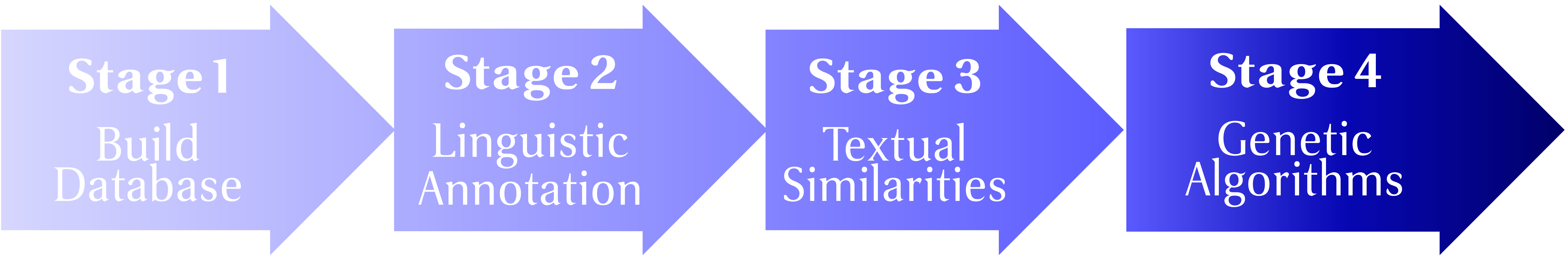
European Research Council

Intelligent Agents and Genetic Algorithms for Tibetan and Chinese Tagging and Alignment

Christopher Handy & Marieke Meelen
c.a.handy@hum.leidenuniv.nl - mm986@cam.ac.uk

Universiteit Leiden The Netherlands
UNIVERSITY OF CAMBRIDGE



Stage 1 Build Database

Stage 2 Linguistic Annotation

Stage 3 Textual Similarities

Stage 4 Genetic Algorithms

Procedure to find similar witnesses in Tibetan & Chinese Buddhist text collections

Stage 1 - Building the 'Open Philology' Project Database

Challenge 1: Measure the similarity of two (cross-linguistic) textual witnesses.
Challenge 2: Retrieve most similar matches in a large collection of Buddhist *sūtra*s.

Digitising Texts
& Dictionaries
Manual Alignment

'Alignment' = the most-similar match (in form & meaning) of verses or entire textual witnesses in: Tibetan-Tibetan/Chinese-Chinese/Tibetan-Chinese.

Creating a Knowledge Base

PostgreSQL database with Django web framework to manually add & edit possible alignments:

- digitised texts in Classical Chinese/Tibetan
- English/German/Japanese translations
- (bi-lingual) dictionaries
- manually aligned material, see [1]

