moose
an R package for mining training sets from dichotomous keys (among other things)

Jeremy Leipzig
Drexel MRC

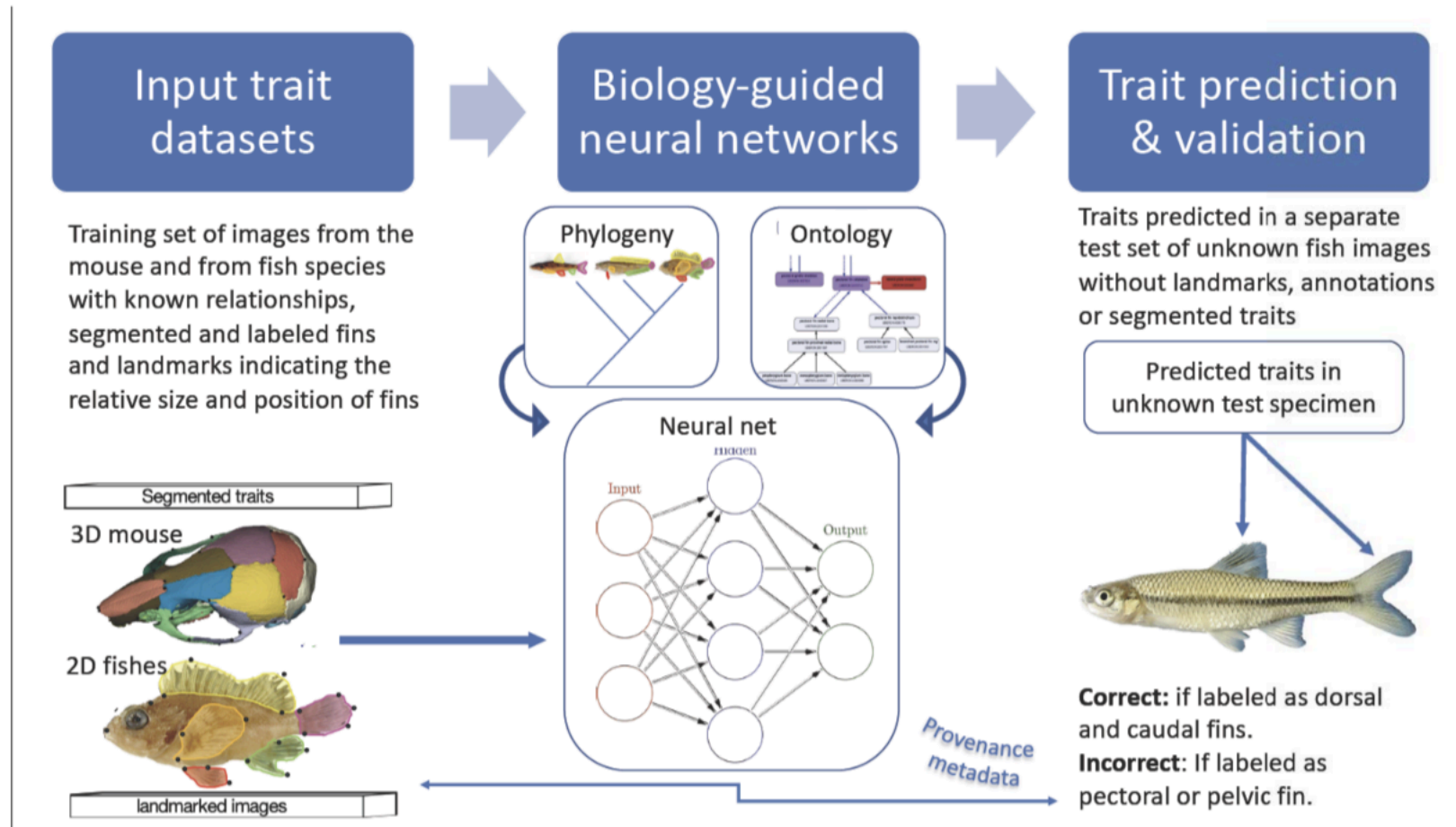# HDR/BGNN: Genotype -> Phenotype

Figure 2. Overall scheme of workplan. Training data, which will focus initially on mouse and then fish species, with a focus on fin