1. What is the current number of sequences in Genbank? What was it 5 and 10 years ago? Which approximate function does the growth of sequences in Genbank follow (especially until 2010)? Is that faster or slower than which other famous "law" related to an essential component of computer hardware?

2. By which keyword/identifier/... would you retrieve all primate (lat. primates) sequences, possibly using other NCBI resources?

3. Given the nucleotide sequences (fragments) here (https://rostlab.org/~tmhpred/ex2_seqs.fasta), give the common name of the encoded gene(s) (or respective protein), the organism it is from and accession number of their highest scoring entry in Genbank.

4. The databases available through the NCBI E-utilities change over time. Therefore, first of all, use EInfo to get all currently available databases. Next, use ESearch to search for "aquaporin" in the database of proteins. Limit the results to just 1 and next retrieve the resulting database entry in xml format with EFetch. Check the sequence IDs besides the primary accession number. For example, what is the (deprecated) GI number of this entry? What organism is the protein from and what is the taxonomic identifier (number) associated with that organism?

Now perform the same analysis computationally. Be mindful of the usage guidelines mentioned in the lecture. You can make use of any existing libraries, in particular ‘Entrez’ from biopython.

Good luck,
L. Richter