Bioinformatics 401
General Outline

Instructors:

Dr. David Wishart

Dr. Gane Wong

Potential Projects

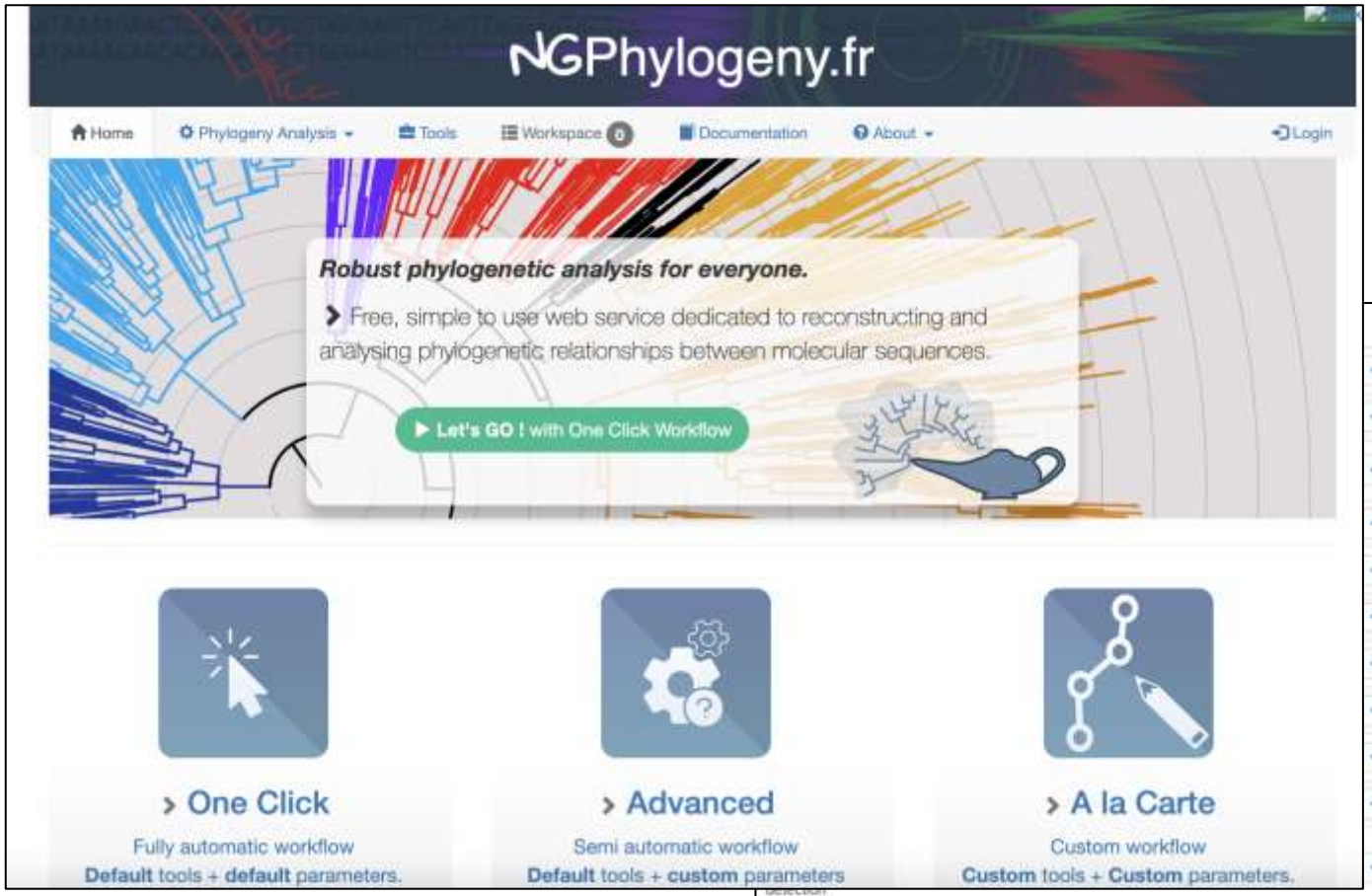
- This is a project-based course
- You have several choices of pre-approved projects (1 project with 1, 2 or 3 persons per project)
- A project of your choosing (needs to be well thought out, original, reasonable and vetted by the instructors)
- The pre-approved projects are listed on the following slides:

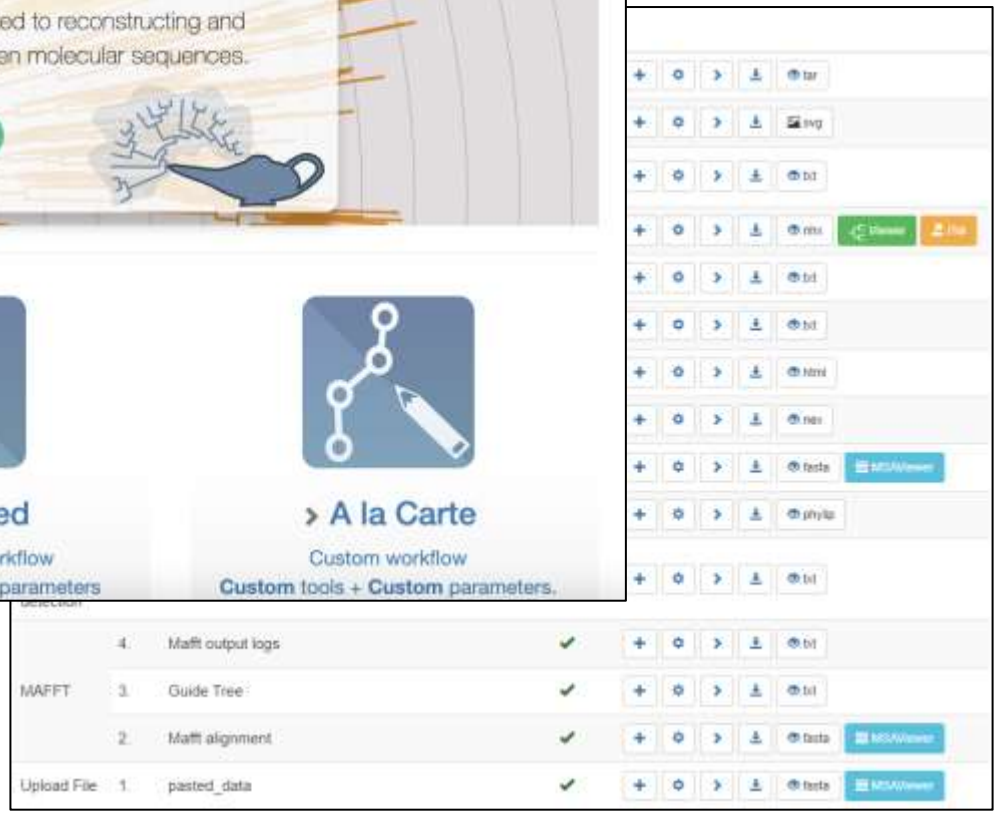
Potential Projects

Project #1

- **Develop a phylogenetic webserver to assemble and illustrate evolutionary trees for Bioin 301**
- **Would be similar to https://ngphylogeny.fr/ (and builds on IQ-Tree) but with more functionality; FRONT-end to create a dataset of orthologs for any given gene and taxonomic range; ENGINE to allow full-range of phylogenetic inference methods; BACK-end can link to iToL but might also compare trees from different genes**
- **A fair bit of programming skill will be required with knowledge of JavaScript, Ruby-on-Rails, and web design being beneficial – although you will have an opportunity to learn such skills on-the-job**

