

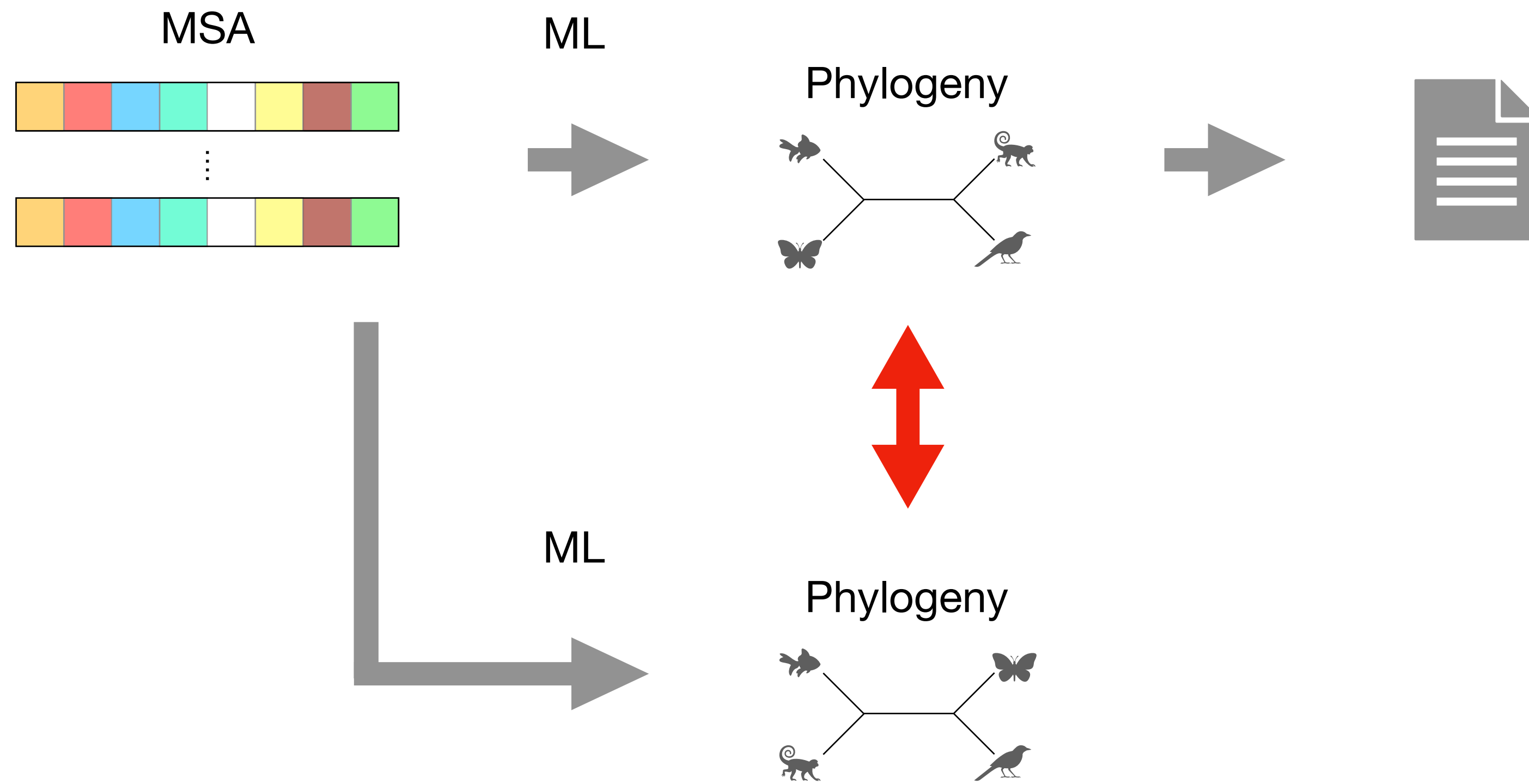
Educated Bootstrap Guesser

Predicting Phylogenetic Bootstrap Values

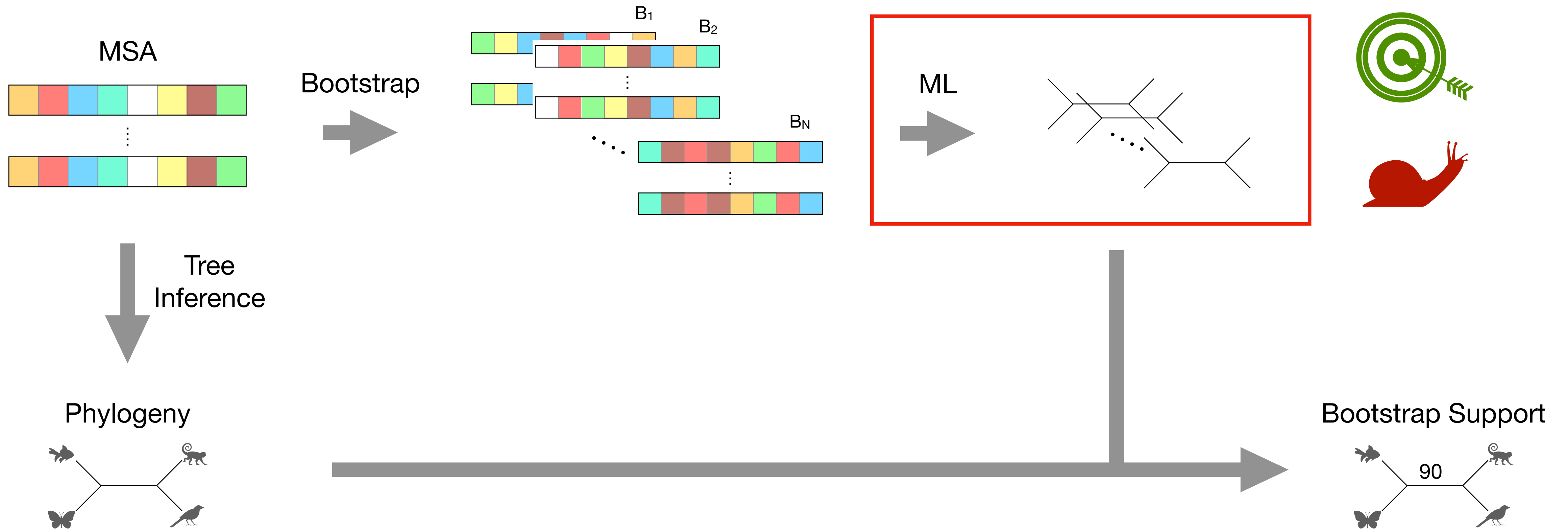
Julius Wiegert, Dimitri Höhler, **Julia Haag**, Alexandros Stamatakis



Motivation



Felsenstein Bootstrap (SBS)



