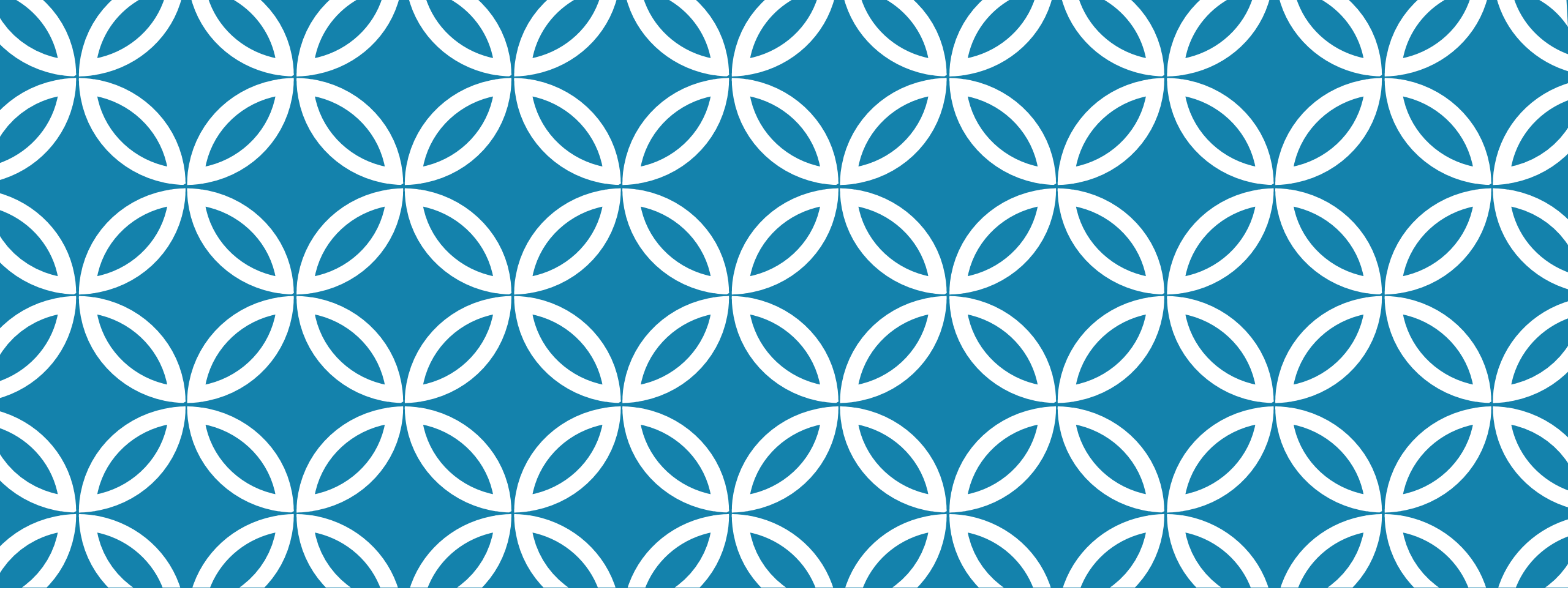
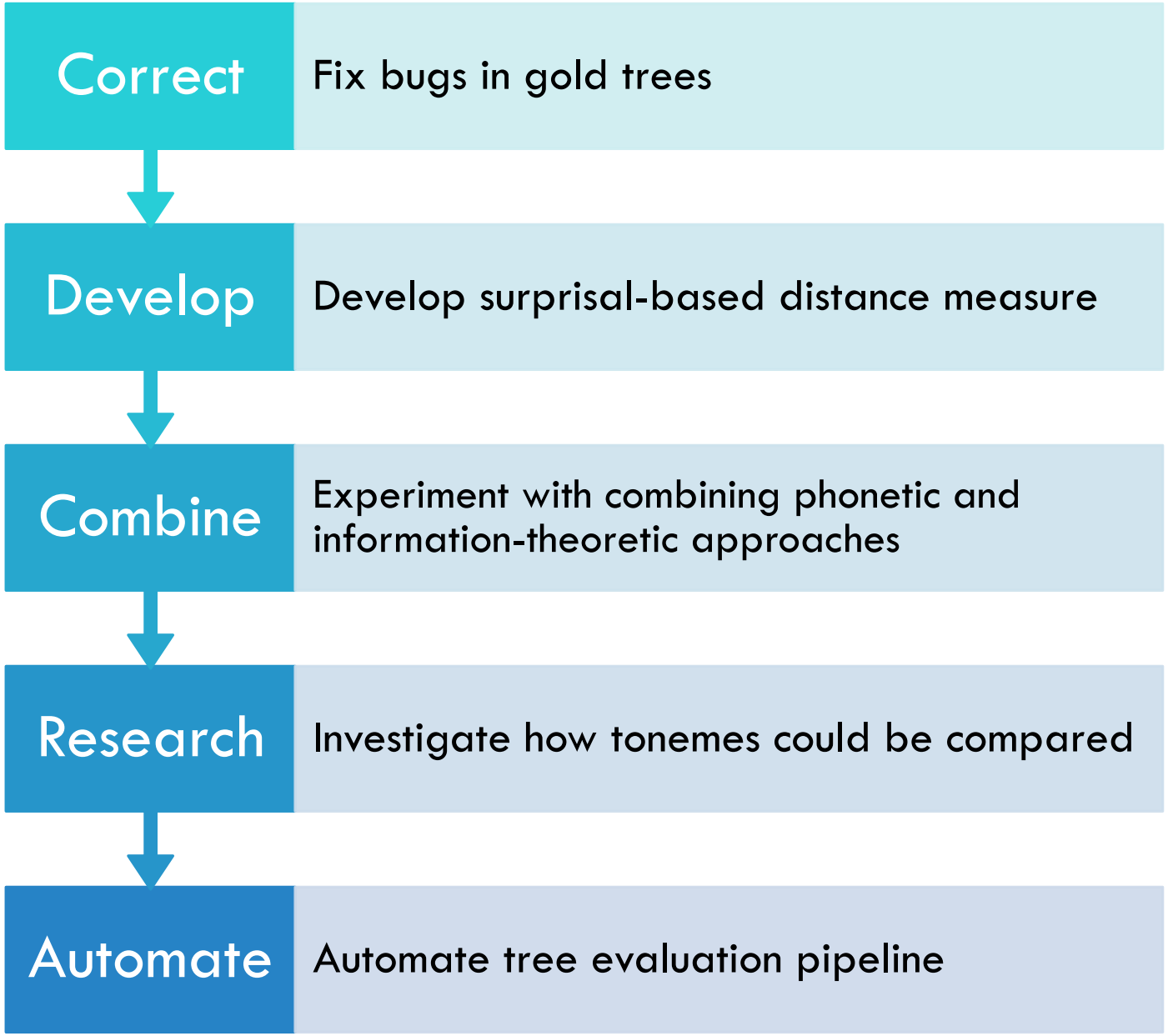