NGphylogeny Server

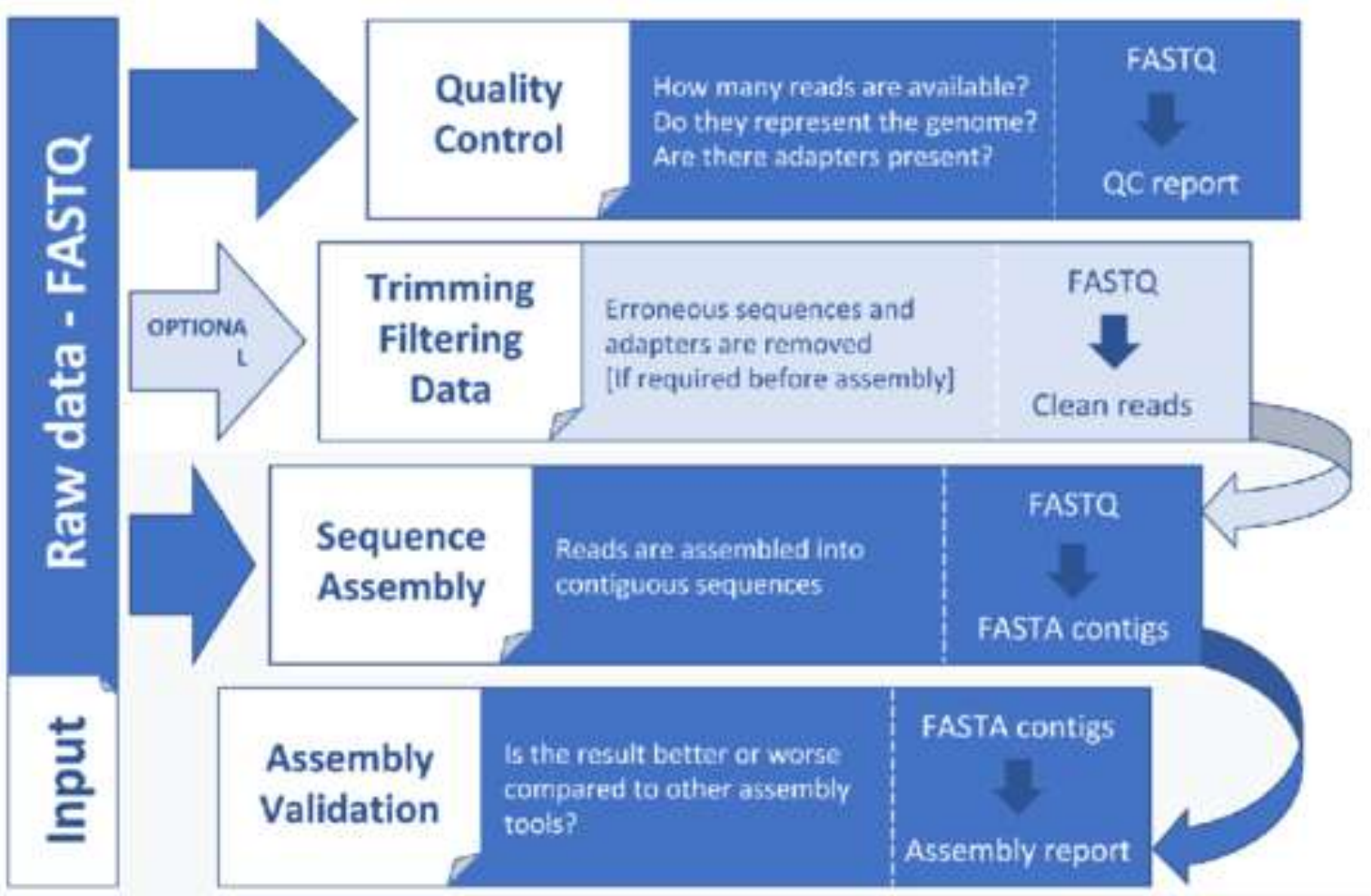




Project #2

- **Develop a genome assembly webserver for students in Bioin 301 to experience all the pitfalls of sequence assembly (assisted by Scott Han and Sukanta Saha)**
- **Server would have a FRONT-end to simulate fake reads from a reference sequence, an ENGINE in the middle to do assembly by various methods, and a BACK-end to analyze the result and compare it with the known reference; will also create DOT-plots to identify repeat motifs in reference**
- **A fair bit of programming skill will be involved with knowledge of JavaScript, Ruby-on-Rails, and web design being beneficial – although you will have an opportunity to learn such skills on-the-job**

Project #2 - Genome Assembly



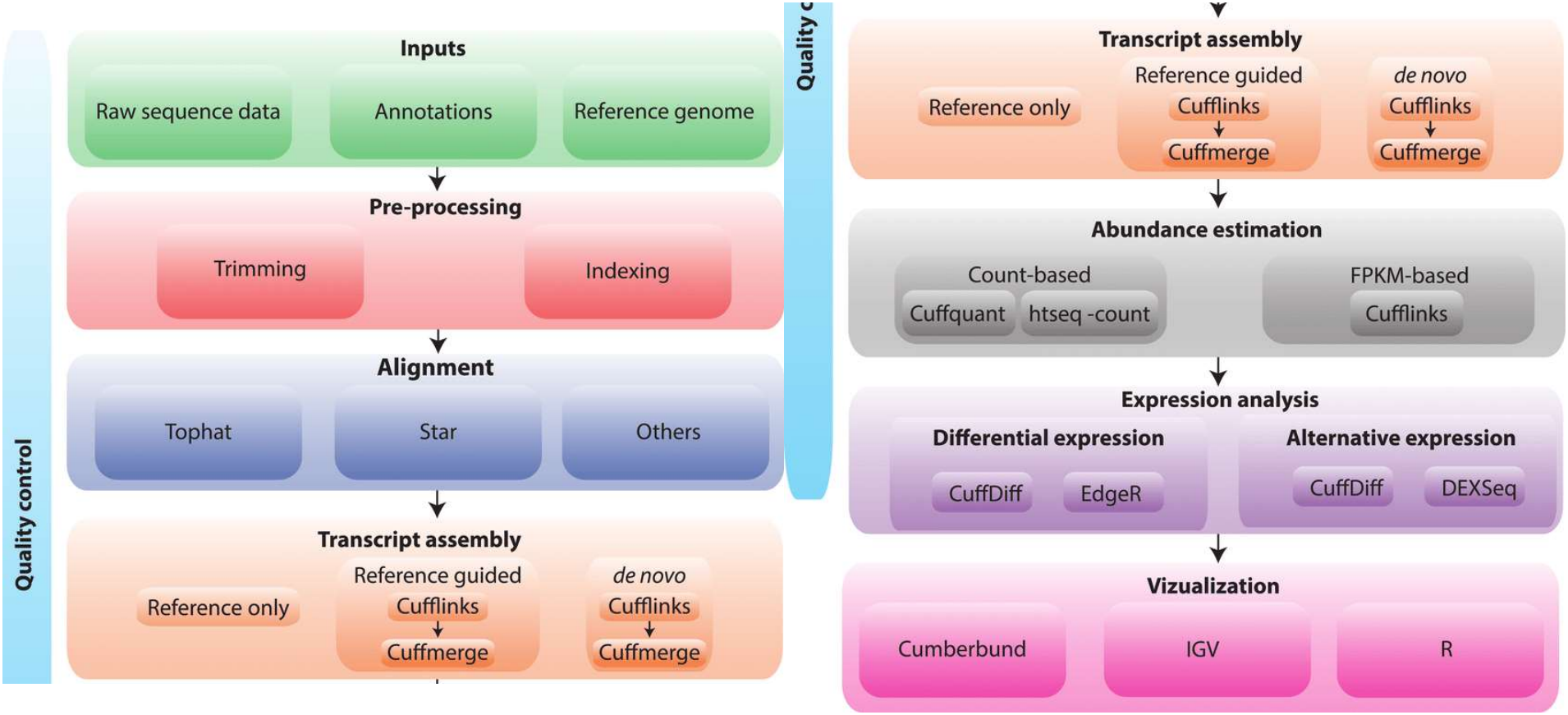
Project #3

- **Developing a web-based RNA-seq pipeline (webserver) called "RNAseek" (you will be assisted by Scott Han and Jenna Poelzer)**
- **RNASeek will accept FASTQ input files and runs FASTQC + Trimmomatic to assess and clean up data, then runs HISAT2 or STAR to compare to reference genome, then runs FeatureCounts to determine expression levels of genes and transcripts, then runs DESeq2 or edgeR to perform expression analysis**

