



seit 1558

Usage-Based Grammar and Quantitative Corpus Linguistics

Daniel Wiechmann
Summer 2009

Linguistisches Kolloquium - FSU Jena

Nothing is in the mind what is not first in the senses.

(John Locke...somewhere in Essay CHU: Book 2)

- ▶ Corpus ~ Representation of Experience
- ▶ Quantitative Corpus Linguistics and Psycholinguistics (→ Language Comprehension)
- ▶ GOALS of this talk:
 - ▶ Motivate the perspective
 - ▶ Provide some illustration
 - ▶ Point out some issues

*(Some) Key Ideas in
20th Century Linguistic Theorizing*





De Saussure

- Emphasis on the role of analogy
- *It preserves and it alters*

“Any creation must be preceded by an unconscious comparison of the material deposited in the storehouse of language, where productive forms are arranged according to their relations.”

(De Saussure, 1916: 165)





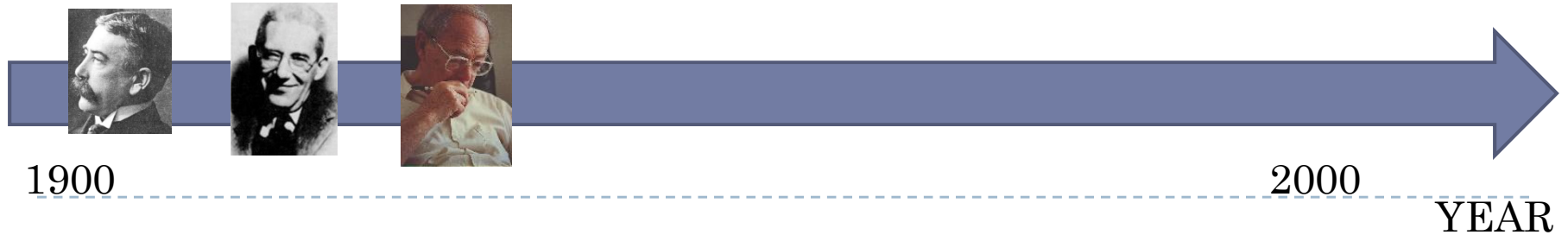
Bloomfield

- ▶ Emphasis on the role of induction
 - ▶ The distilling of general rules or principles from examples

“The only useful generalizations about language are inductive generalizations”

(Bloomfield, 1933:20)



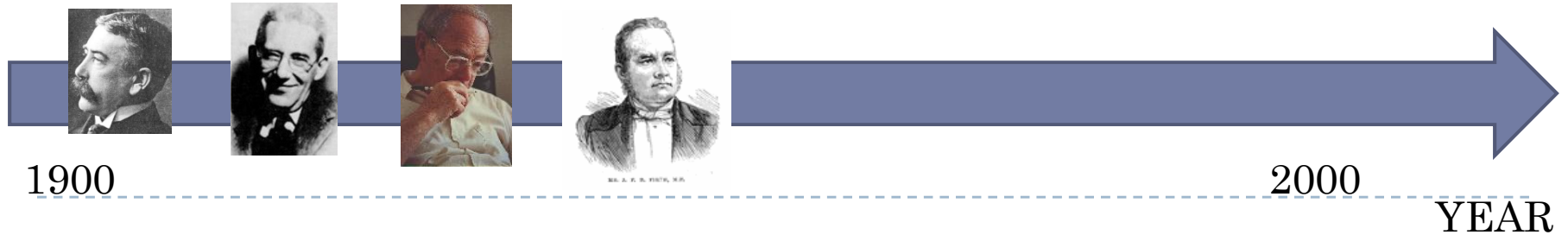


Harris:

- distributional methodology

“[T]he work of linguistics is reducible to establishing correlations. (...) And correlations between the occurrence of one form and that of other forms yield the whole linguistic structure”

(Harris 1940: 704)



Firth

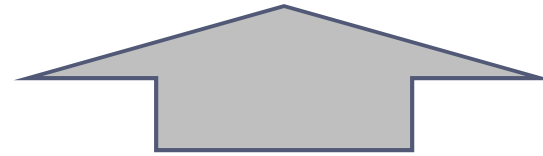
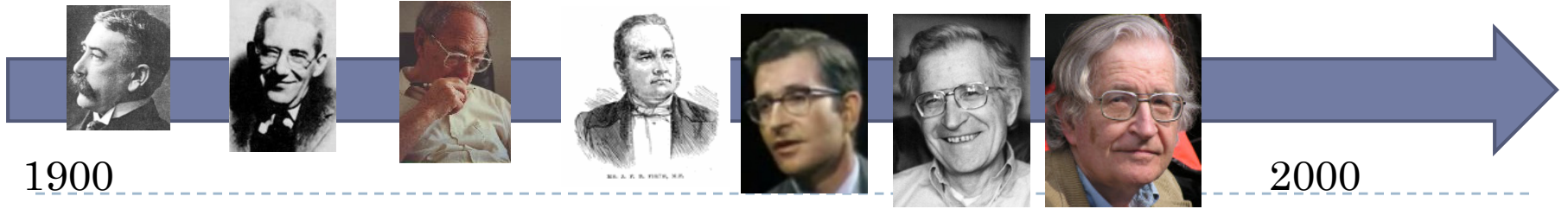
stressed the importance of real-world data