MASTER'S THESIS MEETING

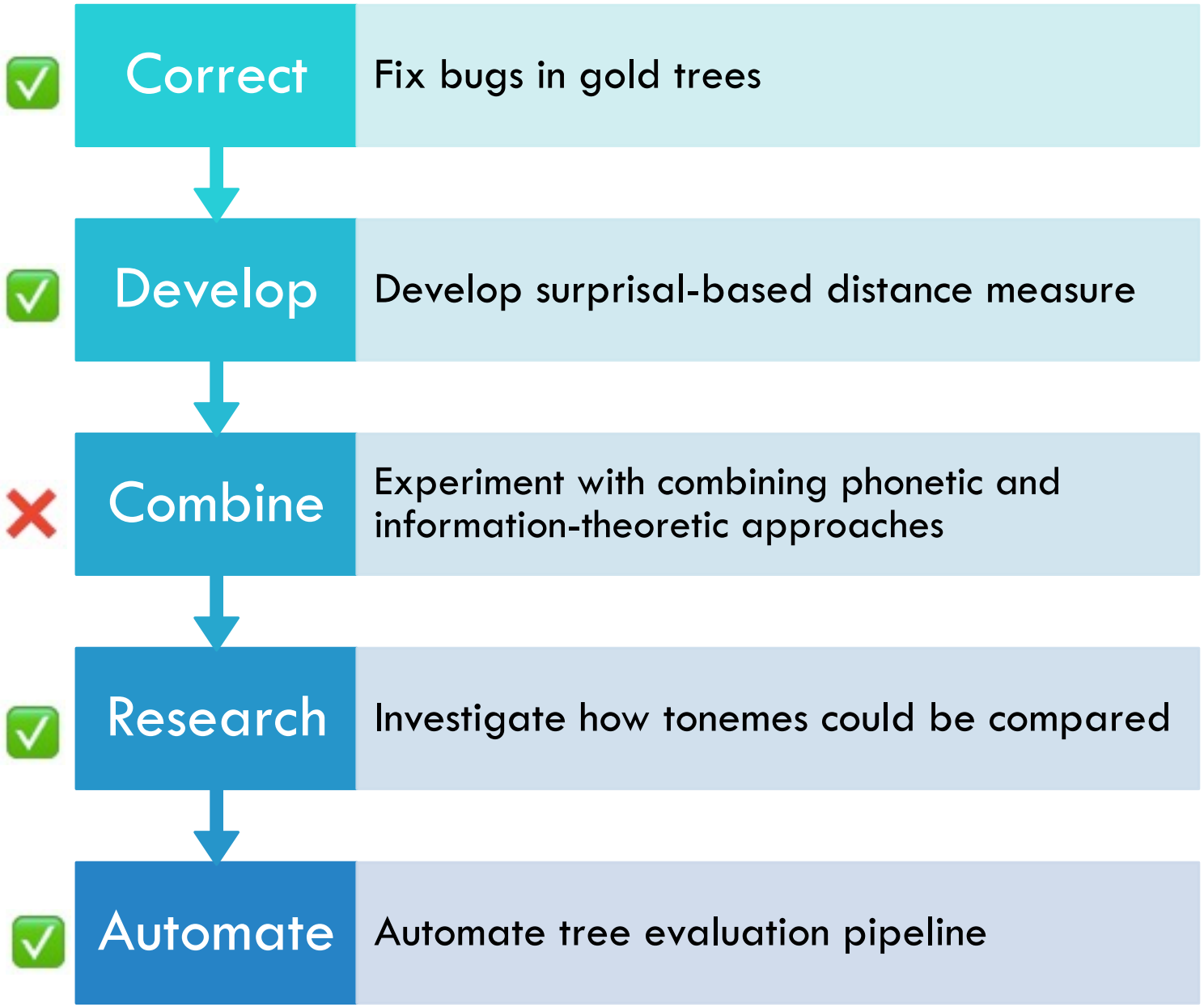
Philip Georgis

September 29, 2021

CURRENT TASKS



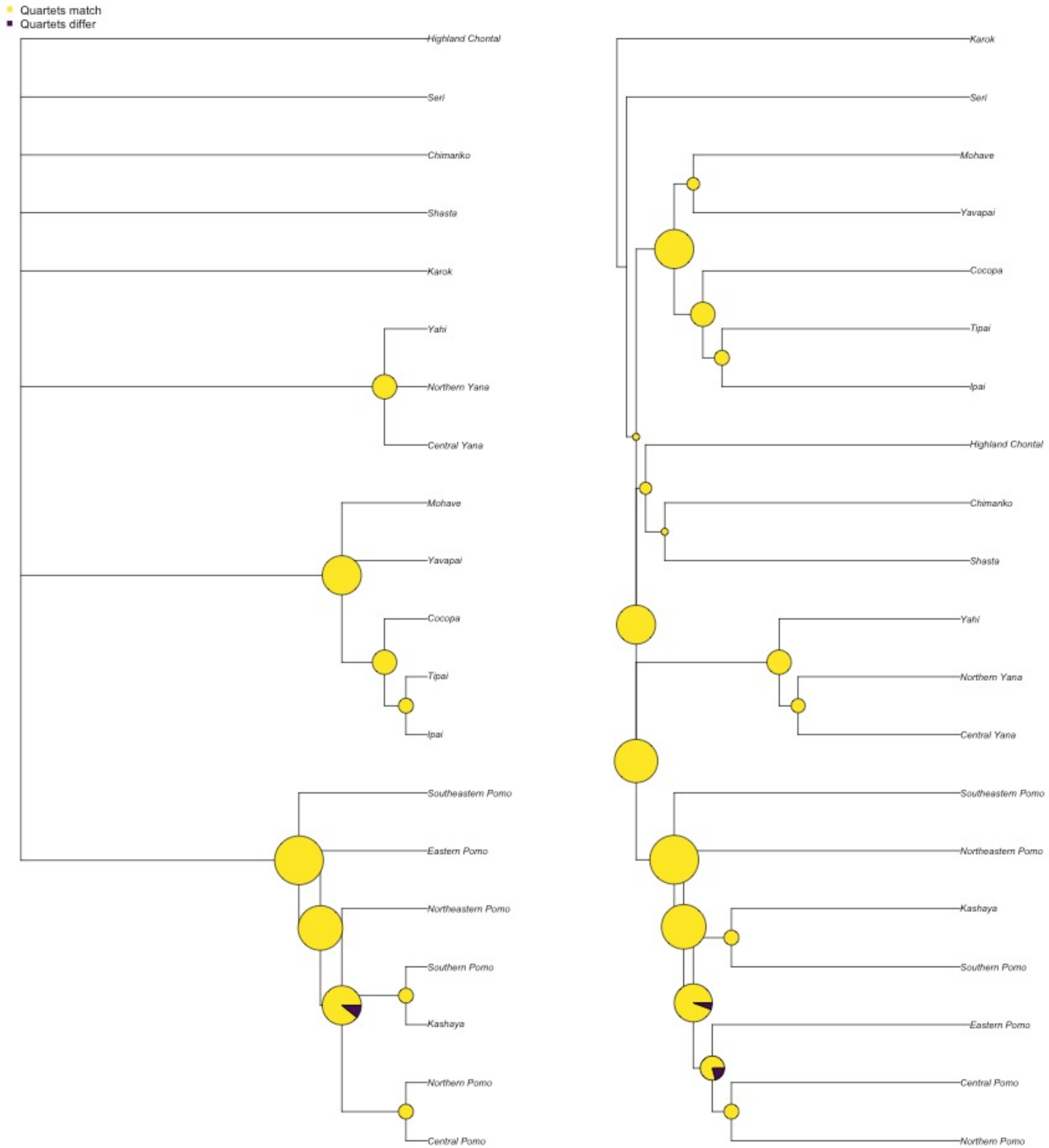
CURRENT TASKS



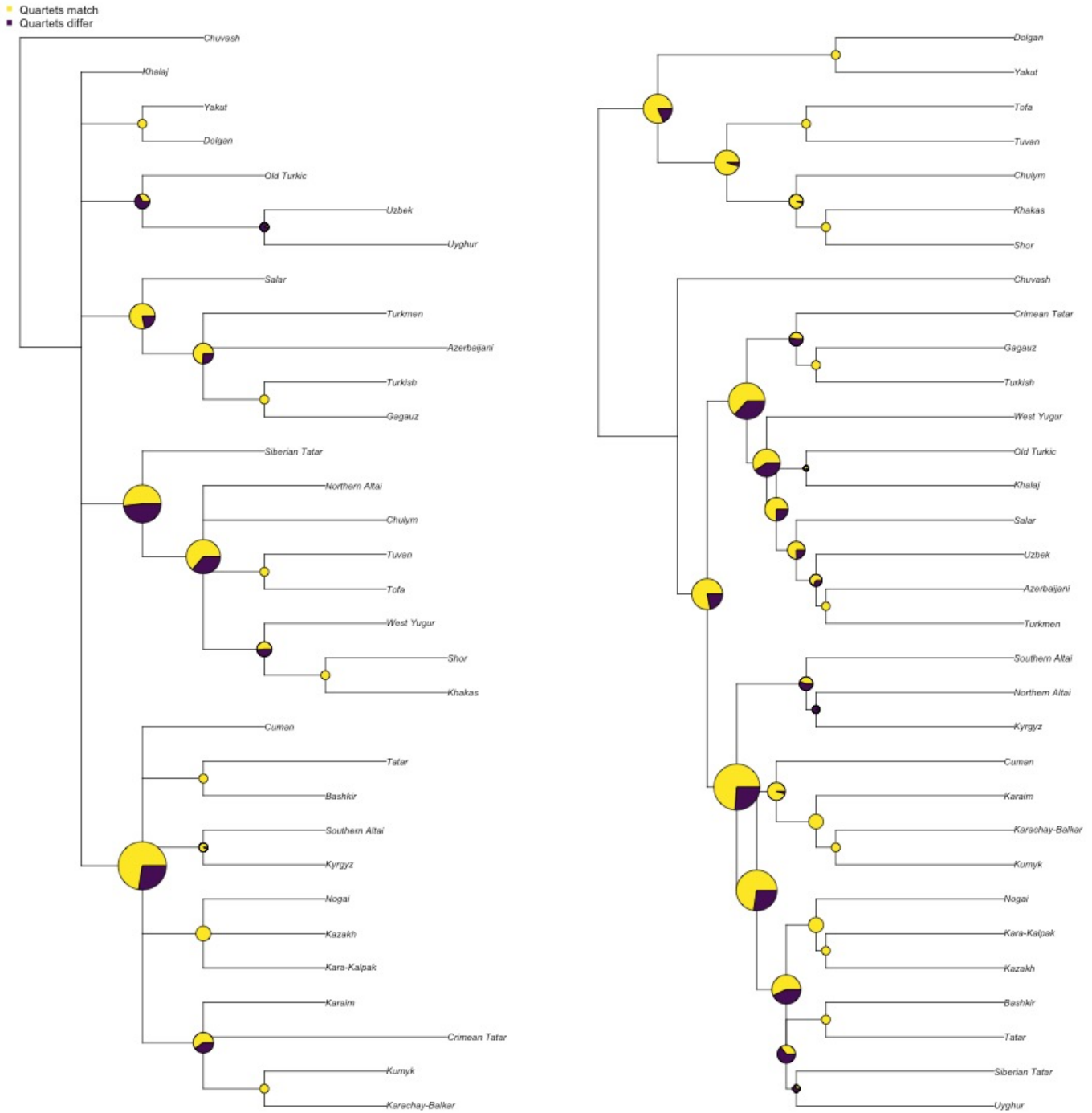
PRUNED GLOTTOLOG TREE BUG FIXES

- Arabic ✅
  - Modern Standard Arabic missing from tree → simple bug in code
- Italic ✅
  - Piedmontese and Ligurian: certain doculects mapped to specific Glottolog dialects, others to the whole language because couldn't find the dialect → slightly off tree structure
  - Piedmontese: all remapped to Glottolog's Turinese dialect, except Vercellese → Low Piedmontese
  - Ligurian: all remapped to Glottolog's Genoese dialect
- Turkic ✅
  - Baraba Tatar: ambiguous doculect, "Baraba" appears twice in Glottolog Turkic tree so it was extracted in both positions in pruned gold tree
  - Either dialect of **Siberian Tatar (Central Siberian Turkic branch)**, or of Tatar (North Kipchak branch)