Project #3 (cont'd)

- Then use GSEA and PathBank to ID enriched pathways or functional groups, then visualize output with HeatMapper2
- You can draw inspiration from www.expressanalyst.ca
- Most required programs will be freely available or freely downloadable
- Stitching things together will require a fair bit of programming skill with knowledge of Python, JavaScript, Ruby-on-Rails, and web design being beneficial – although you will have an opportunity to learn such skills on-the-job

Project #3 (cont'd)



Project #4

- Generating 30+ standard (machine readable) pathways describing protein signalling (mTOR, AMPK, Jak/Stat, Wnt, etc.) using a new pathway drawing tool called PathWhiz for a database called PathBank (Eponine Oler and Ray Kruger will assist you with training and project management)
- This is part of an NSERC-funded project supported by OMx Personal Health Analytics Inc., an Edmonton-based drug informatics company
- No programming is required, but good literature research skills are needed, a talent for drawing or depicting biological processes and some patience

Project #4 - Tools

http://pathbank.org/

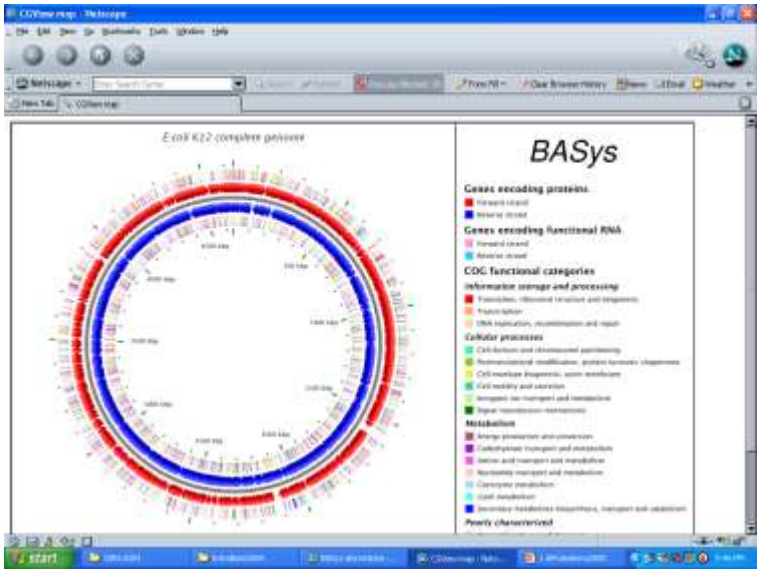
- Multi-organism pathway database specifically designed for multi-omics integration
- 200,000+ pathways from 10 model organisms including Human, Drosophila, S. cerevisae, C. elegans, E. coli, mouse, rat, cow, Arabidopsis
- Supports SBML, BioPAX and other machine-readable formats, with extensive hyperlinks to other DBs
- Shows phenotype, physiology, cell, protein and metabolite details in multiple formats

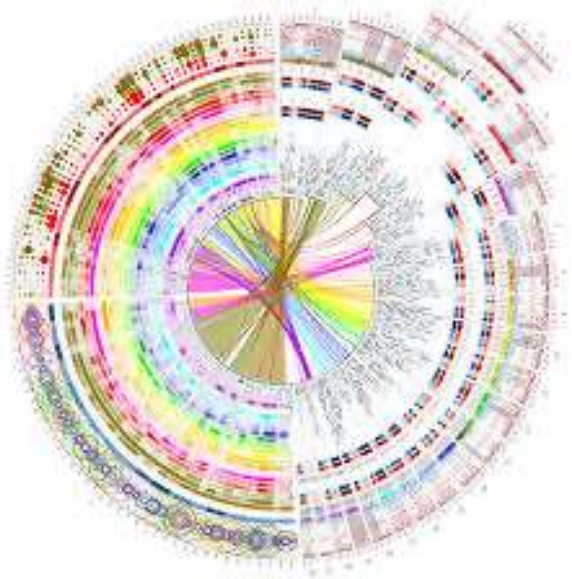
Project #5

- **EUKLID web server (Scott Han will assist you with programing and project management)**
- **EUKLID (EUKaryotic genome Labeling, Identification and Display) will be an interactive genome viewer for viewing annotated eukaryotic genome data**
- **Genome identifications will use comparisons to already annotated genomes (GenBank) while *de novo* genome annotations will be done via AUGUSTUS/Glimmer or GeneMark**
- **Genome/proteome/metabolome annotations will be done using annotation tools previously developed for BASys2**
- **EUKLID will make use of the circular genome visualization tools developed previously for BASys2, CGView and Circos**
- **Good chance that the database will be publishable in Nucleic Acids Research in 2025 or 2026**

Project #5

- Most programs will be freely available or downloadable
- Stitching things together will require a fair bit of programming skill with knowledge of Python, JavaScript, Ruby-on-Rails, and web design being beneficial – although you will have an opportunity to learn such skills on-the-job