→ central to the development of any model of language collocation models

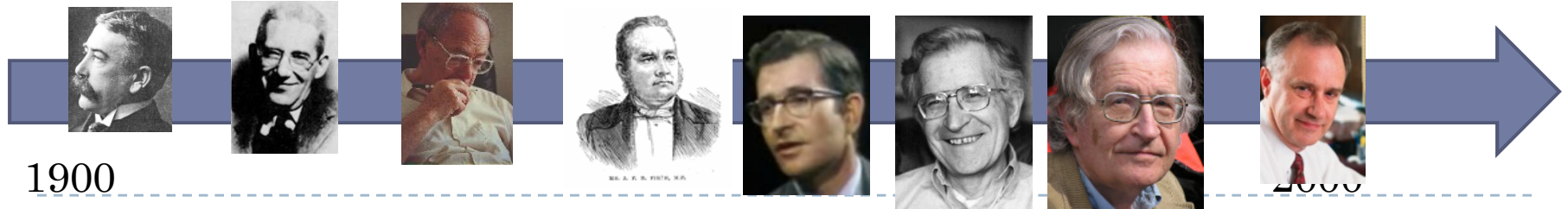
“A theory derives its usefulness and validity from the aggregate of experience to which it must continually refer.”

(Firth 1952: 168)





Rationalist intermezzo



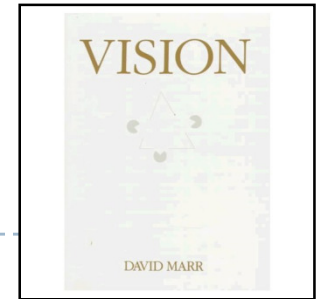
Skousen:

- ▶ proposes operationalization of Harris' ideas
- ▶ Symbolic computational model of analogical processing (Skousen 1989)

- ▶ To predict language behavior and to model language learning, all that is needed is a large database of examples taken directly from language use...
- ▶ ...and a generally applicable method for analogical modeling that is inherently inductive

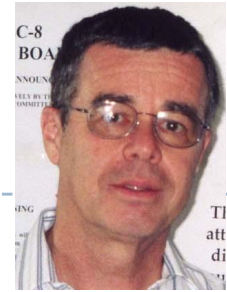


Cognitive Psychology



- ▶ Exemplar-based models are the single most productive model type in psychological domains (Bod 2006)
- ▶ People represent categories by storing individual exemplars in memory (rather than rules, prototypes, or probabilities)
- ▶ Categorization decisions are then based on similarity assessment
 - ▶ Perceived stimulus → stored exemplar

Cognitive Linguistics



- ▶ Usage-based models (Langacker, Bybee, ...) presuppose a ...
 - ▶ Bottom-up
 - ▶ Maximalist
 - ▶ Redundant (patterns and instantiations co-exist)

Interim summary: Key ingredients of a theory of language

- ▶ Stored exemplars
- ▶ Induction
- ▶ Analogy
- ▶ Similarity

- I. Pattern recognition
(precondition)
- II. Concept formation
(induction proper)
- III. Projection
(application)

induction ~ “more-of-the-same” inference
(a pattern is carried forward to new cases)

analogical inference ~ induction

Exemplar-based language processing

21 Century Linguistics: New Heroes

Data Oriented Parsing (DOP)

(Scha, 1992, Bod , 1996)

- Statistical approach to syntactic parsing
- Uses a corpus of parsed utterances as representation of experience
- Frequencies are used to compute probabilities