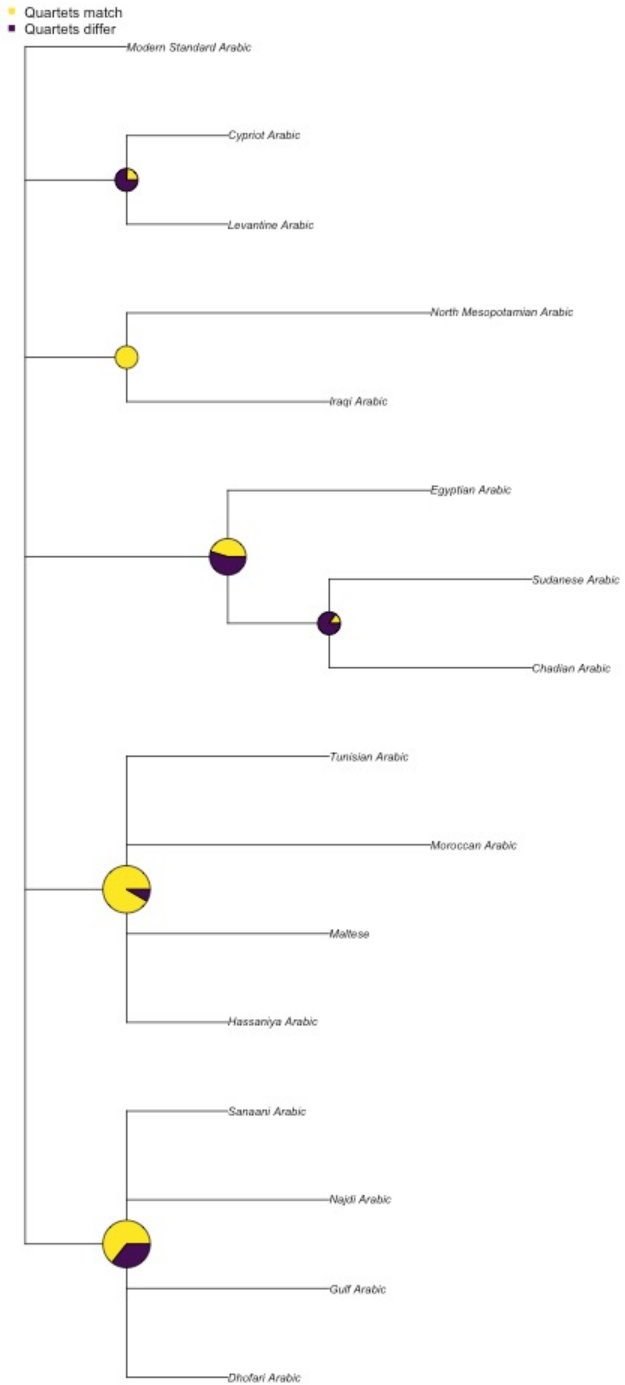
Hokan

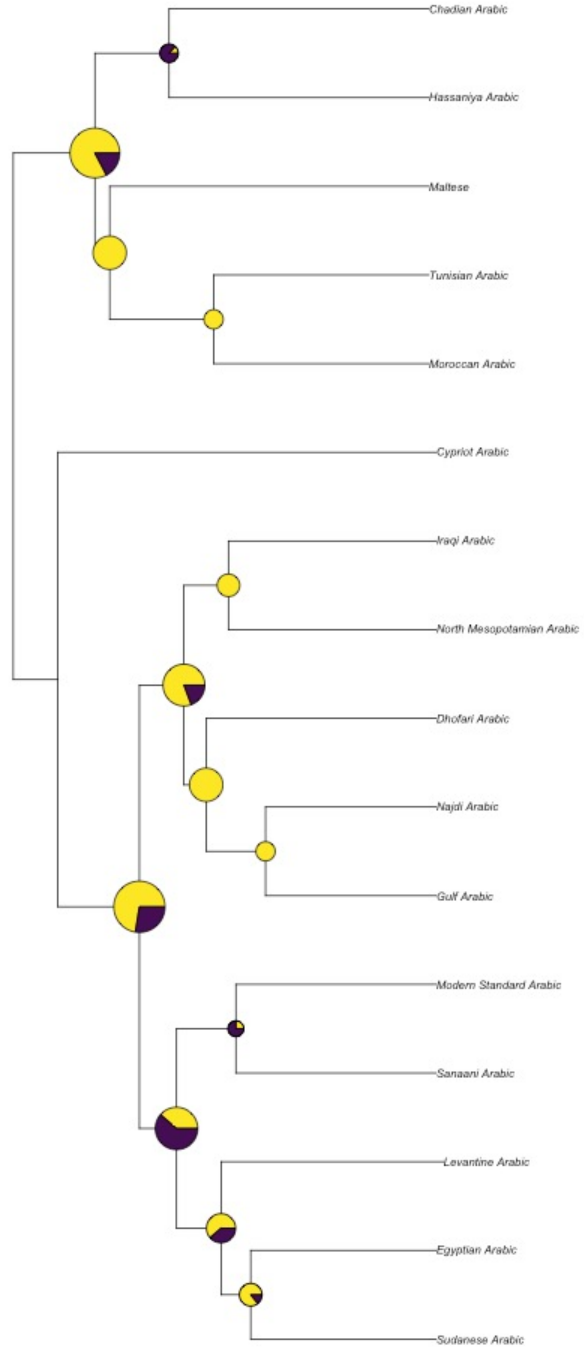

Quartets match
Quartets differ
Highland Chontal
Seri
Chimariko
Shasta
Karok
Yahi
Northern Yana
Central Yana
Mohave
Yavapai
Cocopa
Tipai
Ipai
Southeastern Pomo
Eastern Pomo
Northeastern Pomo
Southern Pomo
Kashaya
Northern Pomo
Central Pomo
Karok
Seri
Mohave
Yavapai
Cocopa
Tipai
Ipai
Highland Chontal
Chimariko
Shasta
Yahi
Northern Yana
Central Yana
Southeastern Pomo
Northeastern Pomo
Kashaya
Southern Pomo
Eastern Pomo
Central Pomo
Northern Pomo

Turkic


Quartets match
Quartets differ
Chuvash
Khalaj
Yakut
Dolgan
Old Turkic
Uzbek
Uyghur
Salar
Turkmen
Azerbaijani
Turkish
Gagauz
Siberian Tatar
Northern Altai
Chulym
Tuvan
Tofa
West Yugur
Shor
Khakas
Cuman
Tatar
Bashkir
Southern Altai
Kyrgyz
Nogai
Kazakh
Kara-Kalpak
Karaim
Crimean Tatar
Kumyk
Karachay-Balkar
Dolgan
Yakut
Tofa
Tuvan
Chulym
Khakas
Shor
Chuvash
Crimean Tatar
Gagauz
Turkish
West Yugur
Old Turkic
Khalaj
Salar
Uzbek
Azerbaijani
Turkmen
Southern Altai
Northern Altai
Kyrgyz
Cuman
Karaim
Karachay-Balkar
Kumyk
Nogai
Kara-Kalpak