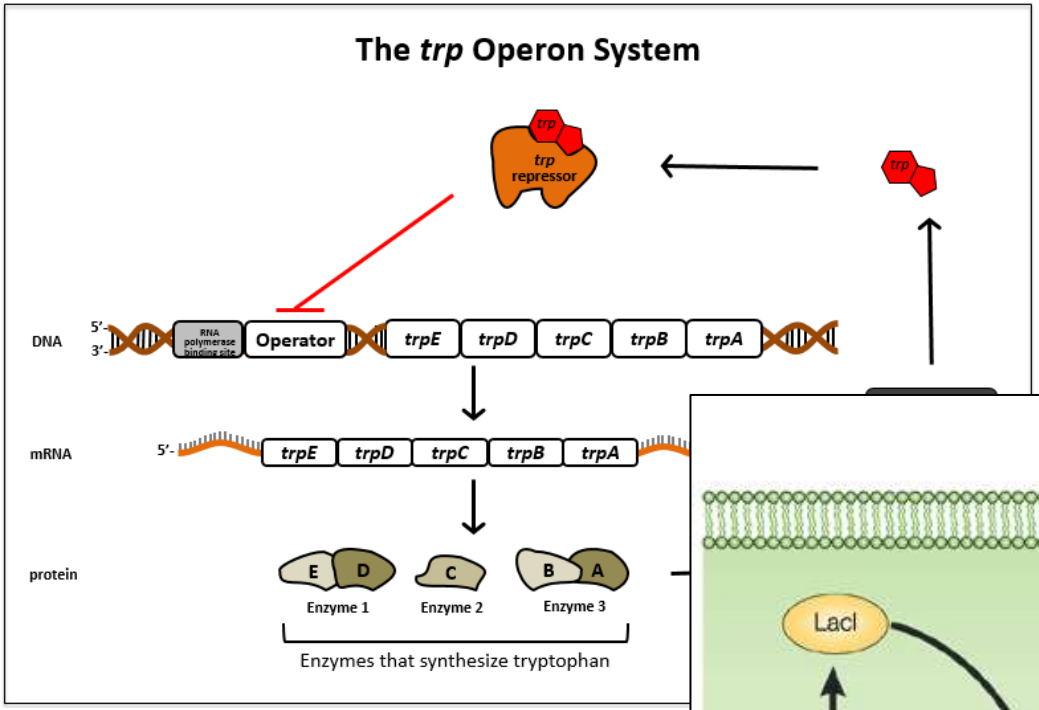


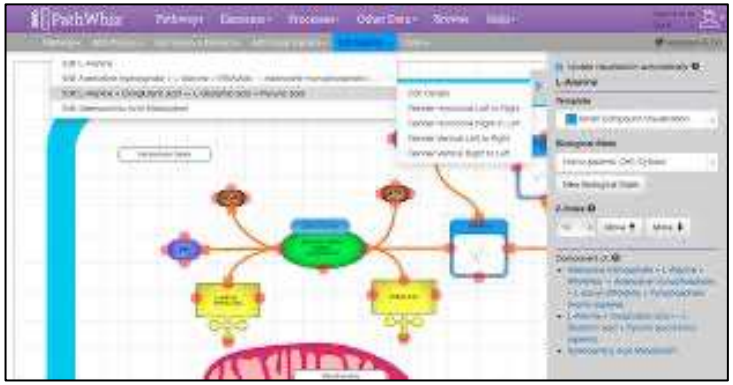


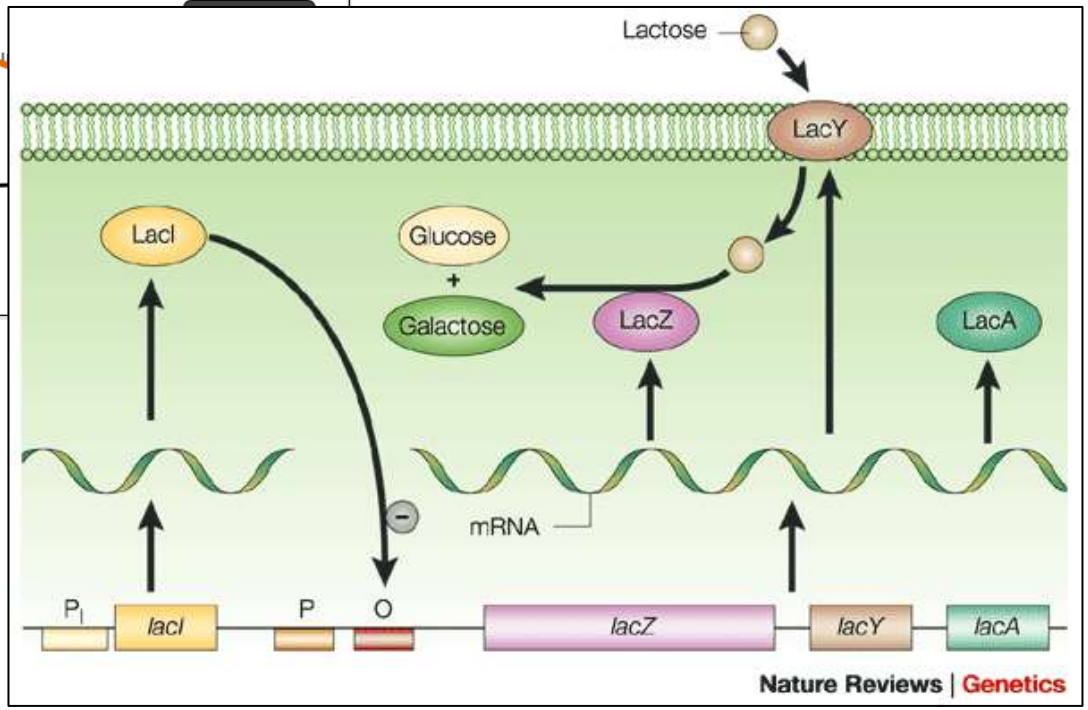
Project #6

- **OPEROD – The Operon Database**
- **This project involves generating 200+ standard (machine readable) pathways describing common bacterial operons using a new pathway drawing tool called PathWhiz (Eponine Oler and Ray Kruger will assist you with training and project management)**
- **The standard operons will be programmatically propagated and replicated to cover operons in 1000s of other bacteria**
- **No programming is required, but good literature research skills are needed, a talent for drawing or depicting biological processes and some patience**
- **Good chance that the database will be publishable in Nucleic Acids Research in 2025 or 2026**

Project #6


The trp Operon System
trp
trp repressor
trp
DNA
5'
3'
RNA polymerase binding site
Operator
trpE
trpD
trpC
trpB
trpA
mRNA
5'
trpE
trpD
trpC
trpB
trpA
protein
E
D
C
B
A
Enzyme 1
Enzyme 2
Enzyme 3
Enzymes that synthesize tryptophan

