Today's topic:

- Computational methods for subgrouping (2)

  Methods from the biological sciences

Can we automate?

- What are some aspects of research in historical linguistics that we could, in principle, automate with computer software?

Can we automate?

- Can we use computers to automate the search for **sound correspondence sets** and sound-change rules by "lining up" the corresponding sounds from all the words in a cognate set?

- Appealing idea, but we're not there yet
  - Similar to the ideas behind **genome comparison** in biology/evolutionary science

Can we automate?

- Can we use computers to automate the process of **subgrouping** within a group of languages already known to be related?
  - This is a much more promising area of research at present

- BUT: The various warnings and points to watch our for in doing subgrouping by hand are still relevant when doing subgrouping by computer
  - Be a cautious consumer when reading reports of this kind of resesarch

Advantages

- Using computational methods in subgrouping may make it easier to...
  - work with large numbers of languages
  - generate and compare multiple hypotheses about subgrouping

Considerations

- Recall our class discussion about subgrouping in Proto-Gazelle-Peninsula
  - What was the trickiest problem we encountered?
  - How did we solve it?
  - What are the implications for programming a software method for subgrouping?

Implementation

- Many computational techniques in historical linguistics have been adapted from biology
  - This does not mean we are using biological information to subgroup languages
  - It's just that the computer programs were originally written by/for biologists (which is relevant because their terminology is used)

- Terminology
  - Taxon, clade
  - "Characters" and coding

Taxon and clade

- **Taxon** — the basic unit of comparison (can be an individual language or a previously proposed group/clade)

- **Clade** — a subgroup

Characters and coding

- Any property that is used to compare across the set of languages is called a **character**
  - A shared cognate could be a character
  - A sound correspondence could be one
  - Etc. — The analyst must choose appropriate or trustworthy characters to code
- Methods for **coding** characters
  - Binary ('has/doesn't have')
  - Multistate ('has option A/B/C...')
- The coded data can then be analyzed with subgrouping software

PIE subgrouping from "% cognates"

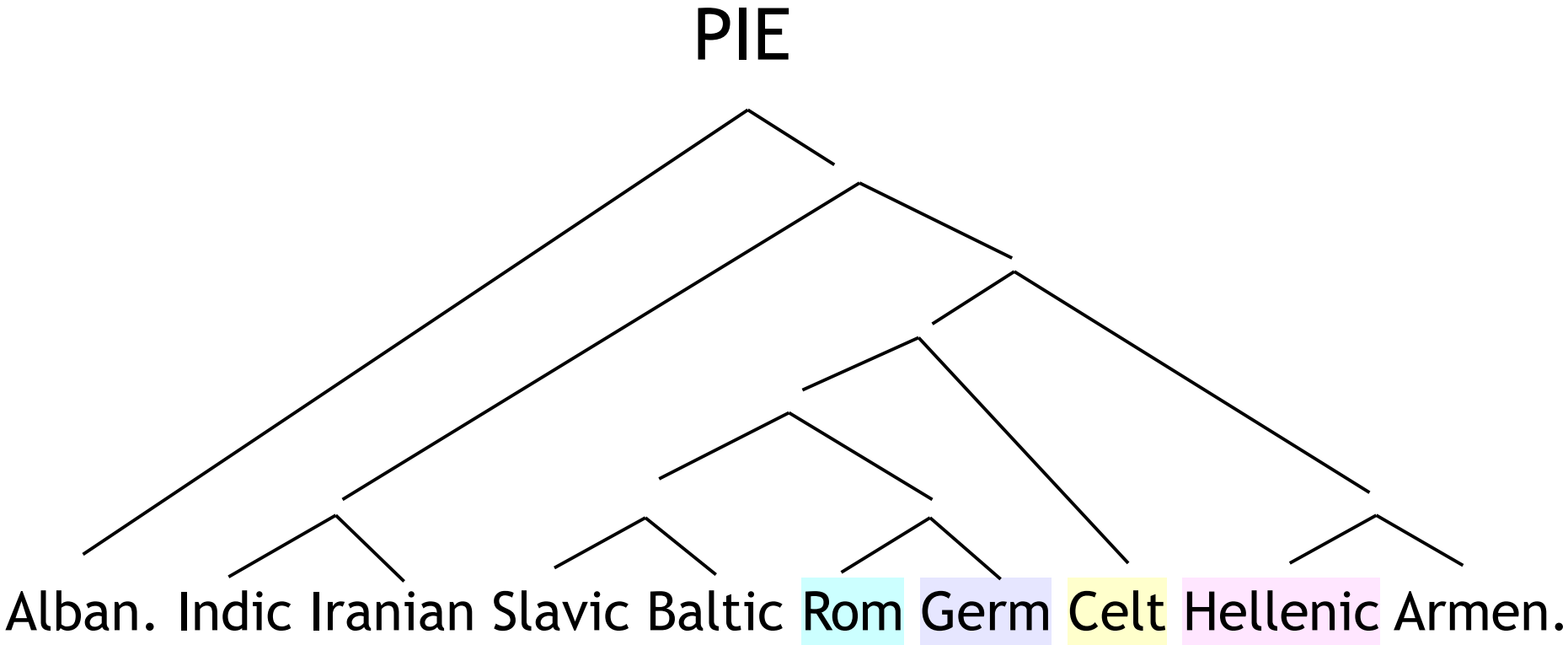
Excerpted from Johnson (2008) discussion, based on Dyen, Kruskal, & Black (1992)

- DKB took Swadesh-200 lists from 84 Indo-European languages and determined how many cognates were shared between each pair
  - How much can we trust these values? (How well is the history of IE known?)
- Method: "single-linkage clustering"
  - When group (A+B) is compared with C, compare the *highest* %-shared of either A or B with C [what alternatives might there be?]