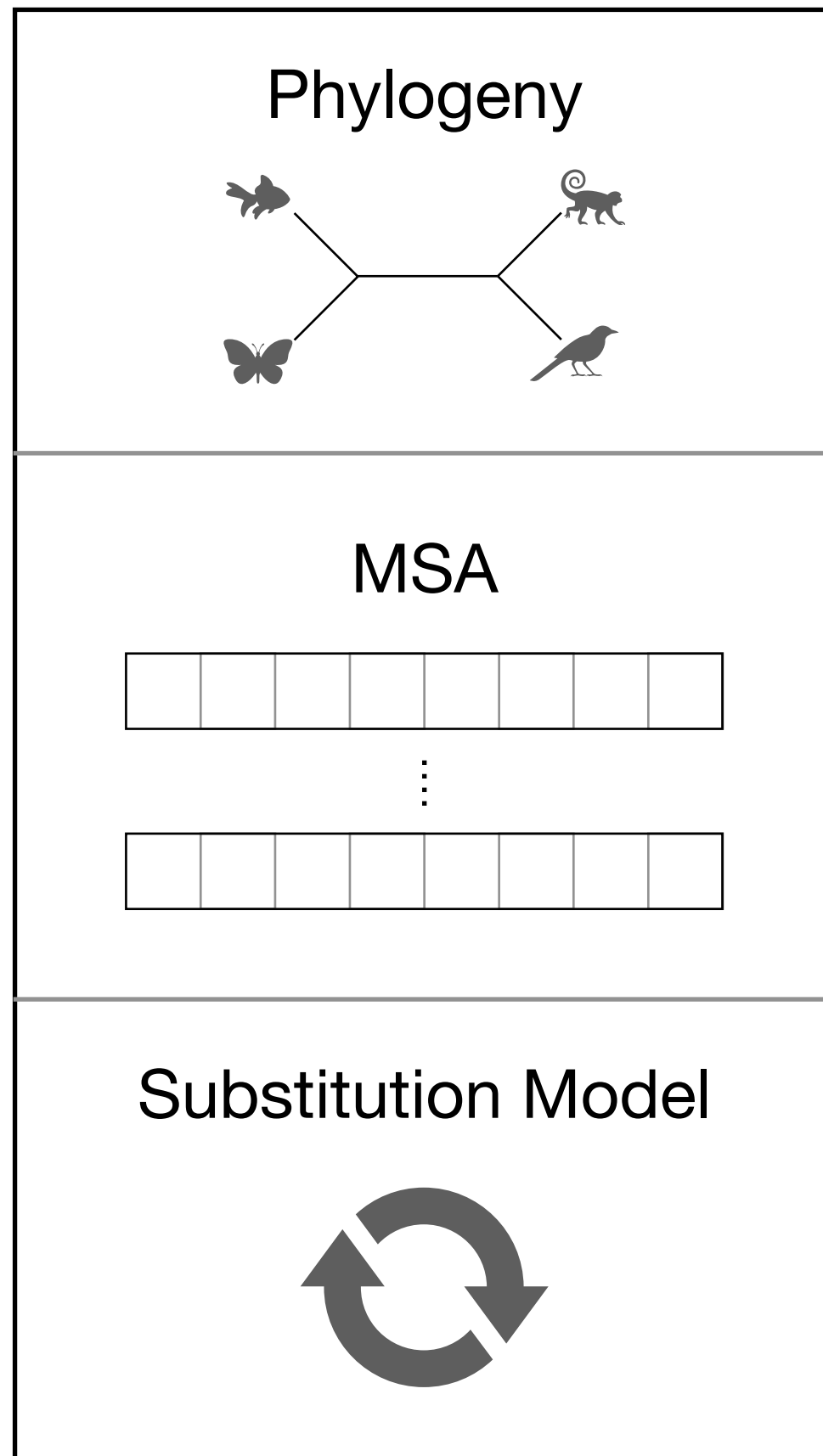
Bootstrap Approximation

- **Rapid Bootstrap (RB)** (Stamatakis *et al.*, 2008)
 - ✓ Accurate approximation + ~15x speedup to SBS
 - ✗ ML
- **UFBoot2 (UBS)** (Hoang *et al.*, 2018; Minh *et al.*, 2013)
 - ✓ ~8x speedup to RB
 - ✗ Different interpretation
- **SH-like aLRT** (Anisimova and Gascuel, 2006; Guindon *et al.*, 2010)
 - ✓ ~5x speedup to SBS
 - ✗ Different interpretation, overconfidence