traits, will be input to BGNN, which are informed by domain knowledge of evolutionary and anatomical relationships, which are available in structured and machine-readable forms (phylogeny, ontology). The output, traits segmented by the machine, will be validated and the inputs and algorithms modified in response. Provenance metadata is collected for specimens and processes.
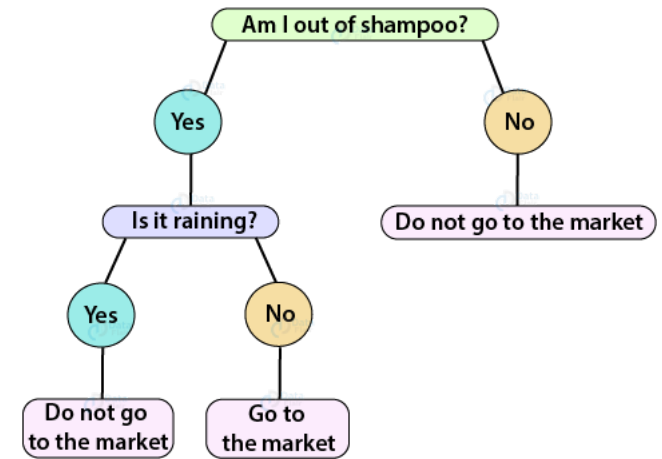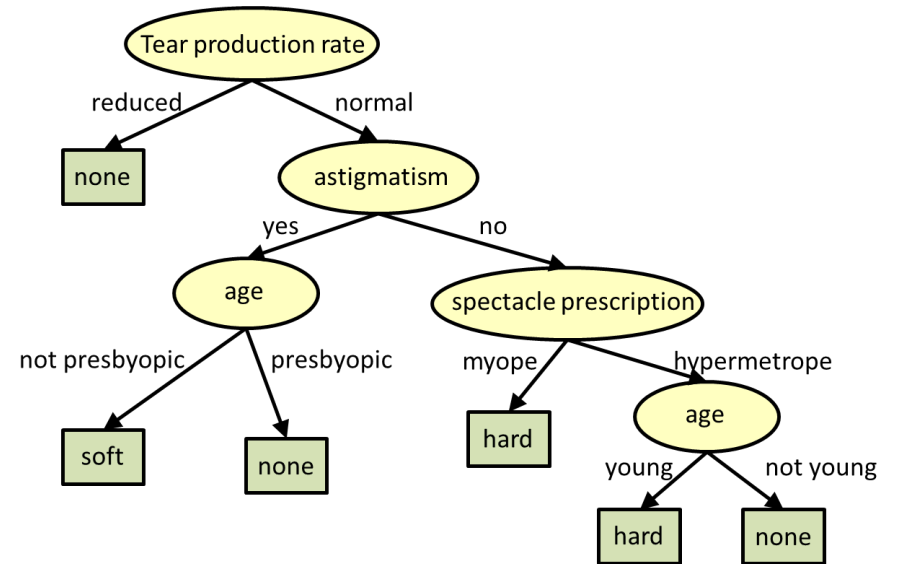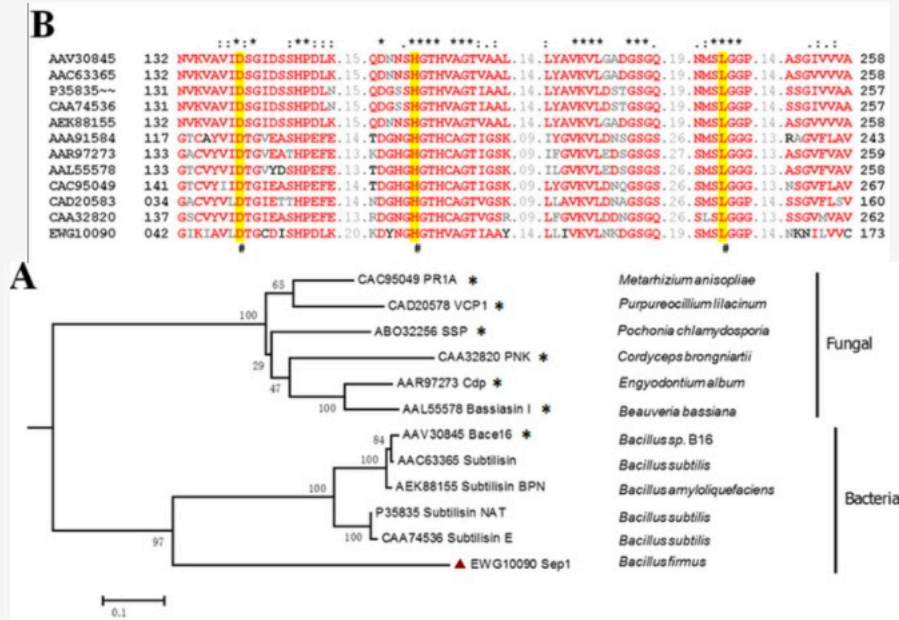
Dichotomous keys
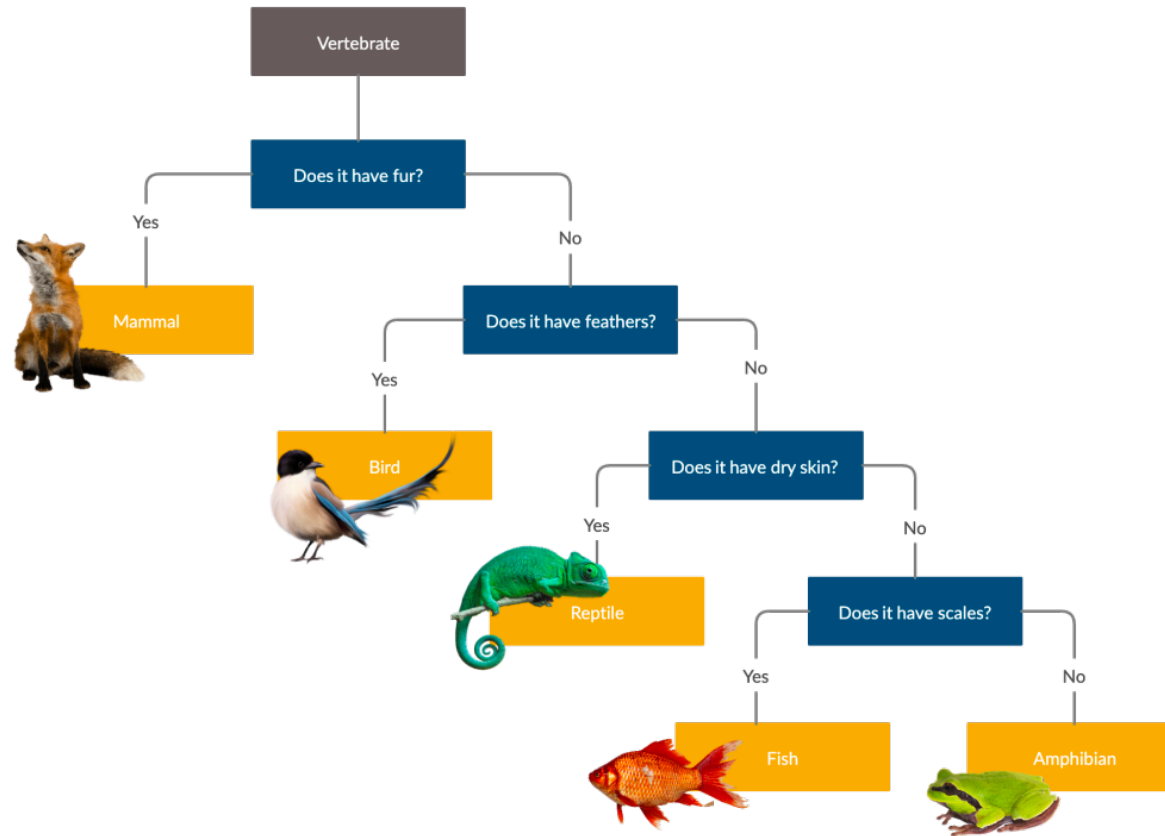
Phylogenetic trees vs Decision Trees
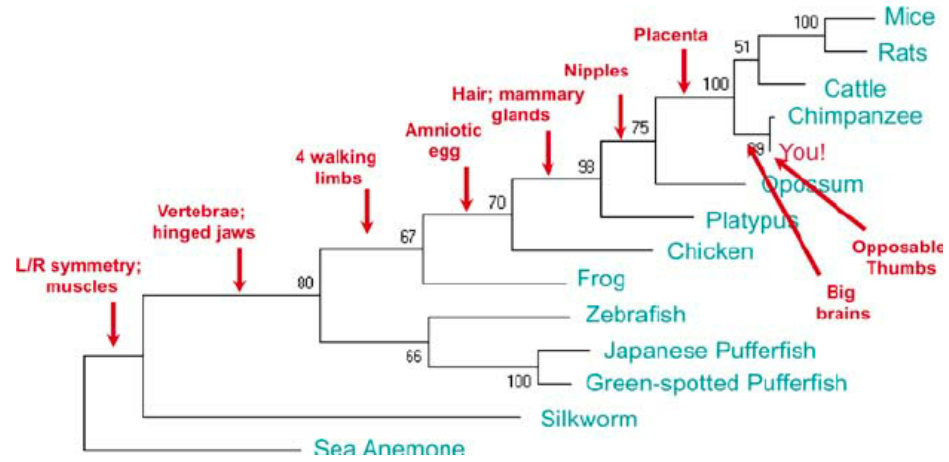
Decision Trees Example