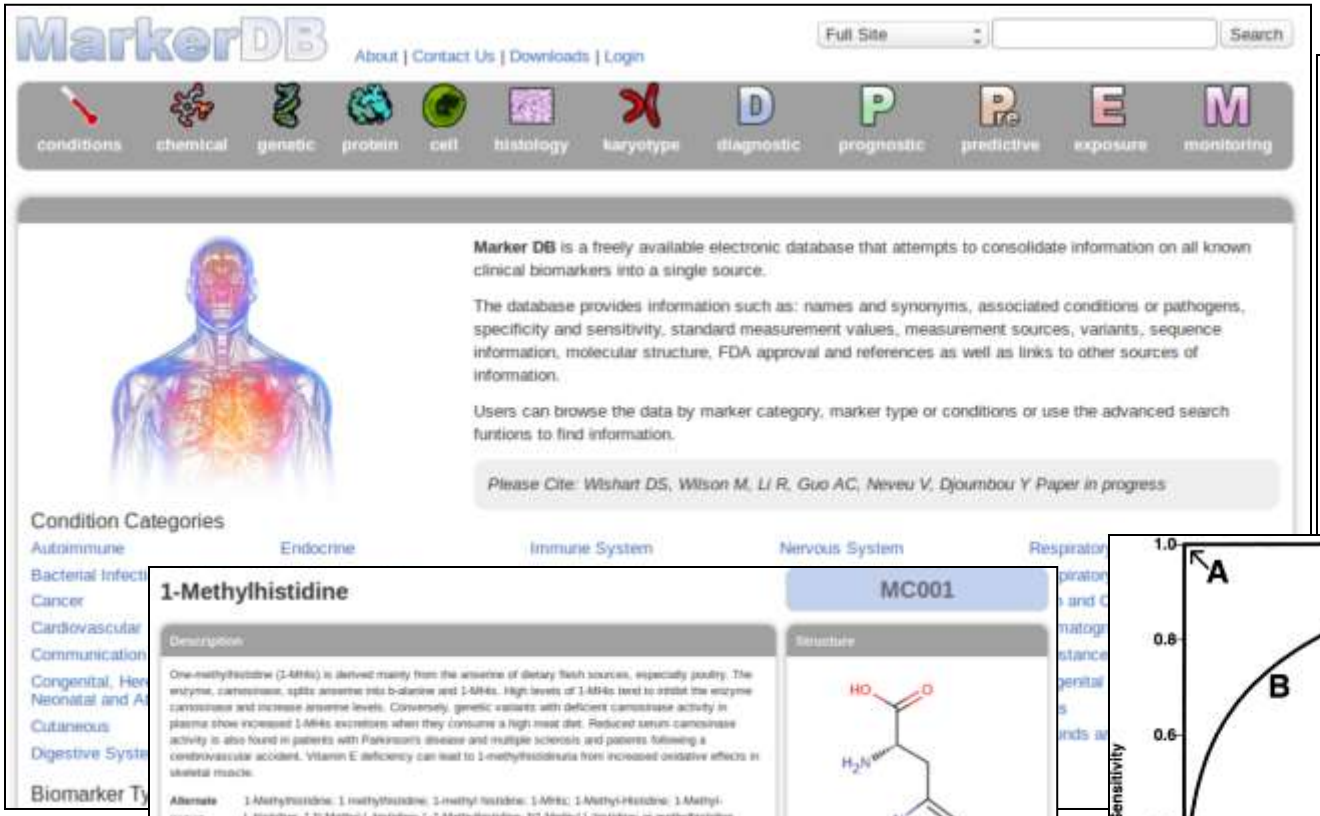


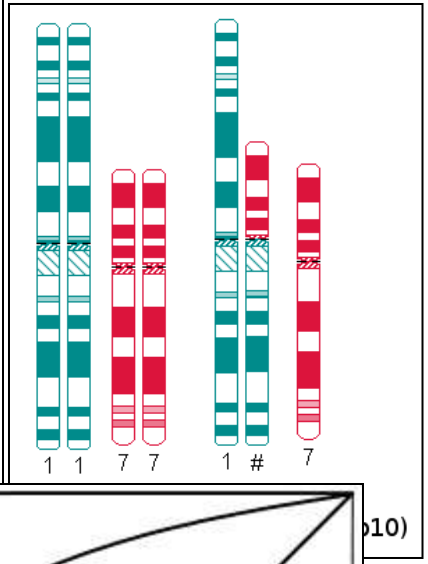

Lactose
LacY
LacI
Glucose
+
Galactose
LacZ
LacA
mRNA
P
I
lacI
P
O
lacZ
lacY
lacA
Nature Reviews | Genetics

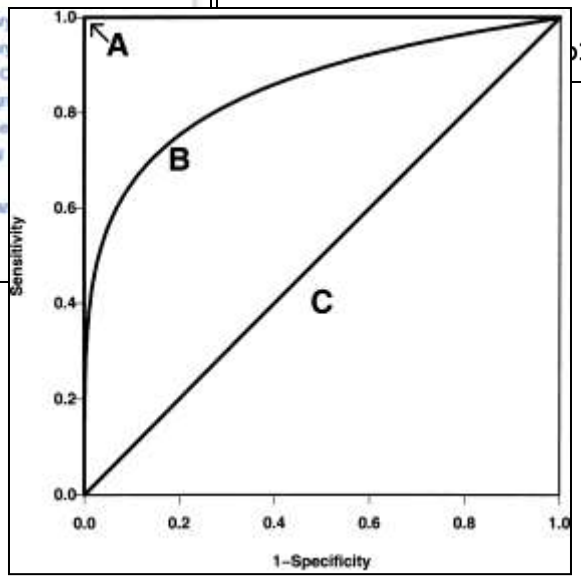
Project #7

- **Upgrading and expanding MarkerDB (version 3.0), a highly visual and interactive human disease biomarker database that includes genetic, karyotypic, protein and chemical biomarkers for 100s of different human diseases. (Eponine Oler and Mark Berjanskii will assist you with project management)**
- **Major focus for version 3.0 is on adding "omics" scale biomarkers which may include disease markers extracted from transcriptomics, metabolomics or proteomics experiments**
- **Some programming will be required, but good literature research skills are also needed. Major focus on data collection, validation and entry.**

Project #7 - MarkerDB







https://markerdb.ca/

Project #8

- Famous Scientist Avatar Project (Scott Mackay)
- Create a digital twin or digital avatar of a famous (dead) scientist using RAG (retrieval augmented generation) modified LLMs of the scientist's writings and works, use text-to-speech tools to mimic their voice and LLM-based video generation to create a mimic of their appearance (as a talking head)
- The avatar should be able to look, sound and respond scientifically as the real scientist would
- Consider: Alberta Einstein, Fred Sanger, Francis Crick, Dorothy Hodgkin
- See: https://thedali.org/exhibit/dali-lives/
- Use AI tools like Metahuman Creator, Blender, Sora ElevenLabs, Descript Overdub, ChatGPT, Inworld AI

Project #8 - Tools

Project #9

- Developing a chatbot version of the HMDB (Mark Berjanskii and Robyn Woudstra will assist you)
- HMDB is a “classic” web-based database on human metabolites, it is viewed >10 million times/year
- Idea is to extract data and knowledge triples using LLMs and/or to use RAG from the text data in HMDB and create a chatbot that can accurately answer and retrieve data from the HMDB
- You will work with Llama 3.3 (70B) to perform knowledge extraction/interpretation/output and use Neo4J, Neo4J Bloom and RDF4J for KG analysis
- Some Python programming is required, interest or knowledge of AI, ML, KG and text analysis is helpful. No need to develop a web tool or GUI

Project #9 - Tools

Project #10

- **Prepare 5 different, well planned, 3-5 minute instructional videos (using AI tools such as Sora, Synthesia, DeepBrain AI, Vyond, Camtasia, Powtoon) describing how to use the following tools: MetaboAnalyst, HMDB, DrugBank, PathBank, MarkerDB, BASys, MetaGenassist, CFM-ID, Heatmapper, Proteus2, etc.**
- **Or Prepare 3 different well planned 5-7 minute instructional videos on selected Bioin 301 topics (sequence alignment, nextgen sequencing, RNAseq, phylogenetic trees, AlphaFold2, MS-based proteomics, etc.) using the above-mentioned tools**
- **Project coordinated with Eponine Oler and Mark Berjanskii**

Project #10

- First step: Script & Storyboard
- Second step: Select avatar
- Third step: customize visuals by adding backgrounds, text and graphics
- Fourth step: select voice, adjust speech pacing, create multi-slide video
- Fifth step: use Camtasia to improve quality of video (screen recordings, music, other effects)

Project #10 - Tools

Project #11

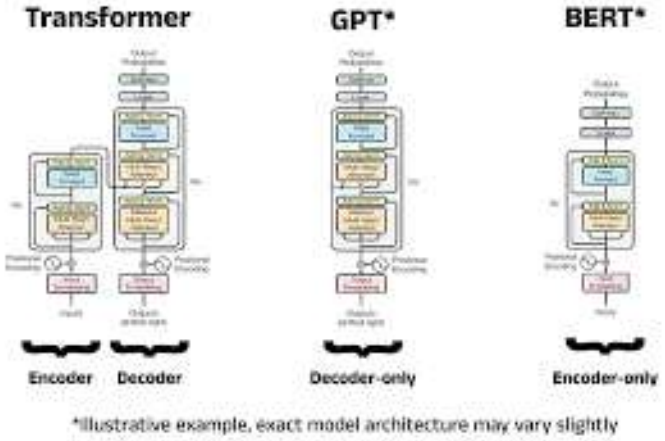
- Converting the Norman Suspect List Exchange (NSLE) to a web-enabled database (Eponine Oler, Ray Kruger, Robyn Woudstra can help coordinate)
- NLSE contains ~120,000 contaminant chemicals with minimal information and no web interface. These contaminant chemicals likely account for 90% of all human deaths (acute and long-term exposure) so they should be of interest to you and your family
- Idea is to use tools developed previously for building HMDB and FooDB along with LLM-based annotation tools to make a more complete database that has all the features, annotations, biological data and health effects for compounds found in HMDB, T3DB & ContaminantDB
- If finished, it could be published in NAR in 2025