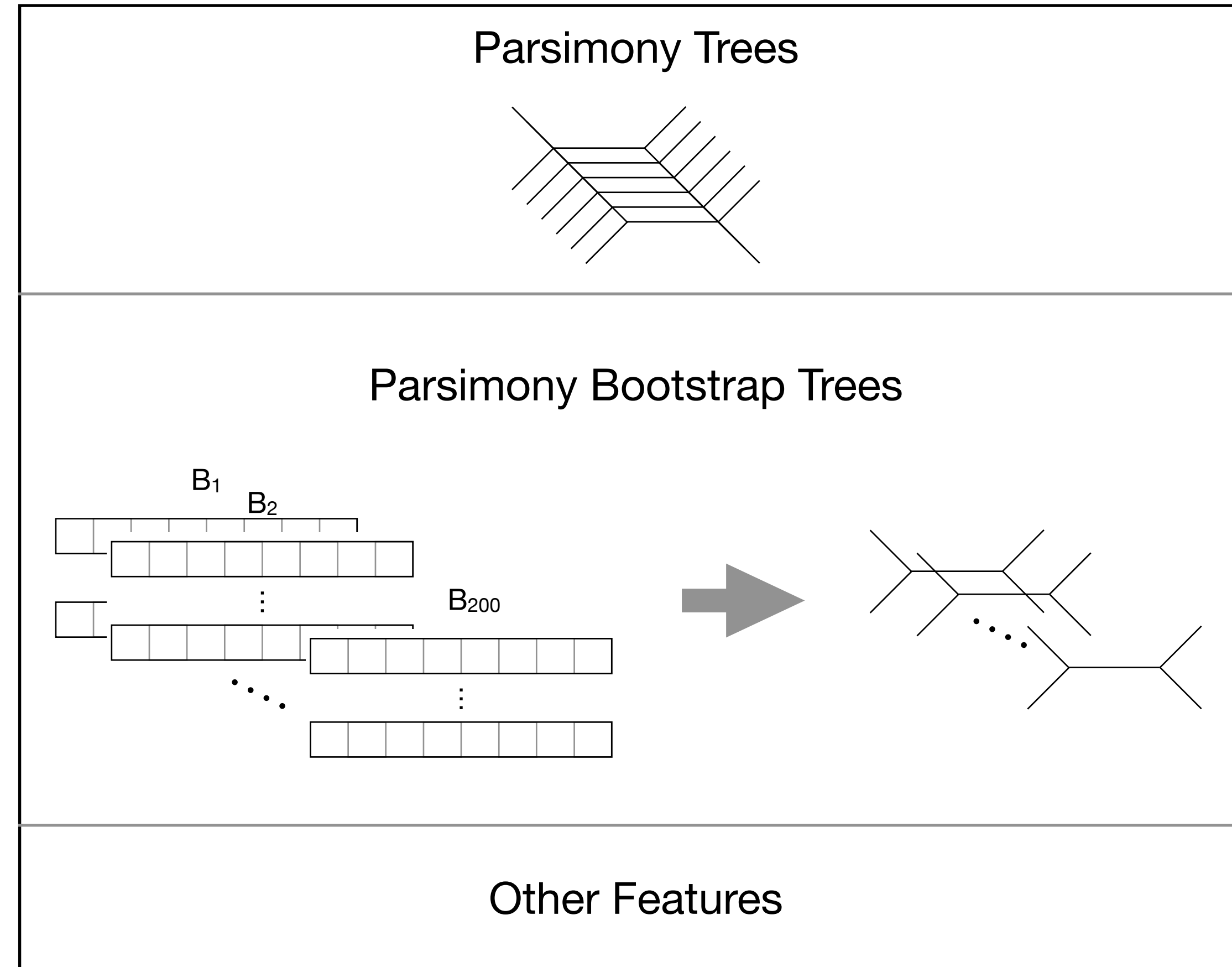
Educated **B**ootstrap **G**uesser

EBG

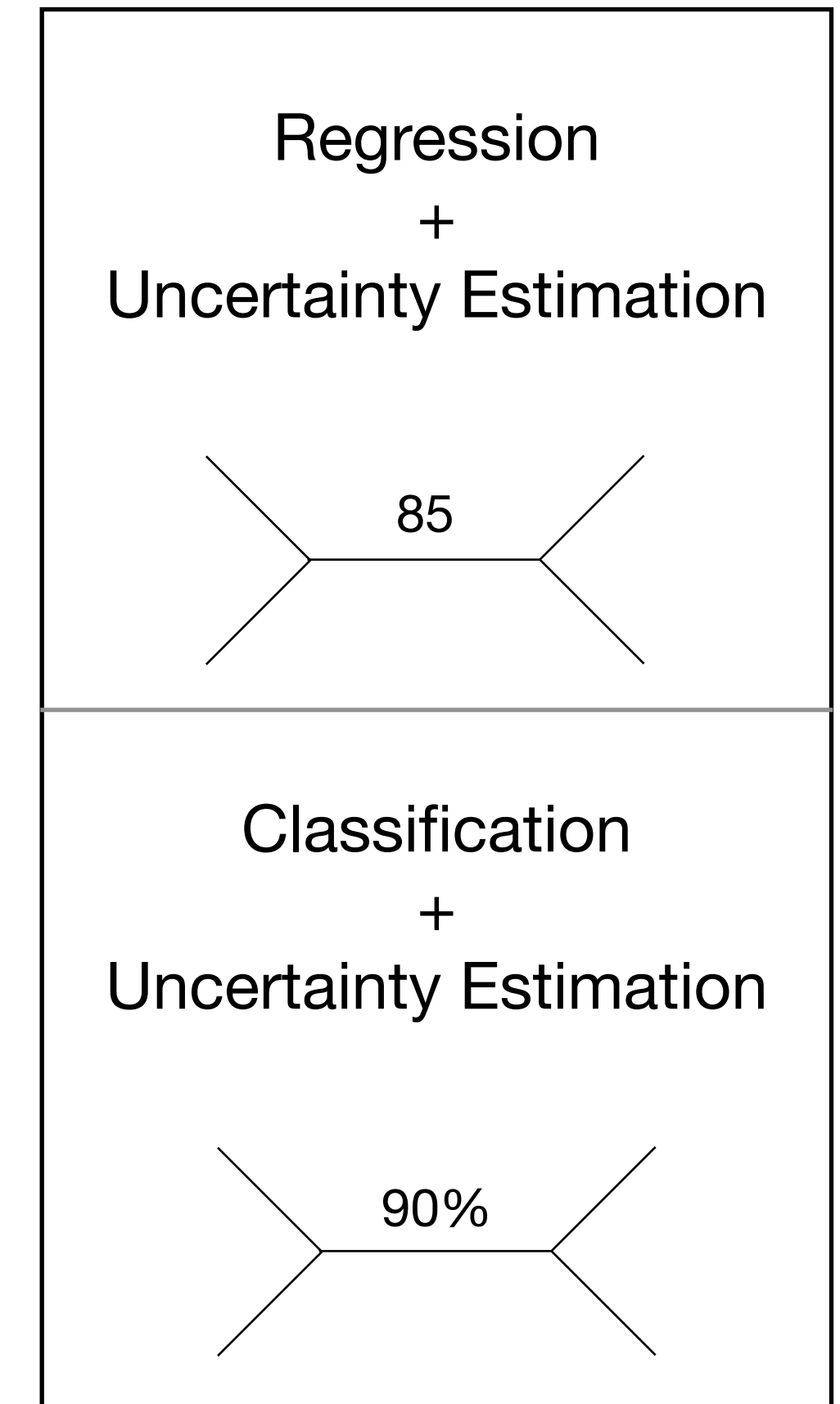