Kazakh
Bashkir
Tatar
Siberian Tatar
Uyghur

Arabic


Quartets match
Quartets differ
Modern Standard Arabic
Cypriot Arabic
Levantine Arabic
North Mesopotamian Arabic
Iraqi Arabic
Egyptian Arabic
Sudanese Arabic
Chadian Arabic
Tunisian Arabic
Moroccan Arabic
Maltese
Hassaniya Arabic
Sanaani Arabic
Najdi Arabic
Gulf Arabic
Dhofari Arabic


Chadian Arabic
Hassaniya Arabic
Maltese
Tunisian Arabic
Moroccan Arabic
Cypriot Arabic
Iraqi Arabic
North Mesopotamian Arabic
Dhofari Arabic
Najdi Arabic
Gulf Arabic
Modern Standard Arabic
Sanaani Arabic
Levantine Arabic
Egyptian Arabic
Sudanese Arabic

BEST AUTOMATIC TREES (NOT USING GOLD COGNATE SETS)

Dataset	TreeDist	Quartet Divergence	Cognate Clustering Method	Word Form Evaluation Method	Linkage Method
Arabic	0.67	0.39	PMI	PMI	Ward
Balto-Slavic	0.19	0.09	PMI	PMI	complete
Dravidian	0.41	0.22	none	Phonetic	average
Hokan	0.34	0.18	none	PMI	average
Italic	0.36	0.19	none	Phonetic	Ward
Polynesian	0.61	0.28	PMI	PMI	complete
Sinitic	0.55	0.28	Phonetic	Phonetic	Ward
Turkic	0.57	0.39	none	PMI	Ward
Uralic	0.32	0.20	none	Phonetic	complete