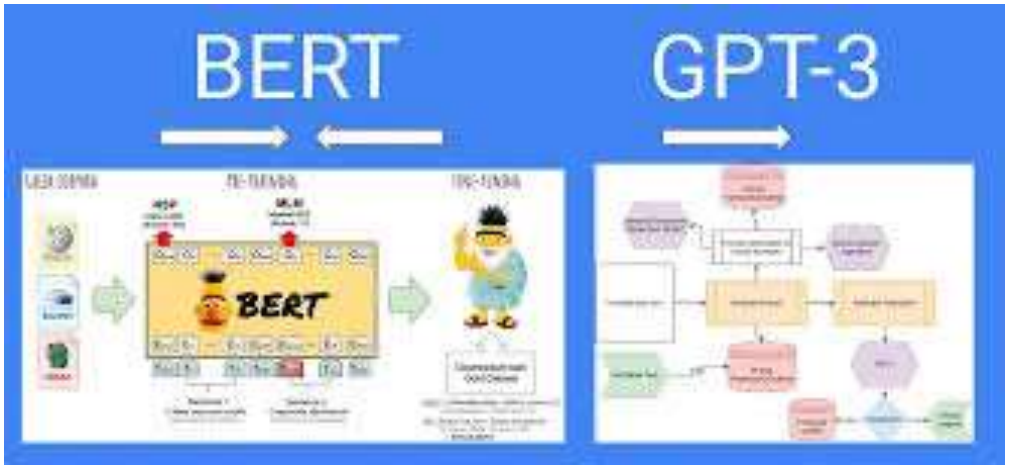
Project #11 - Tools

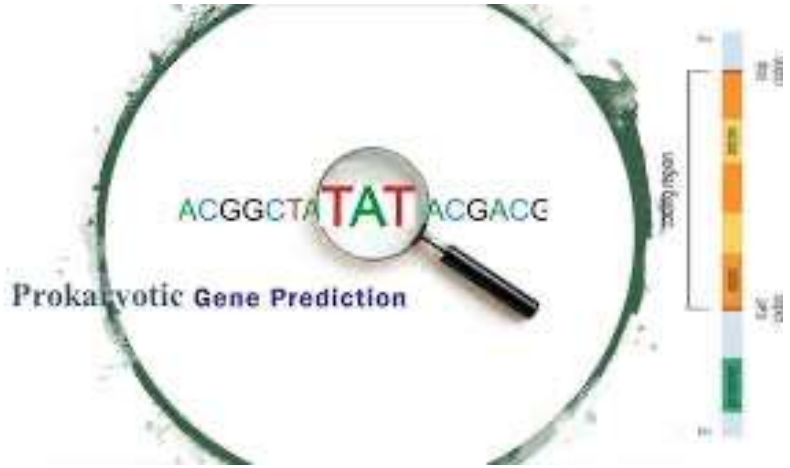
Project #12

- Develop an LLM-based tool for prokaryotic gene finding (Mark Berjanskii and Tanvir Sajed can help coordinate)
- LLMs can be used to find DNA signals in genes just like HMMs and PSSMs. Tools such as DNABERT and DNABERT2 are LLMs that have been trained to find DNA features and appear to do so quite well
- The idea is to use existing LLMs (like Llama and/or BERT or DNABERT) or modify existing LLMs to find prokaryotic genes and demonstrate/learn how LLMs can be used to perform high quality bioinformatic analyses
- Large training datasets available (BASys2, billion genes)
- Some Python programming is required, interest or knowledge of AI, ML, and transformer concepts is helpful. No need to develop a web tool or GUI

Project #12 - Tools



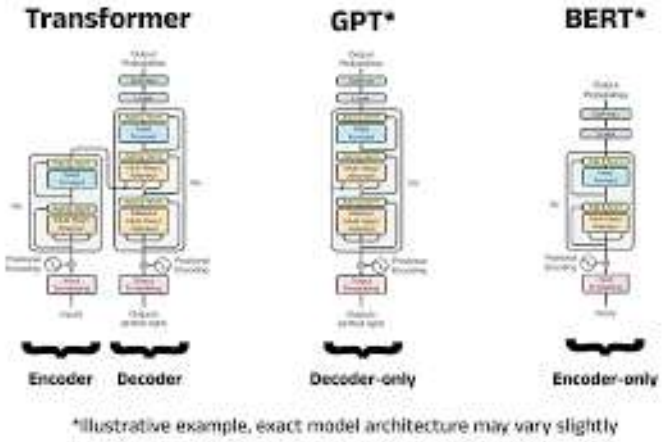


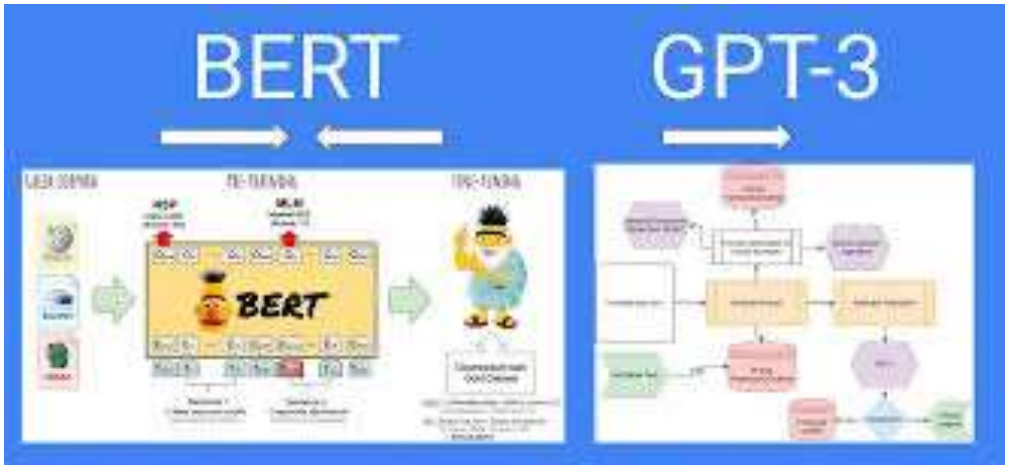


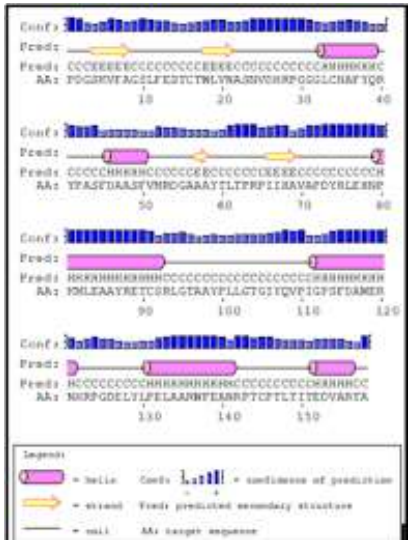
Project #13

- Develop an LLM-based tool for secondary structure prediction (Mark Berjanskii, Tanvir Sajed)
- LLMs can be used to “translate” primary sequence data to secondary structure just like HMMs and ANNs
- Large training datasets are available (AlphaFold/VADAR)
- The idea is to use existing LLMs (like Llama and/or BERT) or modify existing LLMs to perform secondary structure prediction (or primary to secondary text “translation”_ and demonstrate/learn how LLMs can be used to perform high quality bioinformatic analyses
- Some Python programming is required, interest or knowledge of AI, ML, and transformer concepts is helpful. No need to develop a web tool or GUI

Project #13 - Tools



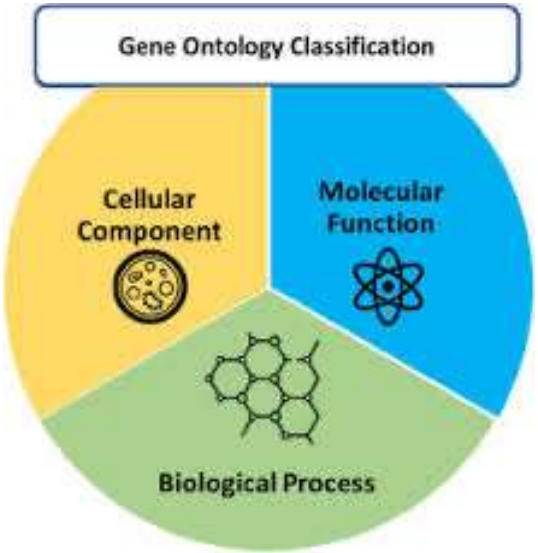




Project #14

- GOLLM - Develop an LLM-based tool for Gene Ontology identification (Mark Berjanskii, Robyn Woudstra)
- LLMs can be trained to recognize ontological descriptors from the literature
- Using a version of PubTator developed in the Wishart lab (which automatically identifies and labels proteins or genes in text documents), develop an LLM that identifies proteins in the literature and extracts Gene Ontology features about these proteins from the literature data found in PubMed
- The idea is to show that LLMs could replace human annotators in extracting GO data
- If successful, this could be published in a Bioinformatics Journal