Input



Feature Computation



Prediction

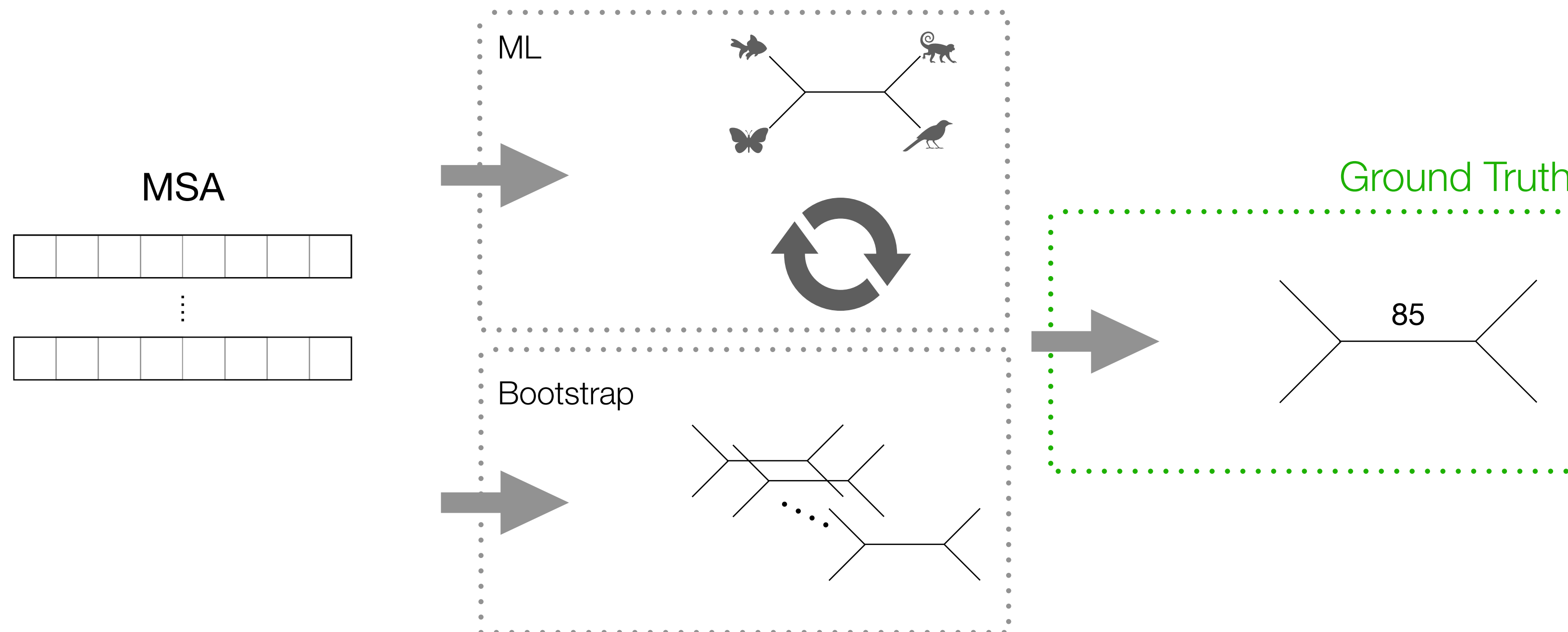


Training

- EBG = 7 machine learning models
 - 3 regression tasks
 - 4 classification tasks
- Predictors = LightGBM Gradient Boosted Tree