BEST AUTOMATIC TREES (NOT USING GOLD COGNATE SETS)

Dataset	TreeDist	Quartet Divergence	Cognate Clustering Method	Word Form Evaluation Method	Linkage Method
Arabic	0.67	0.39	PMI	PMI	Ward
Balto-Slavic	0.19	0.09	PMI	PMI	complete
Dravidian	0.41	0.22	none	Phonetic	average
Hokan	0.34	0.18	none	PMI	average
Italic	0.36	0.19	none	Phonetic	Ward
Polynesian	0.61	0.28	PMI	PMI	complete
Sinitic	0.55	0.28	Phonetic	Phonetic	Ward
Turkic	0.57	0.39	none	PMI	Ward
Uralic	0.32	0.20	none	Phonetic	complete

Notes:

- The best tree for 5/9 datasets was yielded without any cognate clustering
- About evenly split in terms of PMI vs. phonetic as most successful evaluation method
- Tree Distance seems much more sensitive than Quartet Divergence (e.g. Arabic/Turkic, Polynesian/Sinitic)

GOLD TREES? ALTERNATIVE TREES?

- Glottolog trees remain unresolved in many places, or else represent only one view in cases of debated/unclear classifications
- e.g. Turkic
  - Baraba Tatar: Central Siberian Turkic or Kipchak?
    - Glottolog classifies it as Central Siberian Turkic
    - Savelyev & Robbeets (2020) [from which Turkic data were taken] found evidence for Kipchak classification; matches my results also
  - North and South Altai
    - Glottolog: North Altai is Central Siberian Turkic but South Altai is Kipchak
    - Savelyev & Robbeets (2020): classified together in same branch
  - Crimean Tatar: Kipchak or Oghuz?
    - Traditionally classified as Kipchak, Glottolog classifies it this way
    - But southern coastal dialect considered to belong to Oghuz branch → both Savelyev & Robbeets (2020) and my results point to this classification
  - Khalaj: unclear, no consensus

GOLD TREES? ALTERNATIVE TREES?

- Glottolog trees remain unresolved in many places, or else represent only one view in cases of debated/unclear classifications

- Turkic
  - Robeets & Savelyev's (2020) Turkic tree only achieves a slightly better Tree Distance (0.55) with respect to Glottolog tree than does mine (0.59)
  - My tree which most closely matches the resolved R&S tree has a Tree Distance = 0.35

- Polynesian
  - Benedict Kind & Mary Walworth's resolved Oceanic and Polynesian trees
  - Oceanic tree has Tree Distance = 0.53 from Glottolog tree
  - My Polynesian results more closely match Oceanic tree (TD = 0.47) than Glottolog tree (TD = 0.61)

TONAL FEATURES

Added 7 distinctive features for tonal segments (tonemes) according to Wang (1967)

TABLE I

TONES AND THEIR FEATURES

	1	2	3	4	5	6	7	8	9	10	11	12	13
CONTOUR	−	−	−	−	−	+	+	+	+	+	+	+	+
HIGH	+	−	+	−	−	+	−	+	−	+	−	+	−
CENTRAL	−	−	+	+	+	−	−	−	−	−	−	−	−
MID	−	−	−	−	+	−	−	−	−	−	−	−	−
RISING	−	−	−	−	−	+	+	−	−	+	+	+	+
FALLING	−	−	−	−	−	−	−	+	+	+	+	+	+
CONVEX	−	−	−	−	−	−	−	−	−	−	−	+	+

TONAL FEATURES

Results:

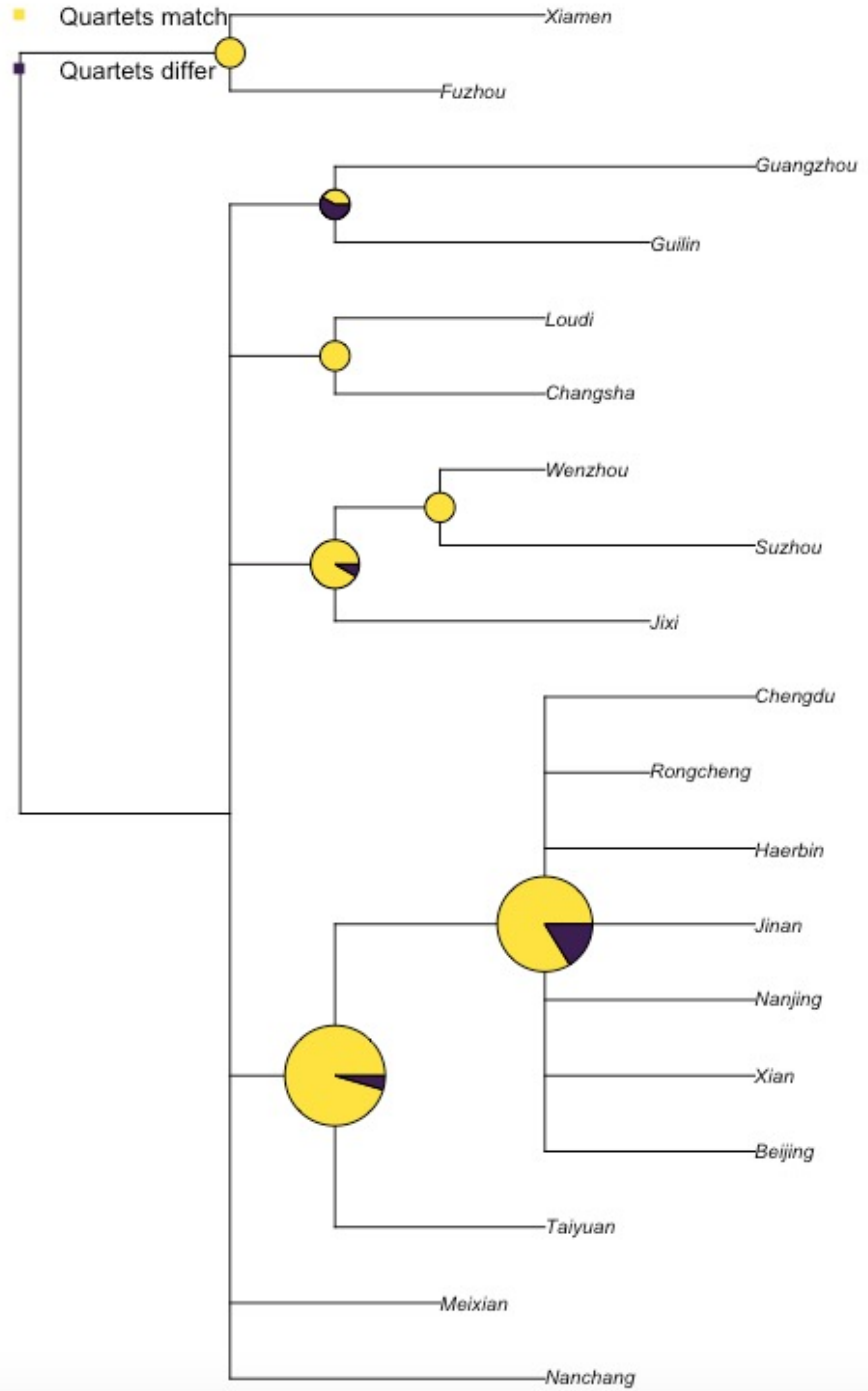
- little impact on accuracy of cognate clustering
- BUT improved the Sinitic tree!