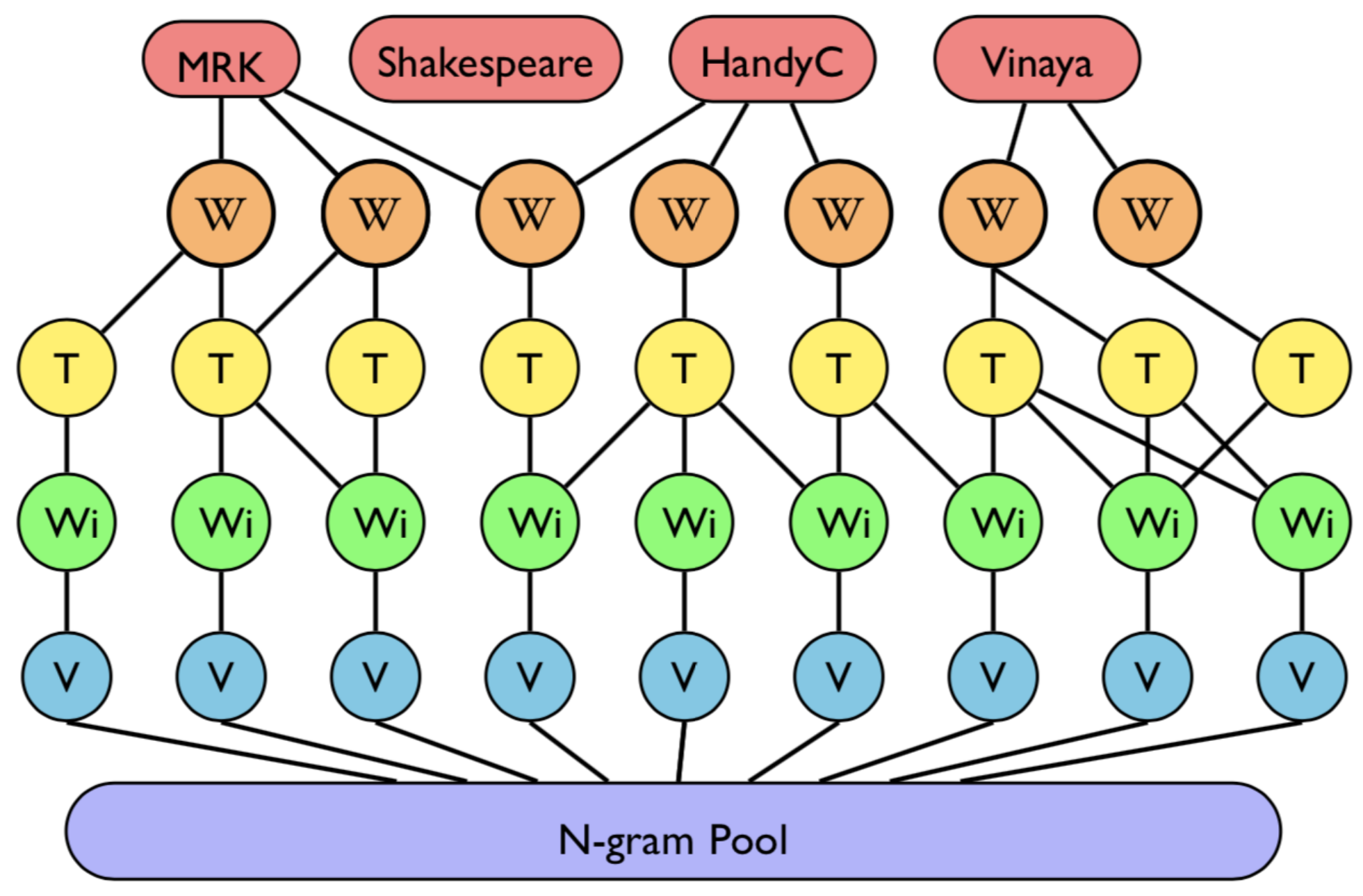


Fig. 1 - Sample structure of textual history of collections

Stage 2 - Adding and Optimising Linguistic Annotation

Classical Chinese and Classical Tibetan are **very low-resource** languages. Many manuscripts containing crucial Witnesses have not been digitised yet, and furthermore:

- Lack segmentation
- Lack clear "word" definition
- Differ in grammar (Chinese vs Tibetan)
- Lack good (bilingual) resources

Segmentation
Tagging & Parsing
Sentence Identification

To address these issues, we are adding the following linguistic annotation:

- Sentence identification based on verbs
- Corrected segmentation
- Corrected POS tagging (including NER)
- Rule-based chunkparsing

For Classical Tibetan, see [2] & [3]; for Classical Chinese, see [4] & [5].

Stage 3 - Retrieving and Measuring (Semantic) Textual Similarities

We developed & tested three methods to retrieve alignments & measure their similarity:
Method 1. Cross-linguistic Information Retrieval & Semantic Textual Similarity with cosine similarity metric for sentence/verse embeddings, see [6].
Method 2. Classic & Statistical Machine Translation combined with search for most similar match measured by BLEU & NIST metrics, see [7].
Method 3. N-gram matching of syllable sequences, solving some remaining issues with these low-resource languages - see this poster ⇒

Freq.	Witness 1	Rank	Witness 2	Freq.
8	恒河上言	1	*bcom ldan 'das kyis bka' stsal pa*	12
4	世尊告言	2	*gang gA'i mchog gis gsol pa*	5
17	恒河上	3	*gang gA'i mchog*	45

Freq.	Witness 1	Rank	Witness 2	Freq.
8	恒河上言	1	恒河上言	9
4	世尊告言	2	河上言	9
17	恒河上	3	恒河	47
8	河上言	4	佛言	46

Matching & Measuring N-gram Pairs

Step 1. Identify language & Extract frequent n-grams as phrase candidates, e.g. "Gaṅgottarā says" (Chi. Rank 1 / Tib. *gang gA'i mchog gis gsol pa*).

Step 2. Identify 'known' alignments in Witnesses from **knowledge base** & (for Tib-Tib/Chi-Chi) String Matching of potential alignments or (for Tib-Chi) Compare linguistic features, e.g. POS/parse.

Cosine Similarity
N-gram Matches

Step 3. Score alignments (measure similarity):
- Minimum Edit & Levenshtein distances [8]
- Ranked Out-of-Place distance (see tables) [9]
- Modified N-gram and R-precision metrics [10]

Step 4. Calculate scores per Witness pair:
- Add score weights for appropriate features
- Normalise overall alignment scores
- Create heatmaps highlighting highly similar matches across Witness pairs (see Fig. 3 below)

Stage 4 - Intelligent Agents & Genetic Algorithms to Optimise Results

To speed up computing time and to allow for testing of a wide variety of variables and scoring systems, a population of agents that operate as independent virtual machines is created to execute each of the tasks in Stages 1, 2 and 3 above in parallel. After extrinsic evaluation, high scoring agents are retained, copied and mutated to create a new population. Over successive generations (see Fig. 2), agents evolve toward ideal alignments, producing increasingly accurate verse matches across source texts with variant verse readings, non-standard spellings, and grammatical peculiarities. These tiny programs can also be manually tuned to focus on specific tasks.