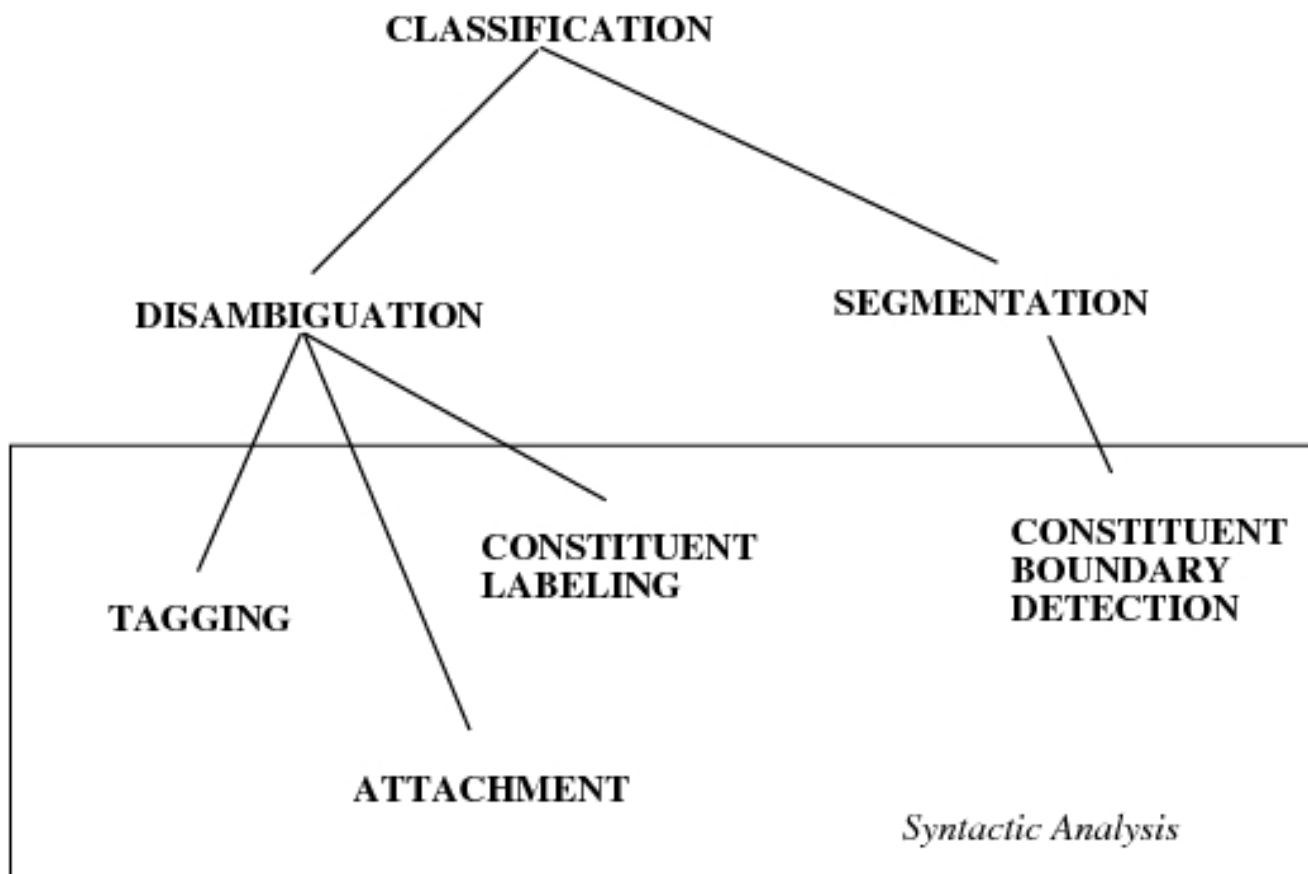
21 Century Linguistics: New Heroes

Memory-based Language Processing (Daelemans, van den Bosch)

- Uses a corpus of parsed utterances as representation of experience
- Similar in spirit, (but maybe) more general than DOP



Exemplar-based linguistics



Data frames

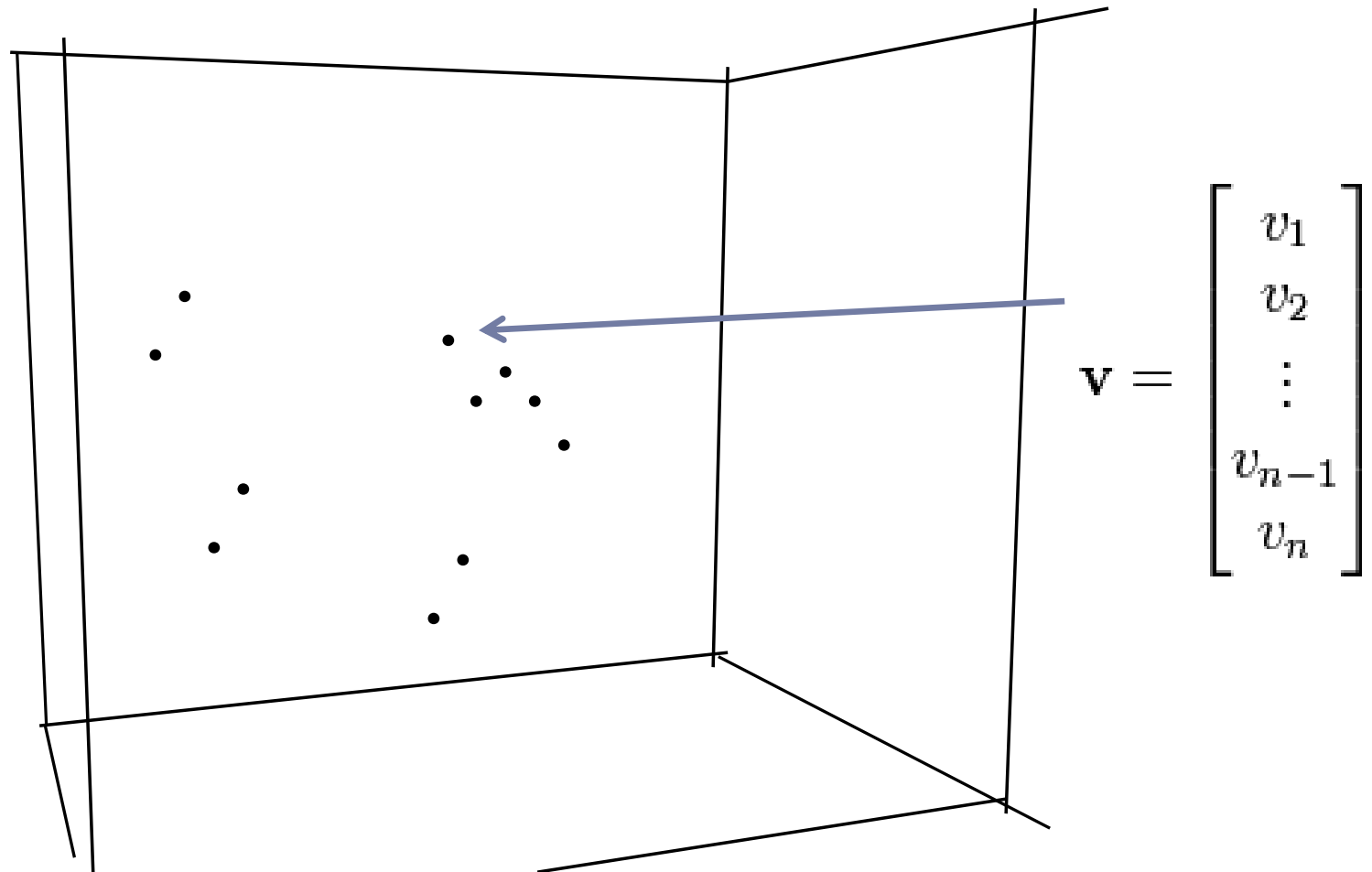
Usage event	string	medium	internal role	transitivity RC	finiteness RC	embedding
1	blablabla	written	subject	arg2	edp	center
2	blablabla	written	subject	arg2	edp	center
3	blablabla	spoken	nonSUBJ	arg2	to.inf	right
4	blablabla	spoken	nonSUBJ	arg2	to.inf	right
5	blablabla	written	subject	arg1	ingp	center
6	blablabla	spoken	nonSUBJ	arg1	to.inf	right
7	blablabla	spoken	nonSUBJ	arg1	to.inf	right
8	blablabla	spoken	subject	arg1	ingp	right
9	blablabla	spoken	nonSUBJ	arg2	to.inf	center
10	blablabla	written	subject	arg2	ingp	center
11	blablabla	spoken	subject	arg1	ingp	center
12	blablabla	written	subject	arg1	ingp	center
13	blablabla	written	subject	arg2	edp	right
14	blablabla	written	subject	arg2	ingp	center
15	blablabla	spoken	subject	arg2	edp	right
16	blablabla	spoken	nonSUBJ	arg1	edp	right
17	blablabla	written	nonSUBJ	arg1	to.inf	right
18	blablabla	written	subject	arg1	edp	center
19	blablabla	written	subject	arg1	ingp	right
20
21
n-1
n	blablabla	spoken	nonSUBJ	arg1	to.inf	center

