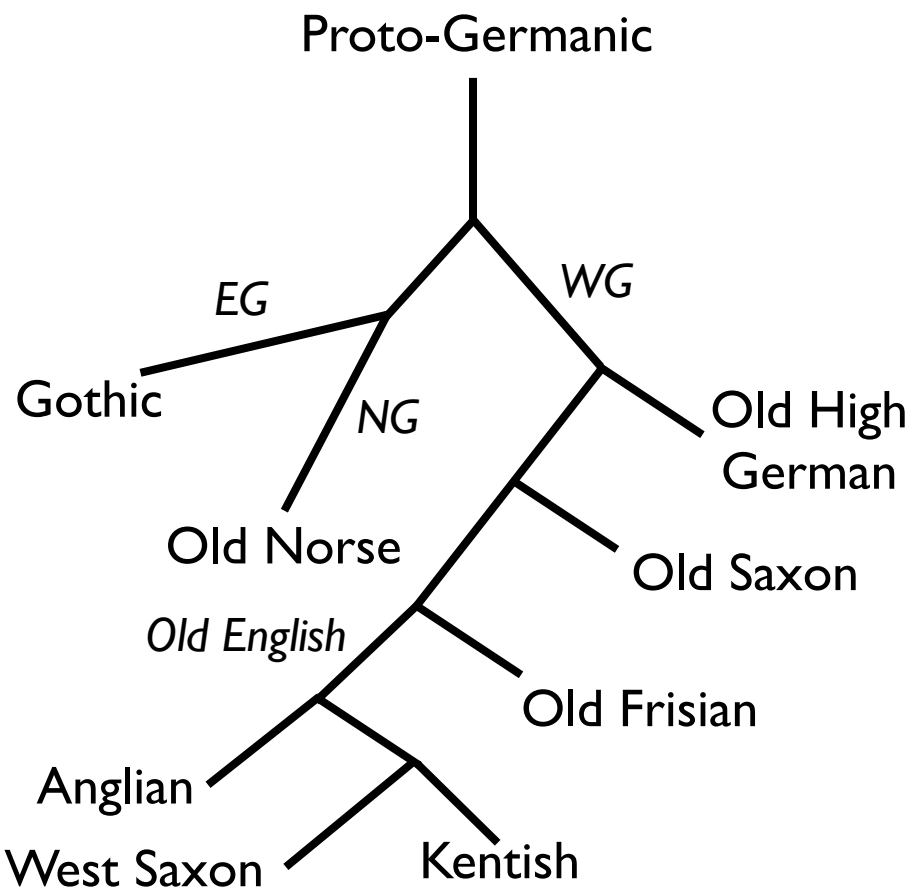
Results


Morphology
ONor
75
OEAn
OSax
Goth
OEKt
OFri
97
OEWS
OHGe


Phonology - vowels
ONor
71
OEAn
OSax
81
80
Goth