Project #14 - Tools



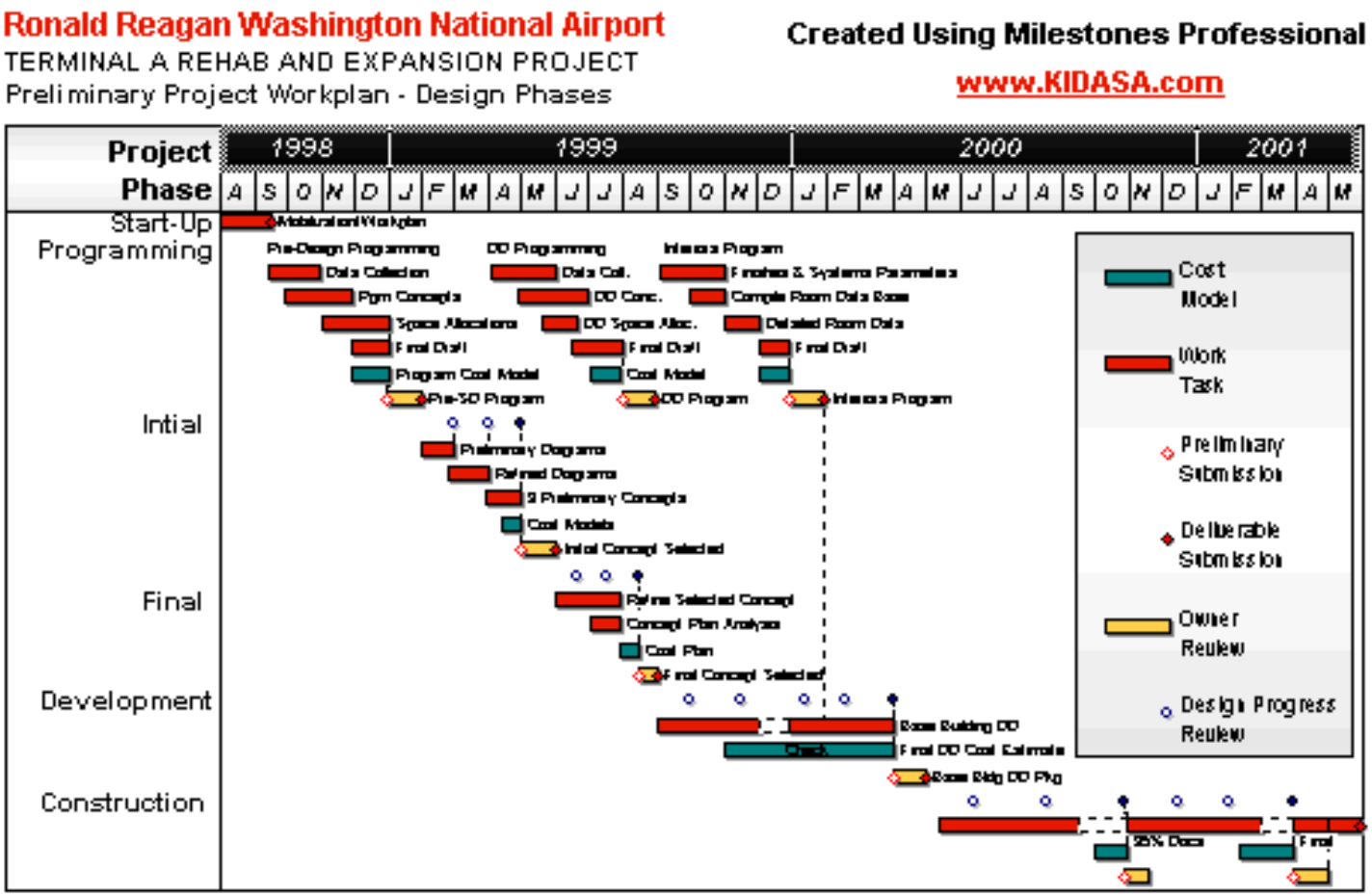
Project Planning

- Need to identify a project within the next week or two (by Jan. 20 at the latest)
- Talk to us if you have questions or need help
- Need to design a work plan that is logical, well-planned and feasible
- Need to set milestones and timelines
- Solution – the GANTT chart