Training Data

- 1496 empirical DNA + AA MSAs (TreeBASE)
- Training data: ~80 000 inner branches + ground truth
- 10-fold CV



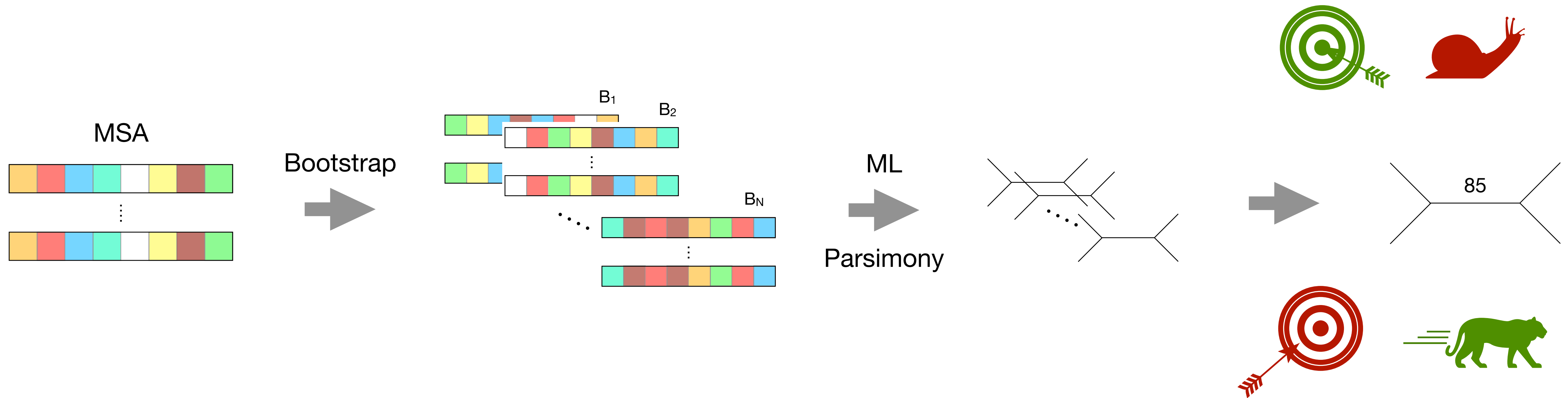
Prediction Results

- Regression: Quantile Regression
 - Point-estimate SBS values
 - 5% and 10% lower bound estimates (uncertainty estimation)
 - Upper bound?
- Classification: 1 - SBS as p-value (Felsenstein and Kishino, 1993)
 - Probability that branch exceeds threshold t
 - One prediction + uncertainty estimate for each t in [70, 75, 80, 85]