98
74
OEKt
OFri
OEWS
OHGe

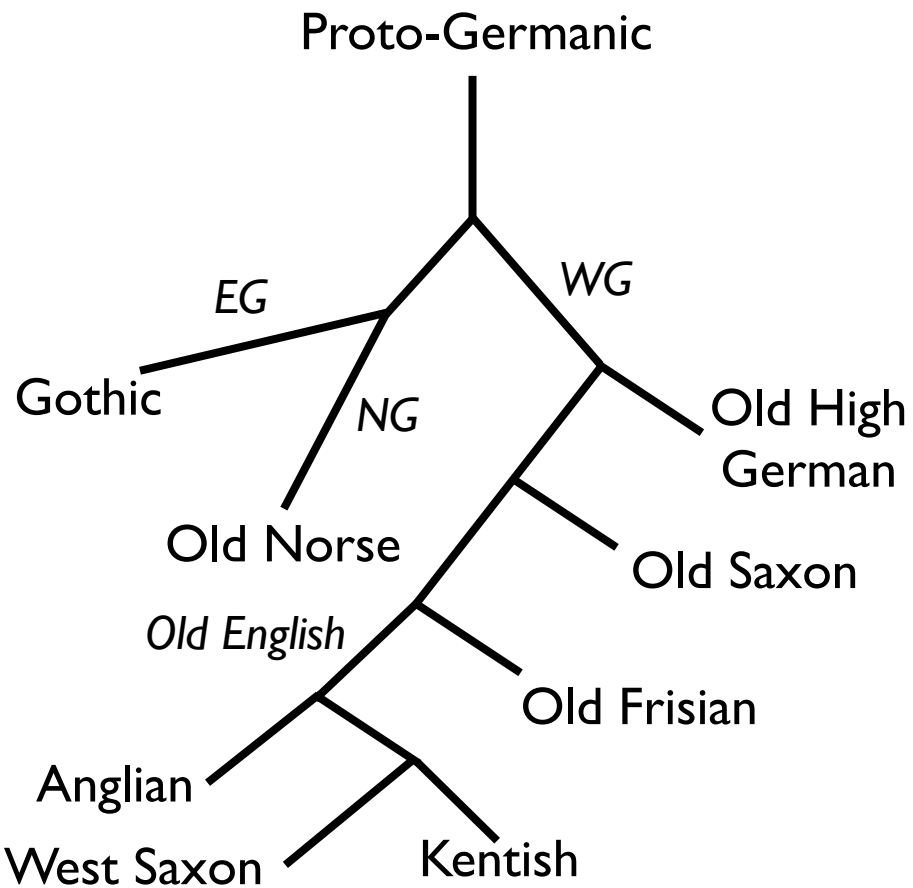
Results

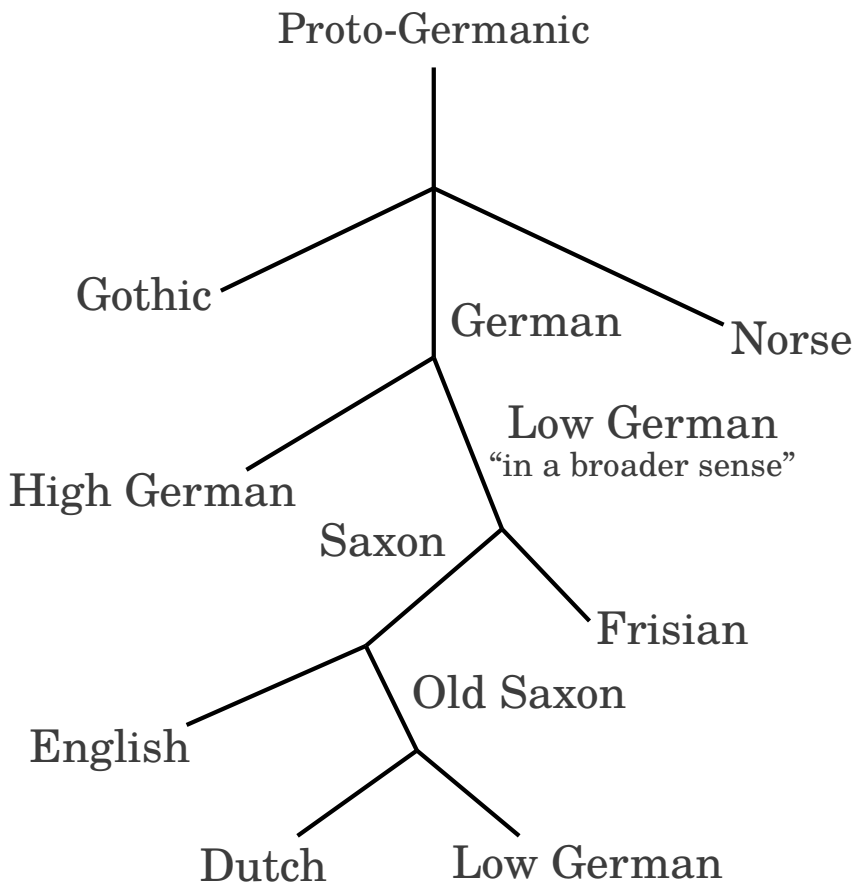
We obtain the following phylogeny...