Data frames, column vectors

Usage event	string	medium	internal role	transitivity RC	finiteness RC	embedding
1	blablabla	written	subject	arg2	edp	center
2	blablabla	written	subject	arg2	edp	center
3	blablabla	spoken	nonSUBJ	arg2	to.inf	right
4	blablabla	spoken	nonSUBJ	arg2	to.inf	right
5	blablabla	written	subject	arg1	ingp	center
6	blablabla	spoken	nonSUBJ	arg1	to.inf	right
7	blablabla	spoken	nonSUBJ	arg1	to.inf	right
8	blablabla	spoken	subject	arg1	ingp	right
9	blablabla	spoken	nonSUBJ	arg2	to.inf	center
10	blablabla	written	subject	arg2	ingp	center
11	blablabla	spoken	subject	arg1	ingp	center
12	blablabla	written	subject	arg1	ingp	center
13	blablabla	written	subject	arg2	edp	right
14	blablabla	written	subject	arg2	ingp	center
15	blablabla	spoken	subject	arg2	edp	right
16	blablabla	spoken	nonSUBJ	arg1	edp	right
17	blablabla	written	nonSUBJ	arg1	to.inf	right
18	blablabla	written	subject	arg1	edp	center
19	blablabla	written	subject	arg1	ingp	right
20
21
n-1
n	blablabla	spoken	nonSUBJ	arg1	to.inf	center

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-1} \\ v_n \end{bmatrix}$$

Constructional/ configural space



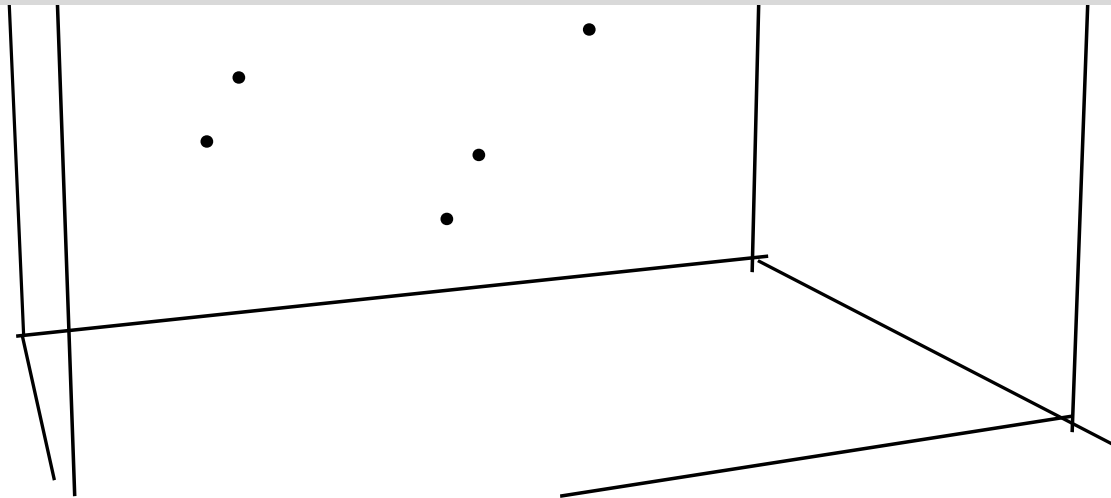
Constructional/ configural space

Nearest neighbor classifiers methods (k -NN classifier)