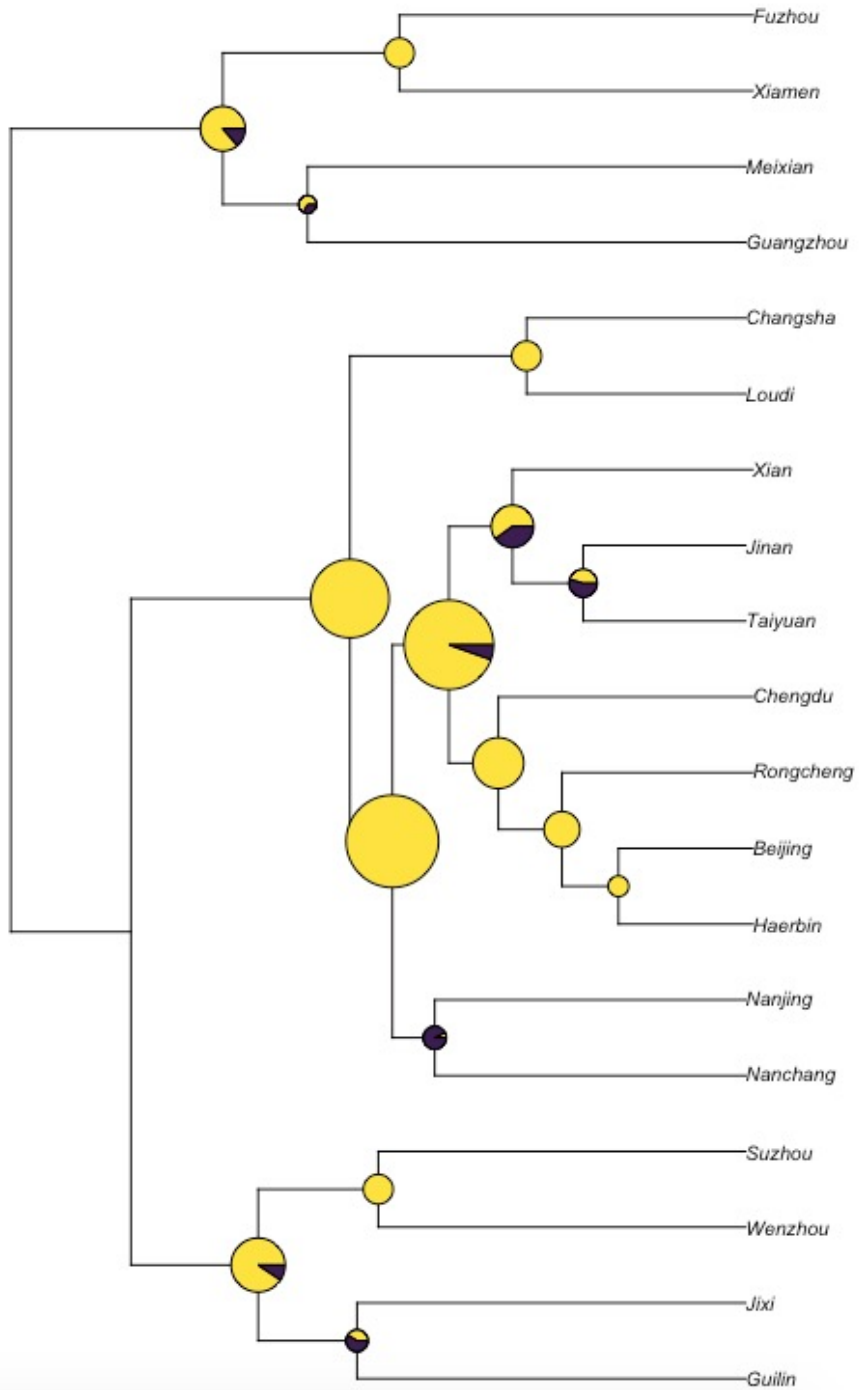
Best tree *without* use of tonal features had TreeDist of 0.65 and Quartet Divergence of 0.36