Conclusions

Compares (mostly) favourably to Schleicher's classification

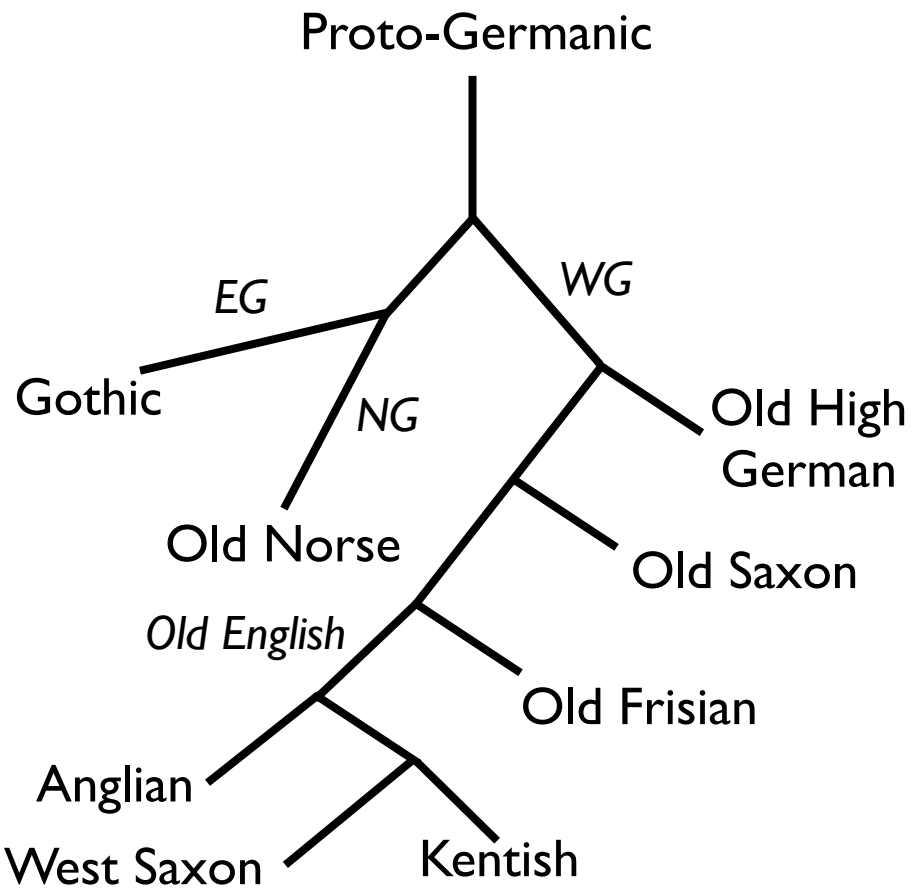


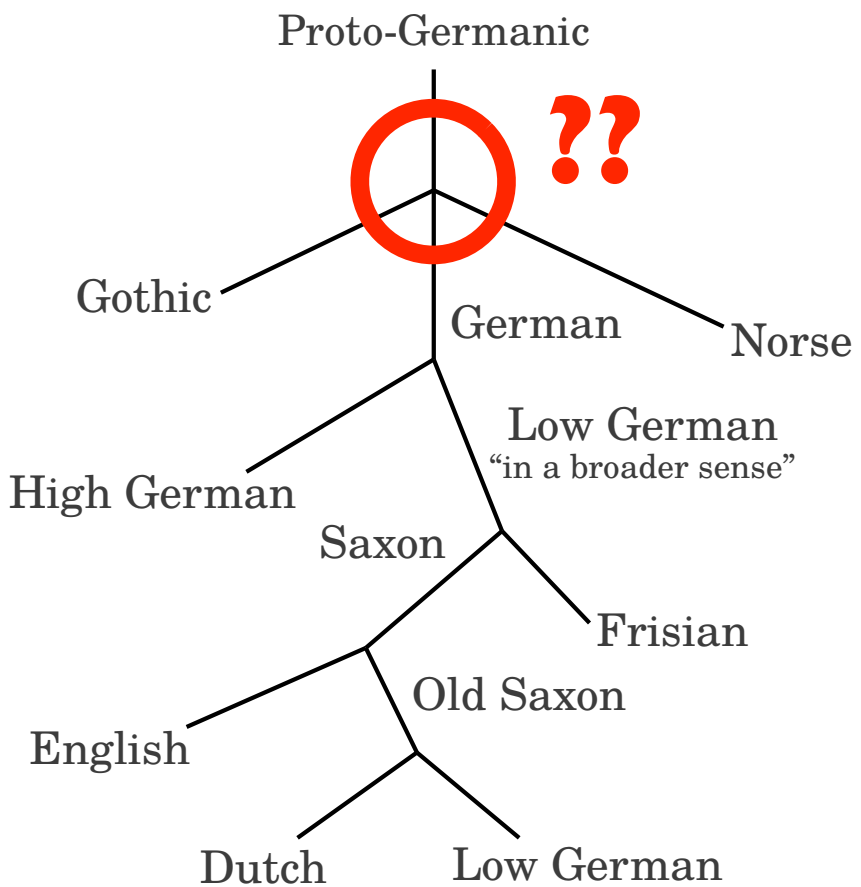


as well as quantitative (lexical) analyses by others

Conclusions

Compares (mostly) favourably to Schleicher's classification

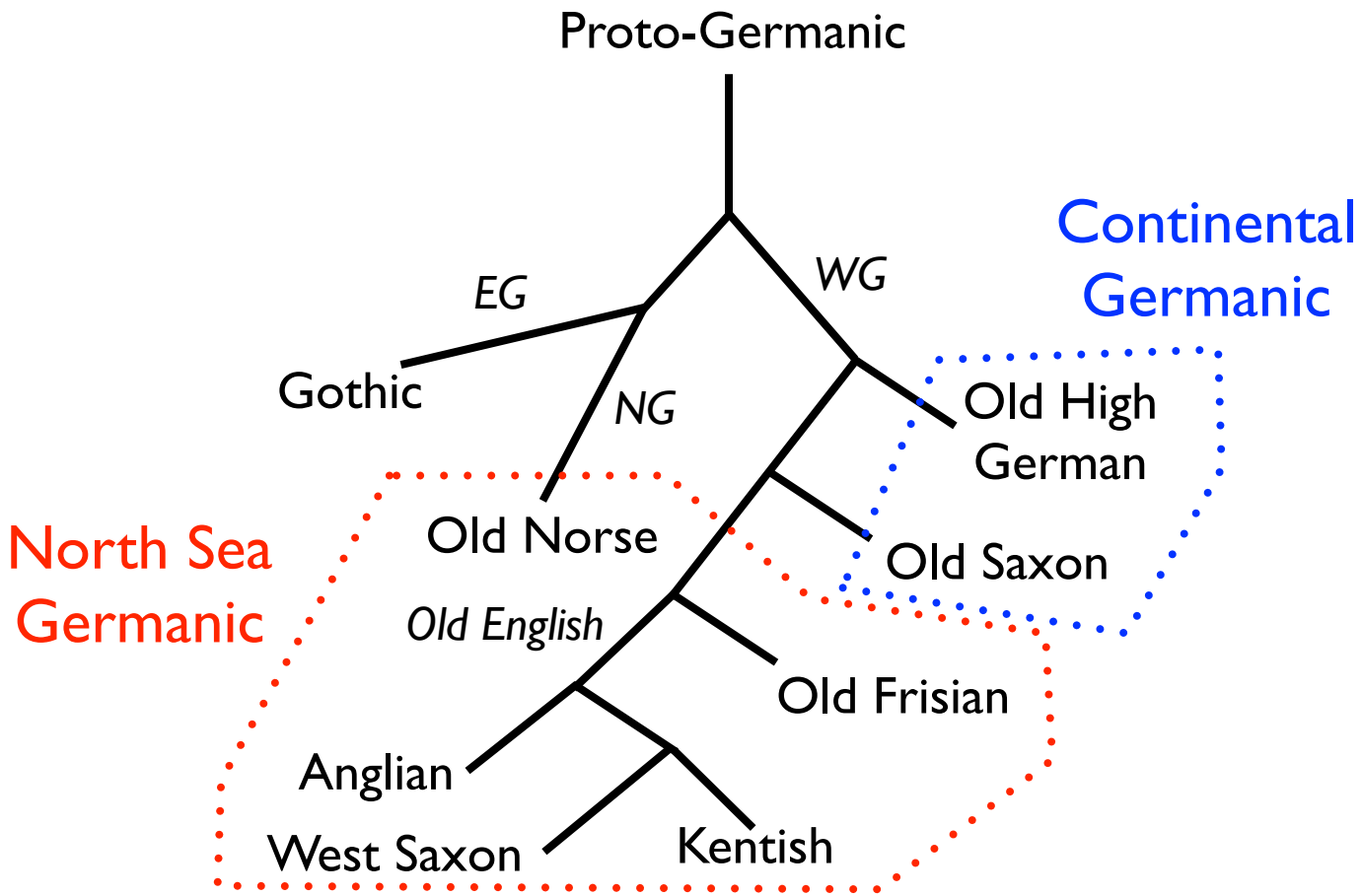




