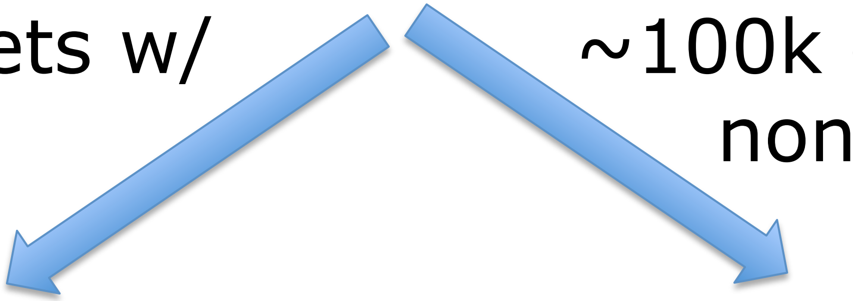
Protein Domains Search Pipeline



~3k datasets w/
assembled
contigs

~100k datasets w/
non-assembled
reads

Inputs:
1) the (assembled) contigs
2) protein domain models
   - ✖ ~300k CDD, no PFAM, POGs, etc. to avoid later work in filtering overlapping hits
   - ✔ ~2k "viral"-enriched set from Rodney

Inputs:
1) the (non-assembled) reads
2) protein domain models
   - but...how to connect?
     i. k-mers – build our own method?
     ii. min-hash, if it would work?