- examples are stored as points in k -dimensional example space
 - with their values (k -dimensional feature vector)
 - with their class membership

Categorisation

- A new example obtains class by finding its position in example space
- And extrapolating its class from nearest neighbor(s)



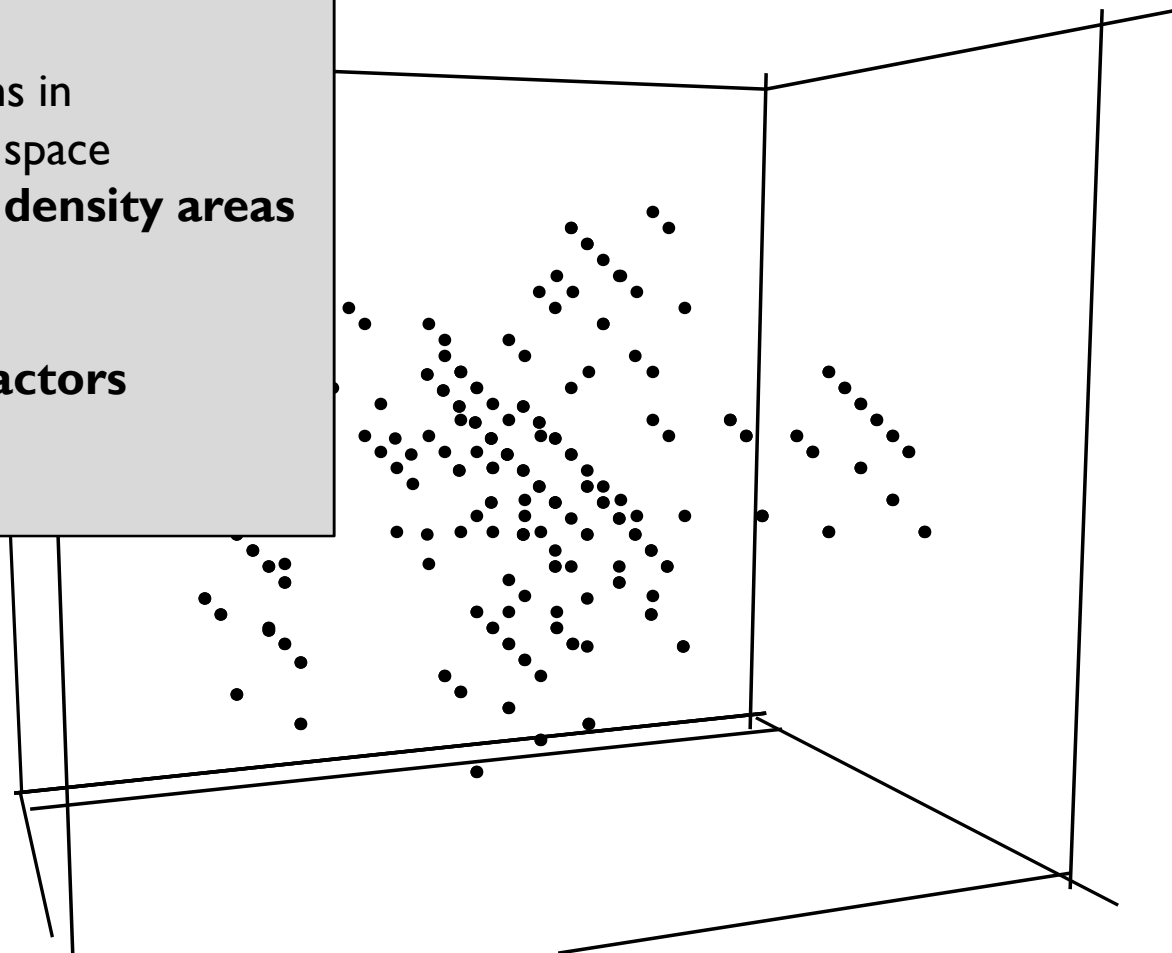
Constructional/ configural space

TASK I:

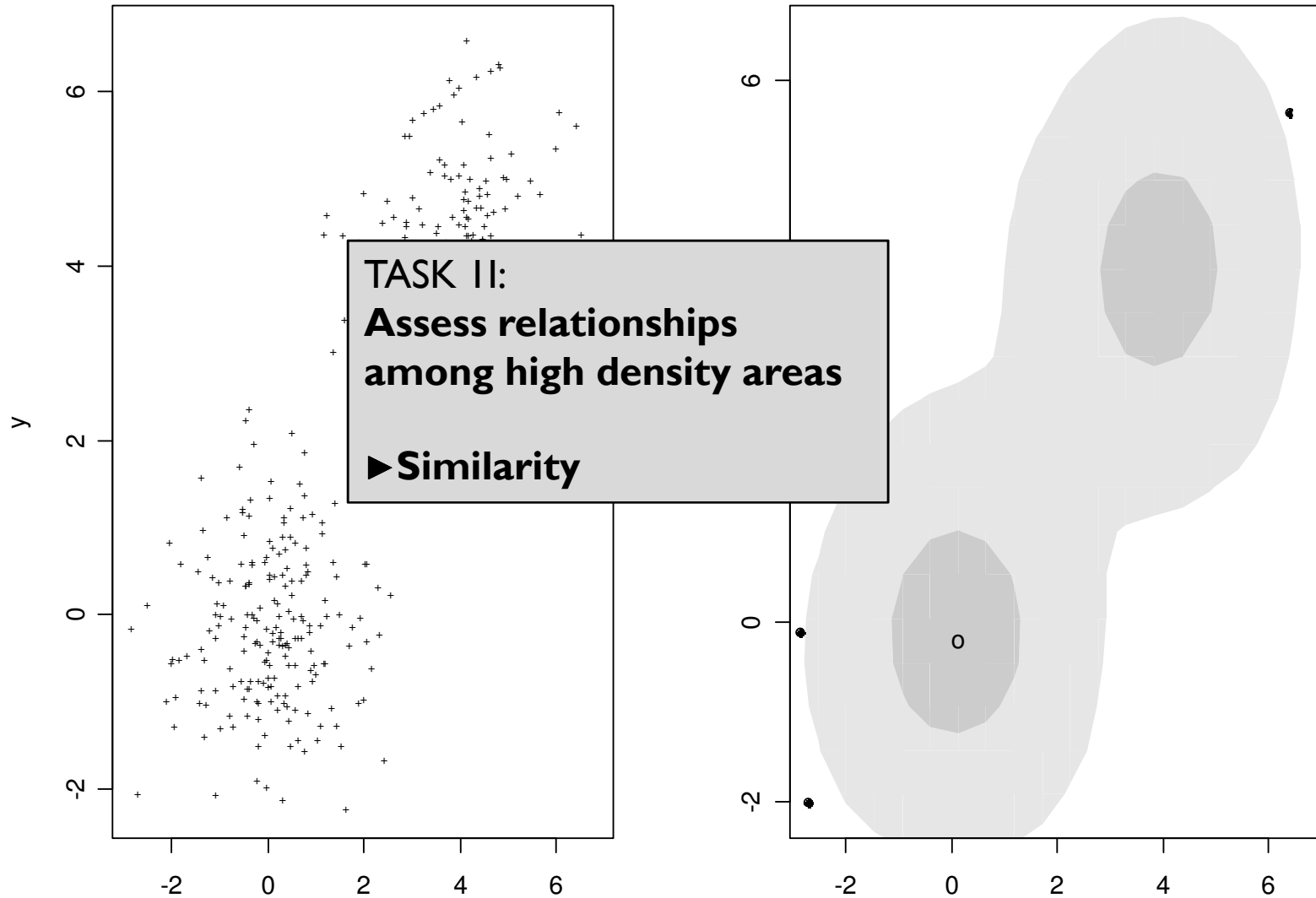
Assign positions in
constructional space

**Identify high density areas
(HDA)**

HDA are **attractors**



High density areas



Similarity assessment


$$\mathbf{v}_i = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-1} \\ v_n \end{bmatrix} \quad \begin{matrix} \longleftrightarrow \\ \longleftrightarrow \\ \longleftrightarrow \\ \longleftrightarrow \end{matrix} \quad \mathbf{v}_j = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-1} \\ v_n \end{bmatrix}$$