→ but *with* tonal features this was reduced to TD = 0.55, QD = 0.28

Phonetic method also outperforms PMI for tree accuracy once tones are added

Sinitic

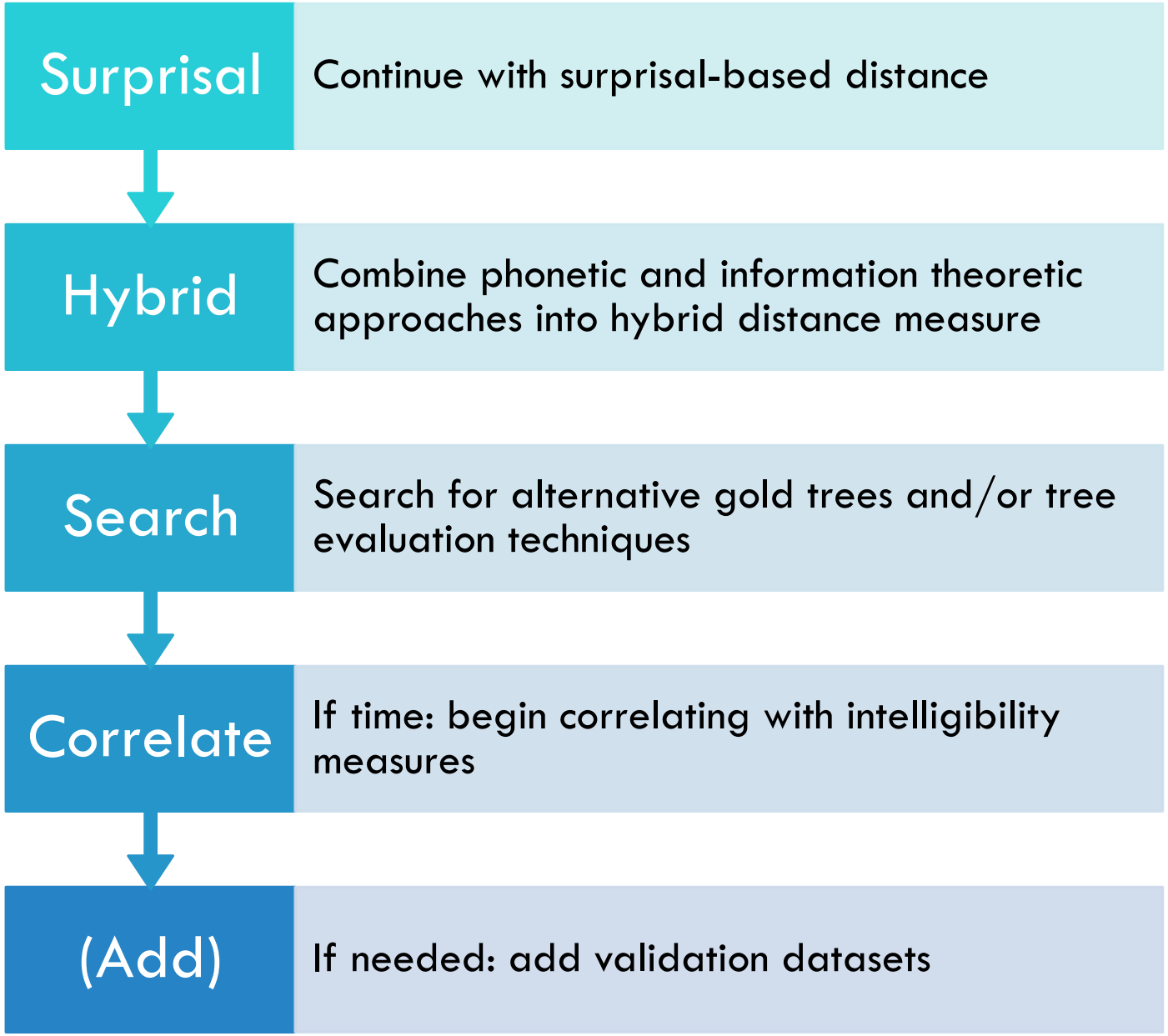

Quartets match
Quartets differ
Xiamen
Fuzhou
Guangzhou
Guilin
Loudi
Changsha
Wenzhou
Suzhou
Jixi
Chengdu
Rongcheng
Haerbin
Jinan
Nanjing
Xian
Beijing
Taiyuan
Meixian
Nanchang


Fuzhou
Xiamen
Meixian
Guangzhou
Changsha
Loudi
Xian
Jinan
Taiyuan
Chengdu
Rongcheng
Beijing
Haerbin
Nanjing
Nanchang
Suzhou
Wenzhou
Jixi
Guilin

VALIDATION DATASETS?

- Is it problematic to use the same datasets for parameter tuning and testing results?
- Could designate validation datasets specifically for parameter tuning (i.e. clustering cutoff thresholds)
- Possible available datasets (not yet examined in detail, but have gold cognate coding)
  - Athabaskan
  - Hellenic
  - Pama-Nyungan
  - Uto-Aztecan
  - Vietic

NEXT TASKS


Surprisal
Continue with surprisal-based distance
Hybrid
Combine phonetic and information theoretic approaches into hybrid distance measure
Search
Search for alternative gold trees and/or tree evaluation techniques
Correlate
If time: begin correlating with intelligibility measures
(Add)
If needed: add validation datasets