as well as quantitative (lexical) analyses by others

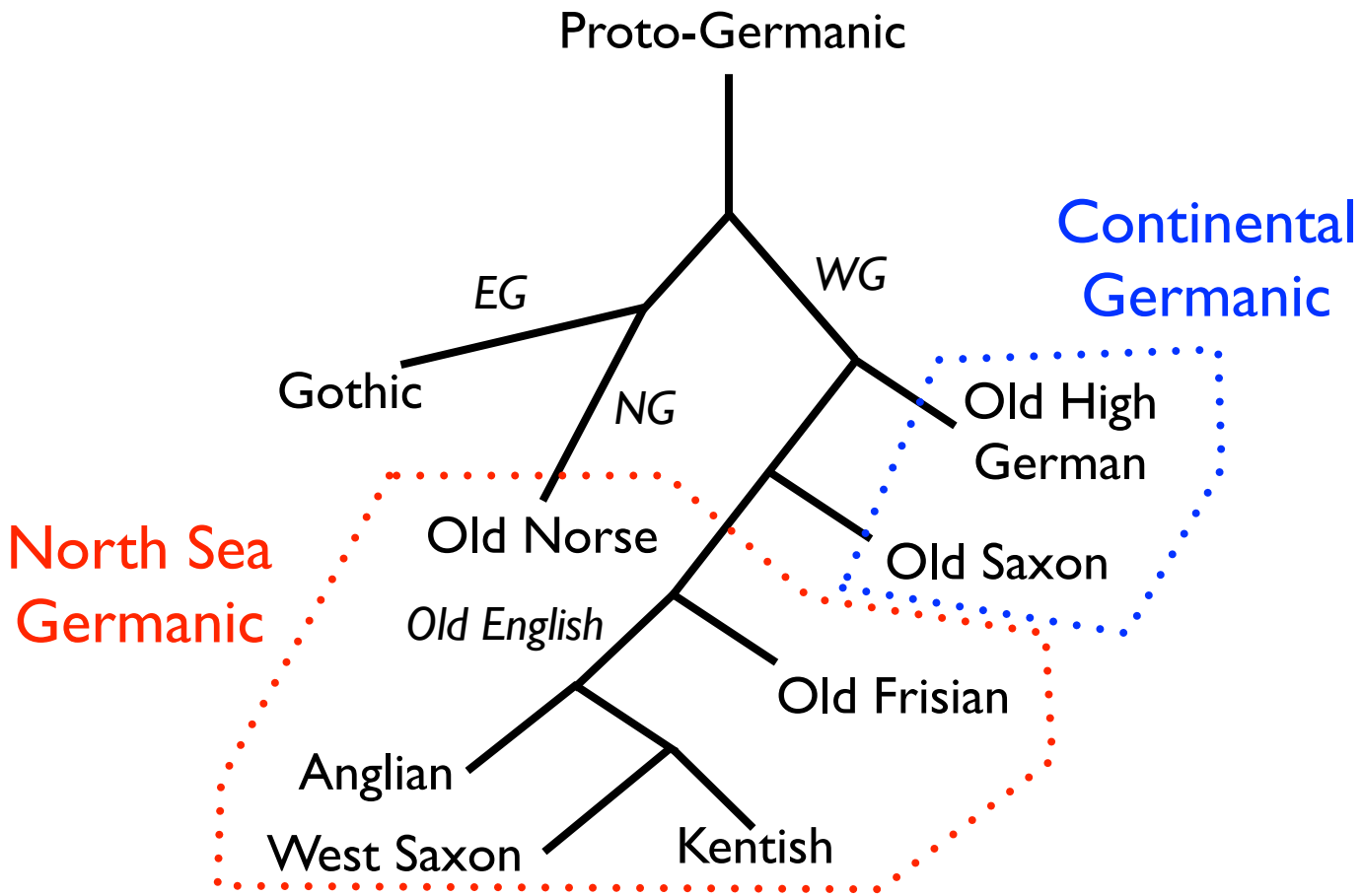
Conclusions

Phonetic, particularly *vocalic*, data emphasise later contact...



Conclusions

Phonetic, particularly *vocalic*, data emphasise later contact...



...criterion to determine breakdown of phylogenic model?

Thanks

Keith Briggs
UWE & BT Research

Dario Spanò
Warwick

Geoff Nicholls
Oxford

Thanks

Keith Briggs
UWE & BT Research

Dario Spanò
Warwick

Geoff Nicholls
Oxford

centre for complexity science

THE UNIVERSITY OF WARWICK

EPSRC
Engineering and Physical Sciences Research Council