Similarity assessment is vector comparison

Similarity: As distance in Euclidean space w/ feature weighting

$$\Delta(X, Y) = \sum_{i=1}^n \omega_i d(x_i, y_i)$$

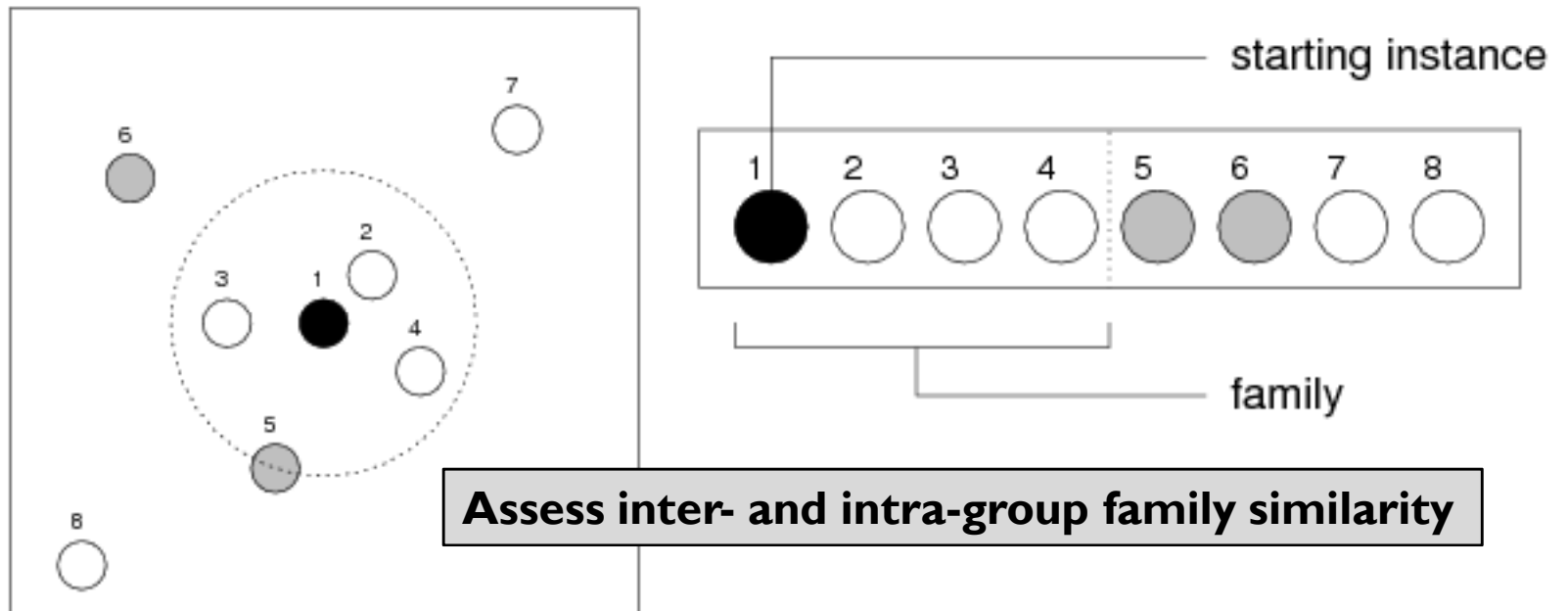
feature weighting
via chi square statistic



$$\chi^2 = \sum_i \sum_j \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

O_{ij} is the observed number of cases with value v_i in class c_j
(mutatis mutandis E_{ij})

Grouping

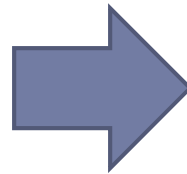


Statistical Analysis

Detect prominent families (schemas)

(hierarchical) **configural frequency analysis**

K-optimal pattern detection (**association rule mining**)