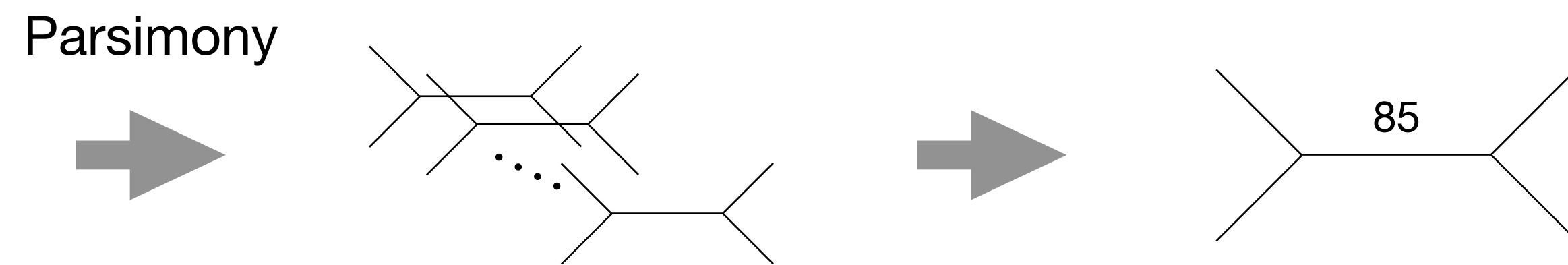
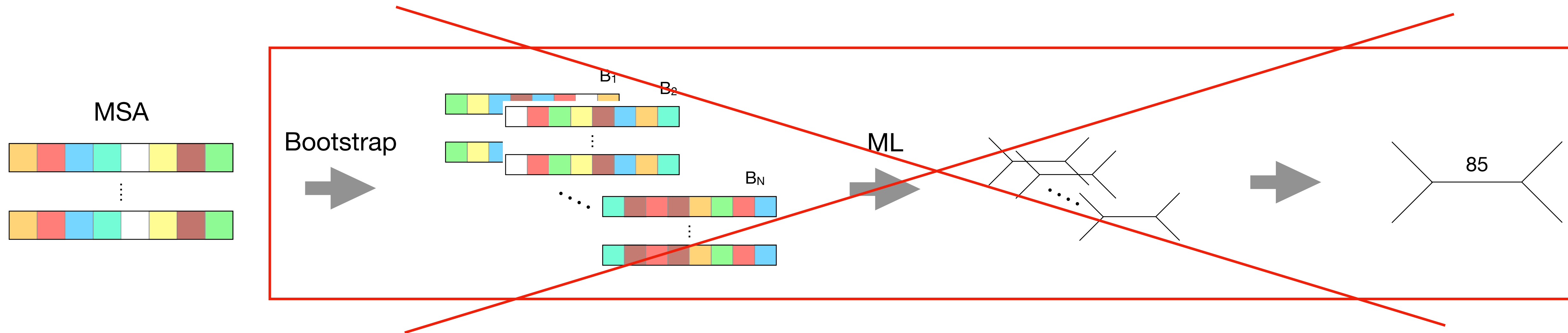
Features

- Experimented with more than 150 features
- Final predictor: 23 features (RFE)
- Features based on
 - Phylogeny
 - MSA
 - Parsimony Bootstrap Support
 - Parsimony Support

Parsimony Bootstrap Support



Parsimony Support



Evaluation

- Prediction:
 - Regression: Percentage values (0 – 100)
 - Classification: Threshold $t = 80$
- Baseline: 200 parsimony bootstrap trees

	Regression		Classification		
	MAE	MdAE	BAC	AUC	F1
Baseline	13.8	8.6	0.85	0.85	0.82
EBG	8.3	5.0	0.91	0.98	0.89

Feature Importance

Feature	Importance
Parsimony Bootstrap Support (PBS)	82.2%
Parsimony Support	3.1%
Normalized branch length	2.0%
# child inner branches	1.7%
Skewness PBS	1.5%