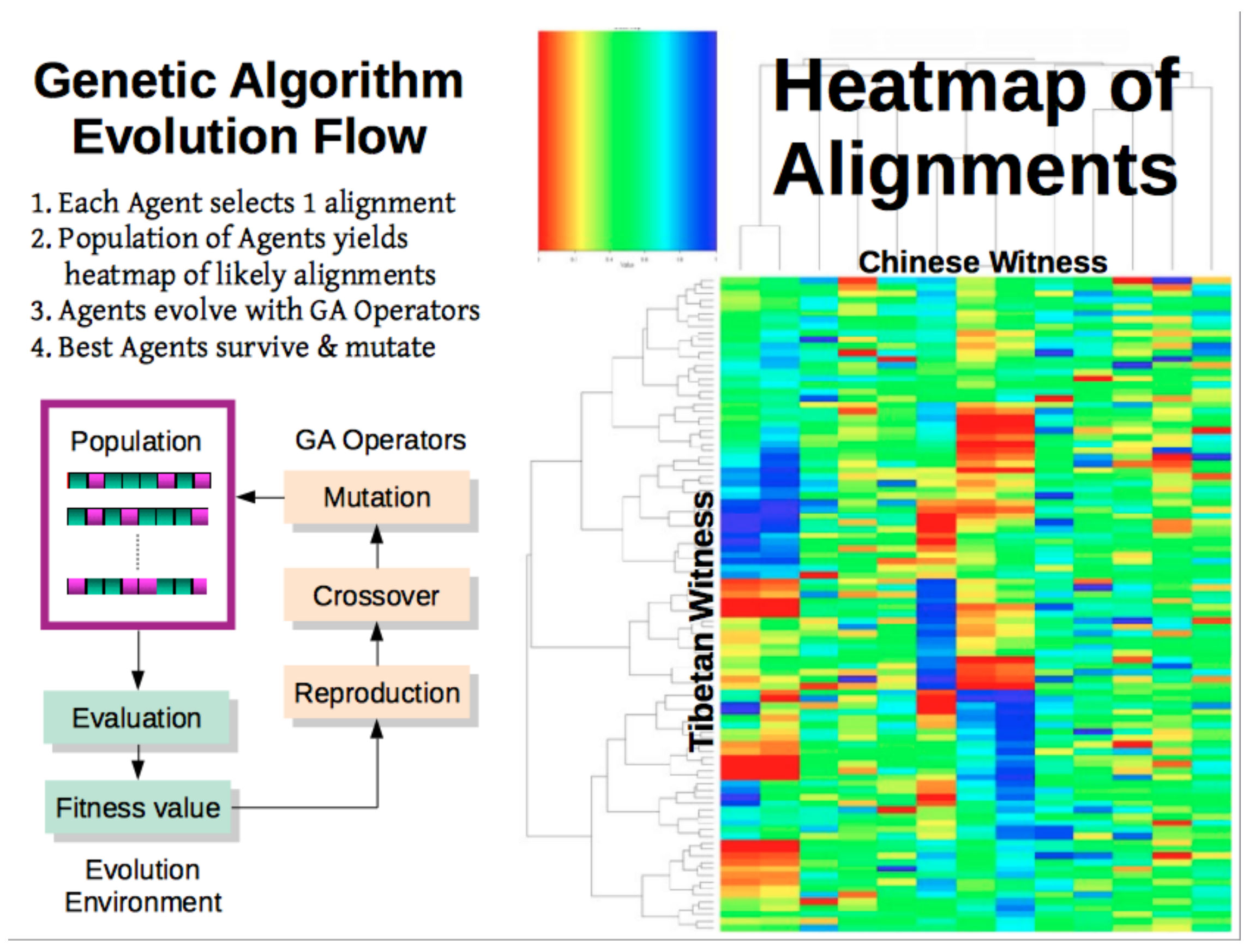

Genetic Algorithm Evolution Flow
1. Each Agent selects 1 alignment
2. Population of Agents yields heatmap of likely alignments
3. Agents evolve with GA Operators
4. Best Agents survive & mutate
Population
GA Operators
Mutation
Crossover
Reproduction
Evaluation
Fitness value
Evolution Environment
Heatmap of Alignments
Chinese Witness
Tibetan Witness

Fig. 2 & 3 - Genetic Algorithm Workflow & Alignment Heatmap

Results & Conclusions

⇒ We have created a unique 4-stage procedure to retrieve & measure philological alignments in a collection of Buddhist Witnesses in very low-resource languages.
⇒ We added linguist annotation and developed an innovative & intricate method of cross-linguistic N-gram Matching to overcome specific challenges for these languages.
⇒ This N-gram Matching can be used alongside existing methods from STS & MT by using intelligent agents as virtual machines to maximize efficiency and retrieve the most optimal results.