PIE subgrouping from "% cognates"

Excerpted from Johnson (2008) discussion, based on Dyen, Kruskal, & Black (1992)

- Top-level results: (good match with traditional results)



PIE subgrouping from "spelling distance"

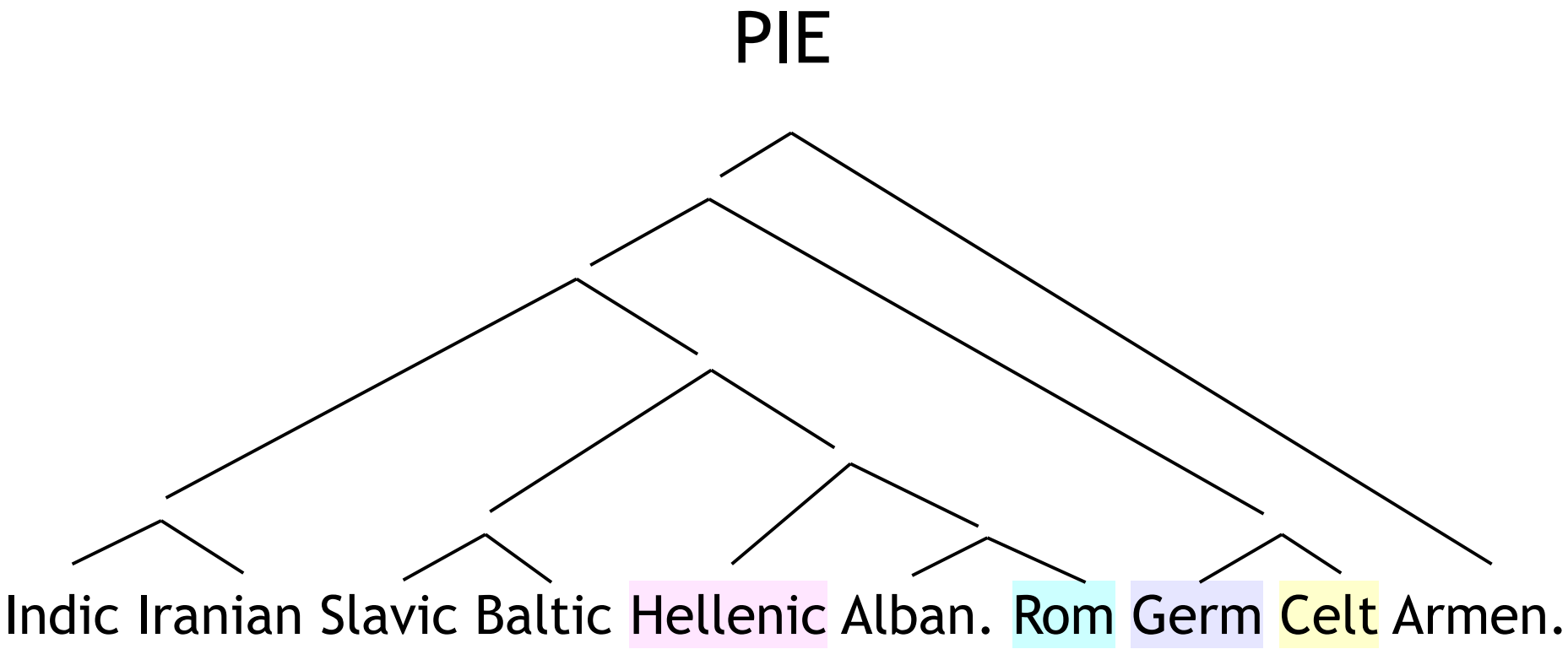
Excerpted from Johnson (2008) discussion

- Proposed as an alternative to try when little about the history of a language group is known
- Calculate degree of "phonetic similarity" between languages and group them into trees on this basis
- For this analysis: J uses the same set of IE words as was used above, but this time calculates only "degree of phonetic similarity" (as represented in *spellings* of words)
  - Comments?

PIE subgrouping from “spelling distance”

Excerpted from Johnson (2008) discussion

- Successful in getting to these top-level groups
- But, some discrepancies here with the previous tree (what might be some reasons?)



An example from the literature

- Class discussion of excerpts from an article on Indo-European subgrouping:

  Peter Forster and Alfred Toth. 2003. Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proceedings of the National Academy of Sciences* 100(15): 9079-9084.

- As historical linguists, how would you evaluate the claims from this article?

- For a more detailed critique, see the links available from this Language Log post