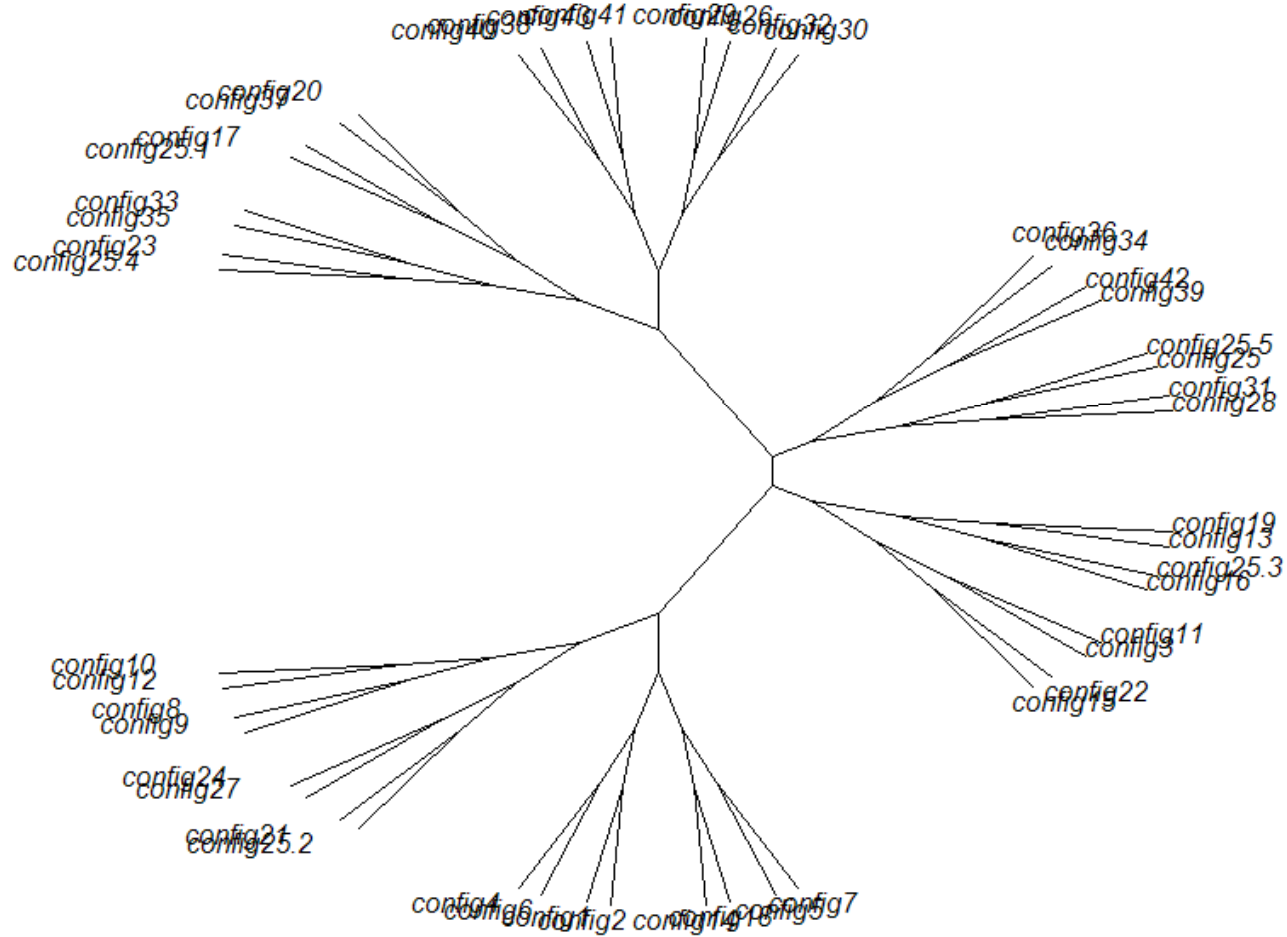
cluster analysis*

Similarity network of RCC

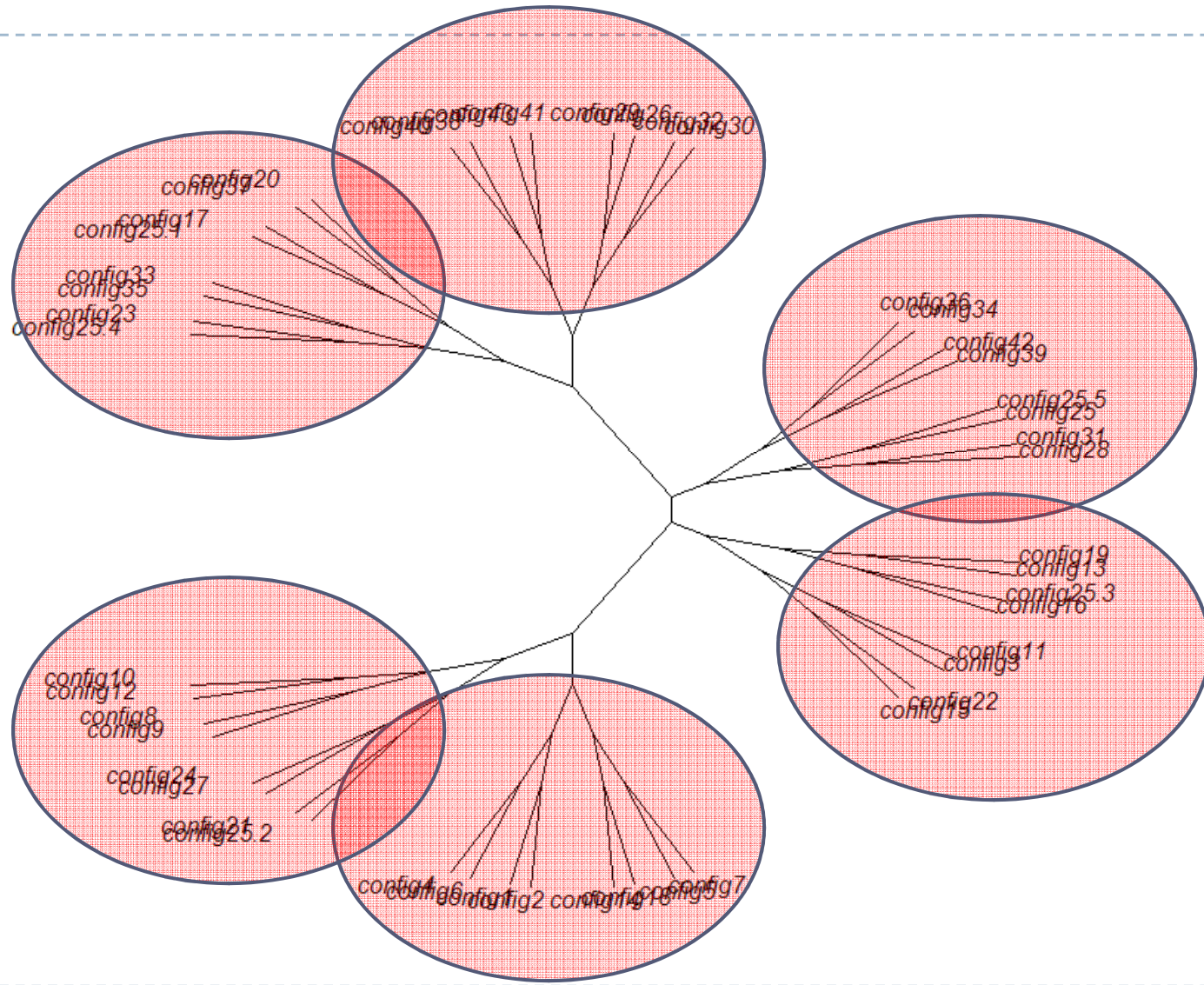
*node joining algorithm to produce unrooted phylogenetic trees



Statistical Analysis



Statistical Analysis



Thank you !