Parsimony = good & fast prediction of ML results

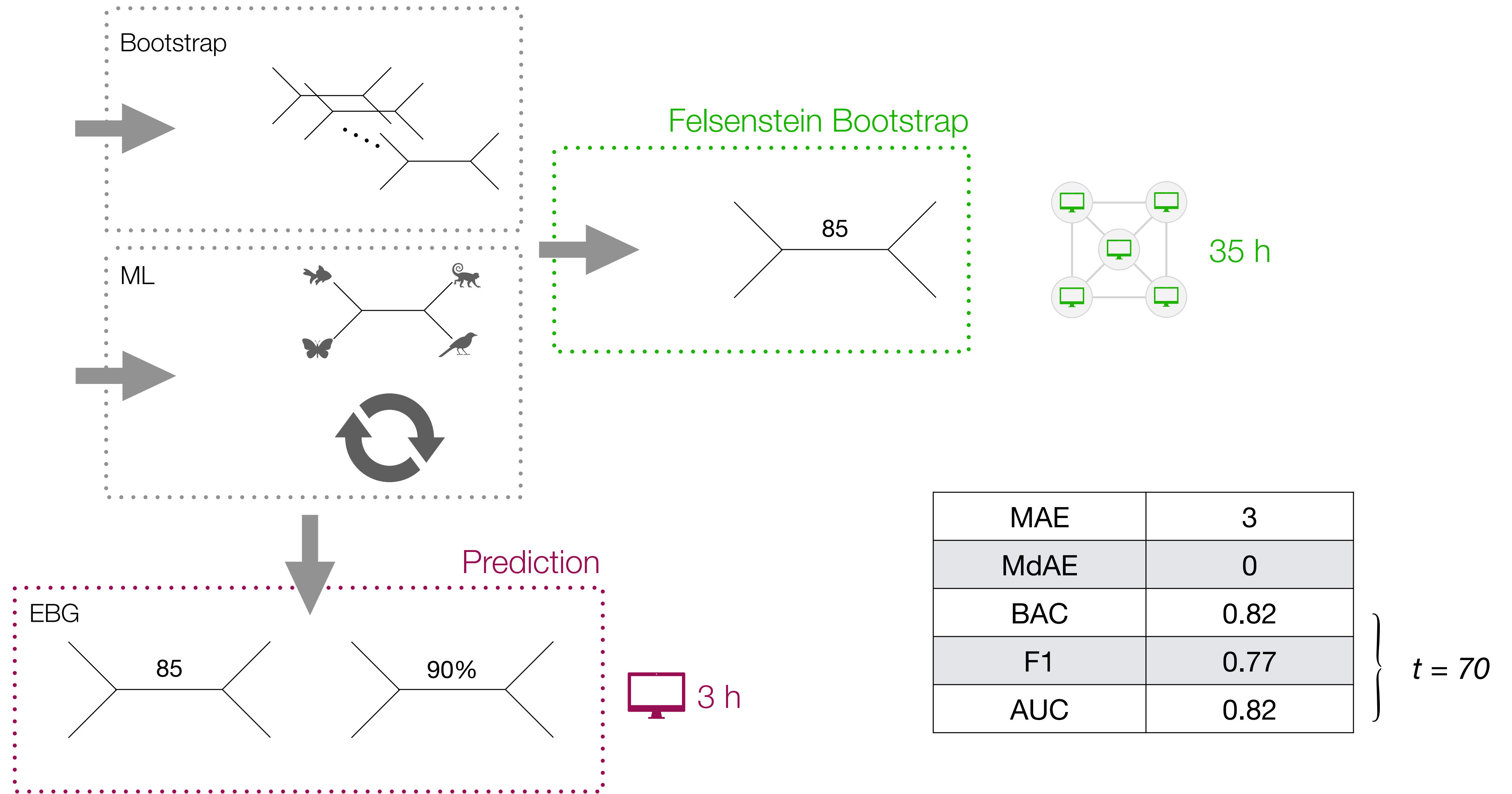
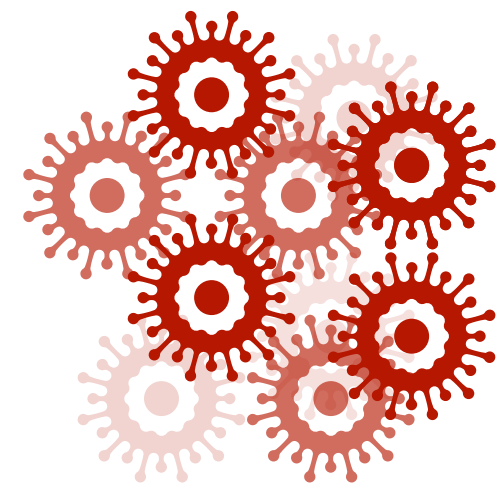
⇒ Renaissance of parsimony-based methods?

Tool Comparison

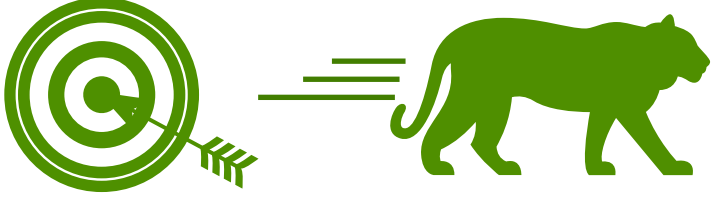
- Alternative tools:
 - RB (RAxML)
 - UFBoot2
 - SH-like aLRT
- Accuracy:
 - Simulated data: EBG performed best
 - Empirical data: RB > EBG* > EBG
- Runtime:
 - ~100x speedup to SBS (RAxML-NG) and ~10x speedup to UFBoot2

Example: Covid Dataset

(Morel *et al.*, 2020)



Summary

- EBG = SBS predictor
 - Regression, Classification, Uncertainty Quantification
- High predictive power, fast prediction 
- Available on conda
 - `conda install ebg -c conda-forge`
- GitHub: <https://github.com/wiegertj/EBG>
- biorXiv preprint: <https://doi.org/10.1101/2024.03.04.583288>