From MSA to Phylogenetic Tree

Pros and Cons of Using Fishbase

- Pros:
  - Large selection of features (1677 keys in Fishbase)
  - Features designed to discriminate between closely related taxa
- Cons
  - Ambiguous, inconsistent, imprecise terms
  - Many keys do not extend to species
  - Teeth
  - Difficult to parse

Tree structures in R

- Ape phylo (S3) & phylo4 (S4)
  - Nodes and edges
  - Tip labels, node labels, edge labels
  - Support for phenotypic data
- data.tree (S6)
  - Just nodes
- Rpart & partykit
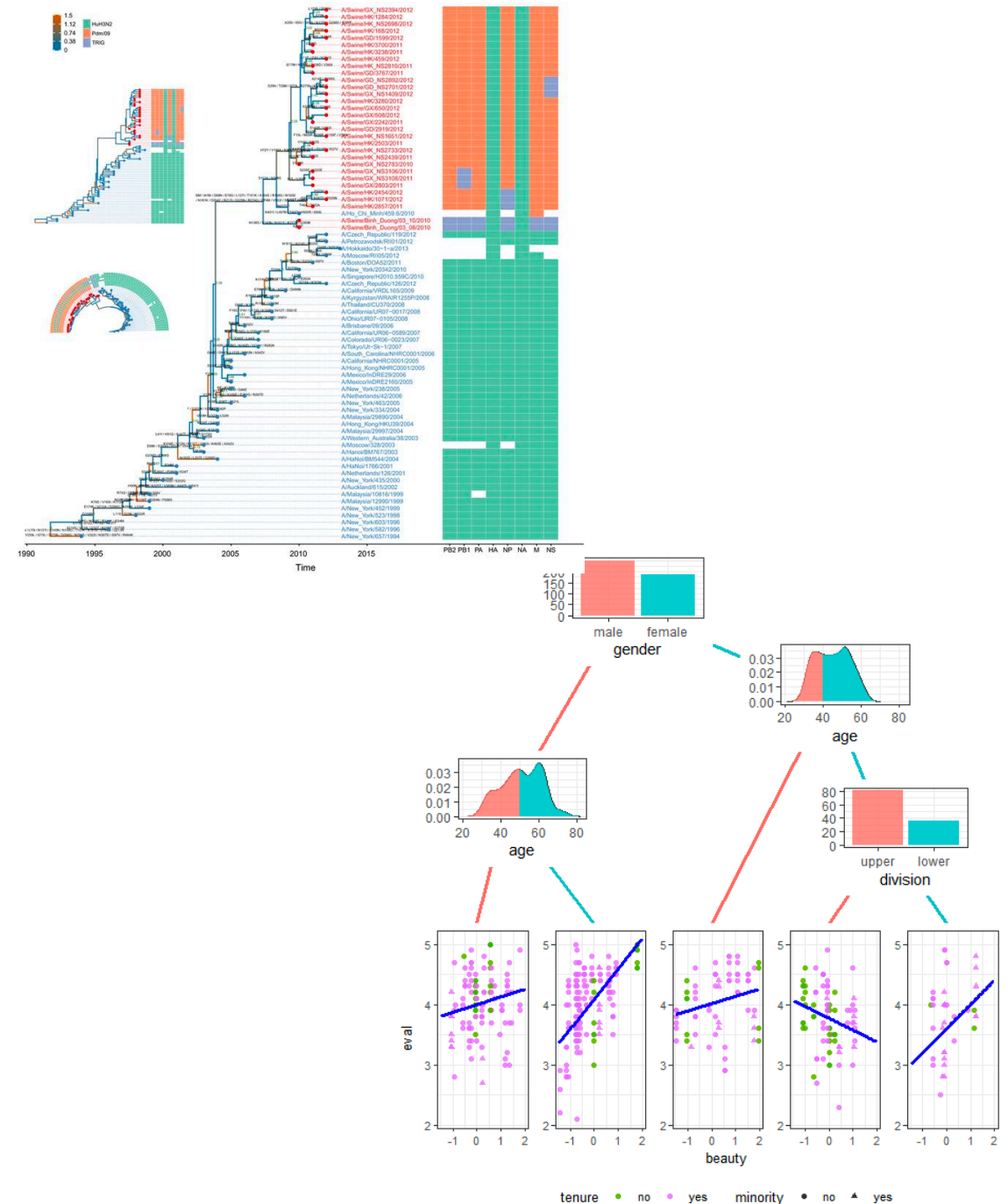
Goals

- Parse existing trees
- convert from and to popular formats including consensus matrices, ape phylo, data.tree, data frames, JSON, YAML
- merge and prune dichotomous keys
- generate interactive dichotomous keys ("wizards")
- produce training data for machine learning using splits
- metadata slots for provenance and supporting guide images
- perform basic lexical analysis, and specialized data types (single nucleotide variations) of criteria and taxa
- ontological annotation, together with labels at the decision nodes and edges can be mapped to ontologies using join tables
- build mixed effects conditional inference trees (mecits)

Where can I see more?

- https://github.com/leipzig/moose (send me a request for a preview)