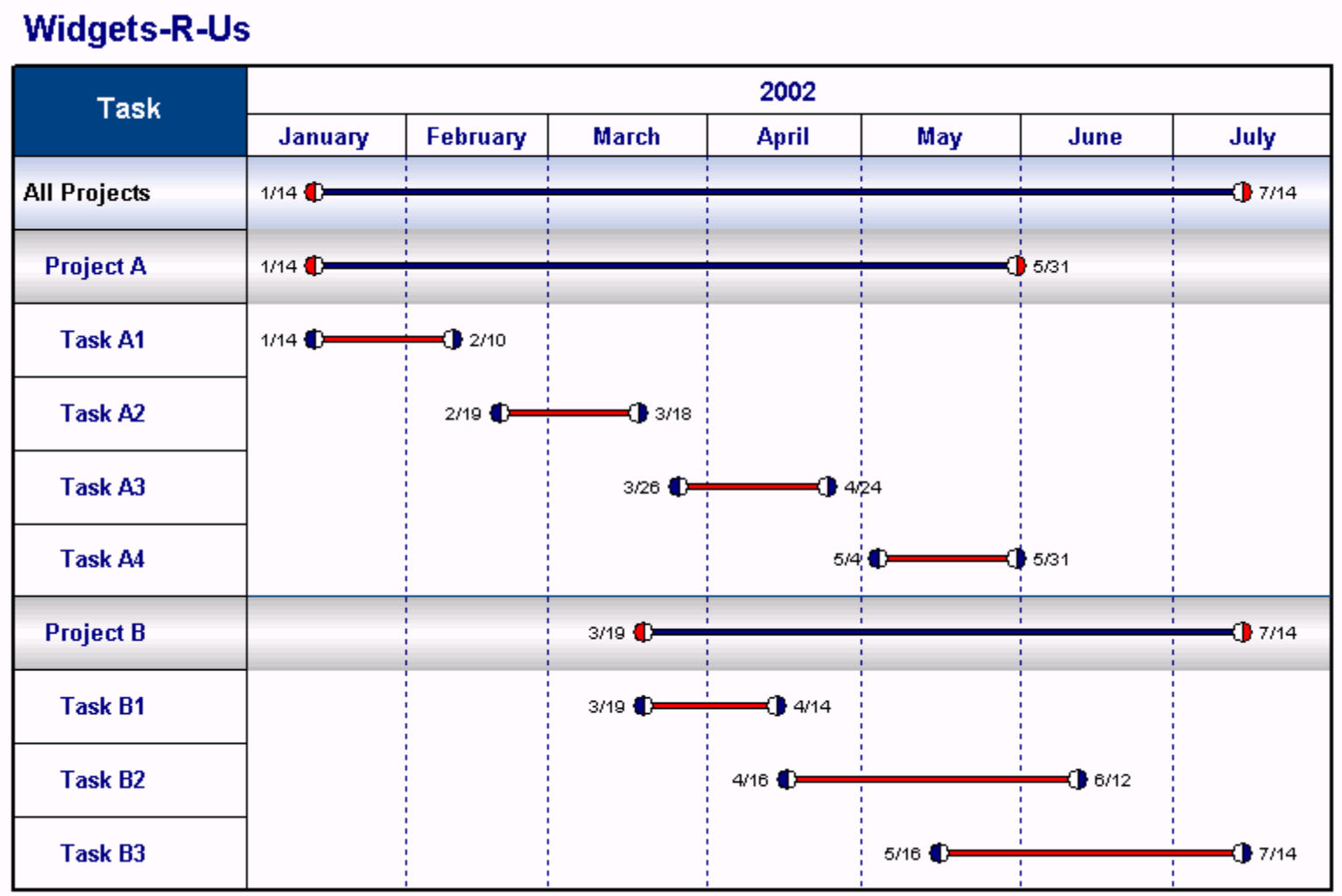
The GANTT Chart

- Invented by Henry Laurence Gantt (1861-1919)
- Mechanical engineer, management consultant and industry advisor
- Henry Gantt developed Gantt charts in the second decade of the 20th century
- Gantt charts are used as a visual tool to show scheduled and actual progress of projects

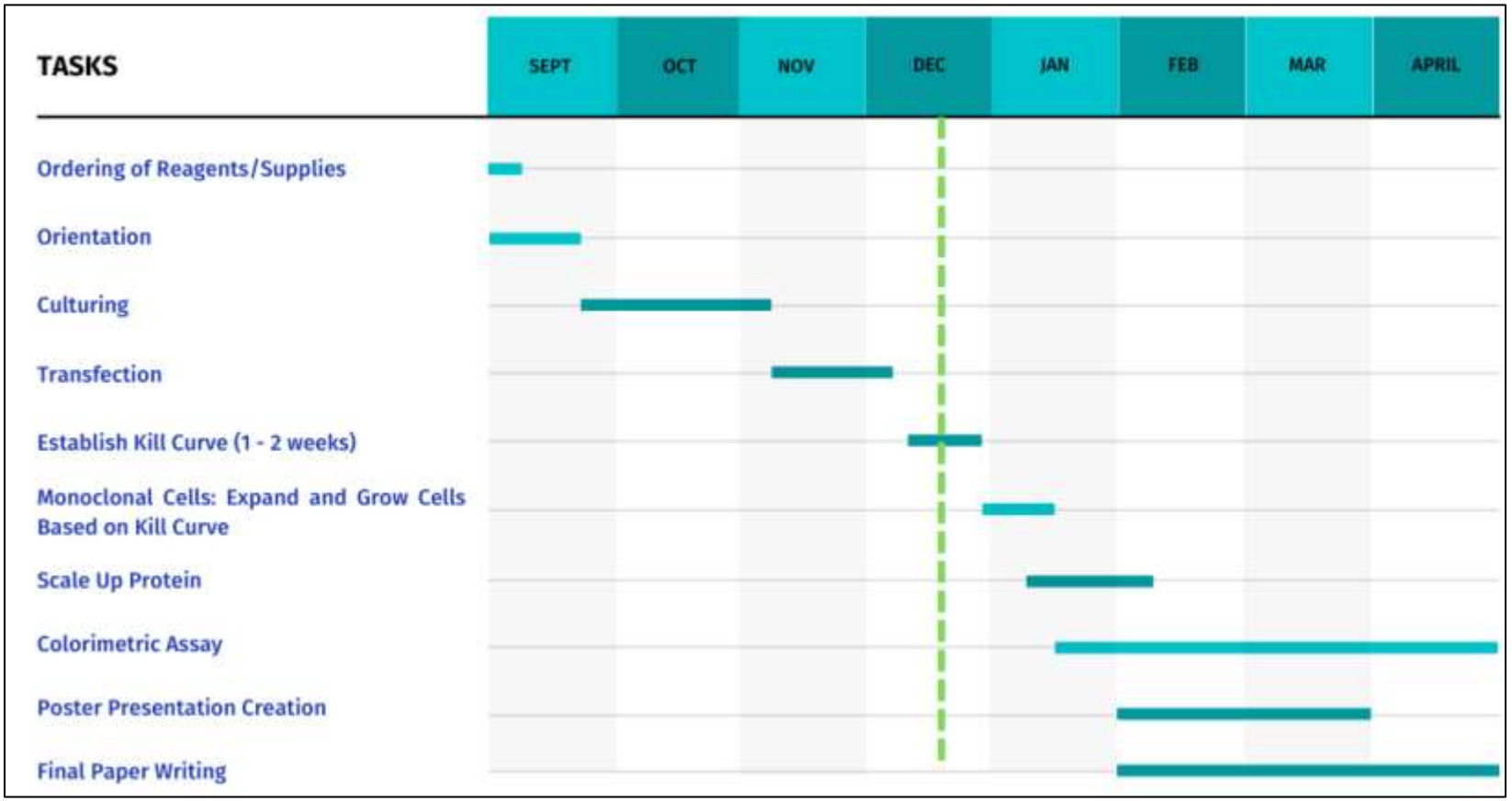
Sample Gantt Chart



Sample Gantt Chart



Sample Gantt Chart



401 Project

- One of the first tasks for your project is to build a GANTT chart for the project
- This is done in consultation with your coworker (and/or your project supervisor) and after reading some of the appropriate background literature on the project
- One of the best and simplest ways of building a good GANTT chart is to use Excel and to put text in some cells and fill other sets of Excel cells with different colors. Others will use the “table” feature in PowerPoint to generate high quality GANTT charts
- We would like to see GANTT charts for next class

Sample PPT presentation

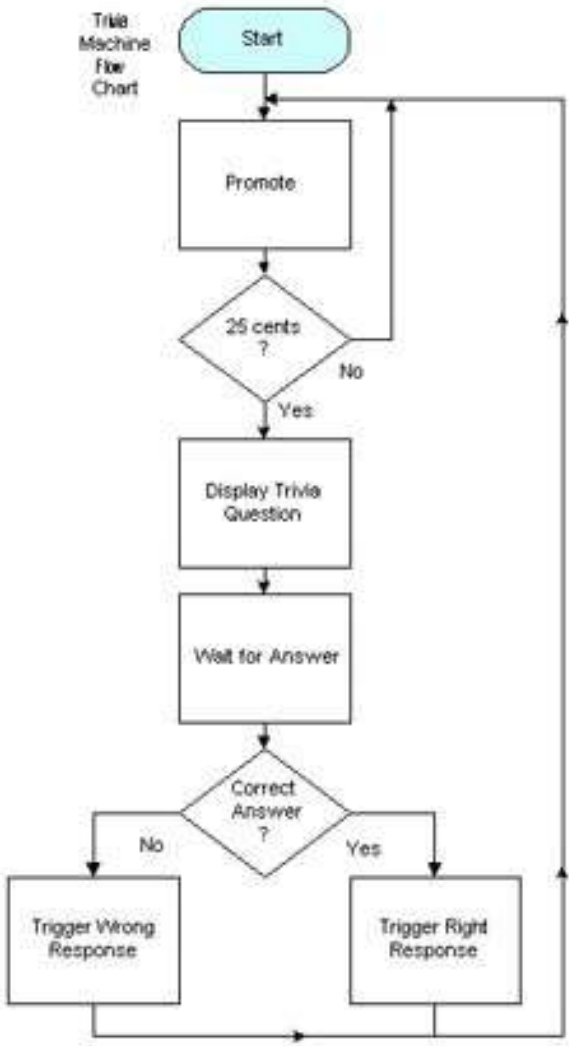
Your Name goes here

Date

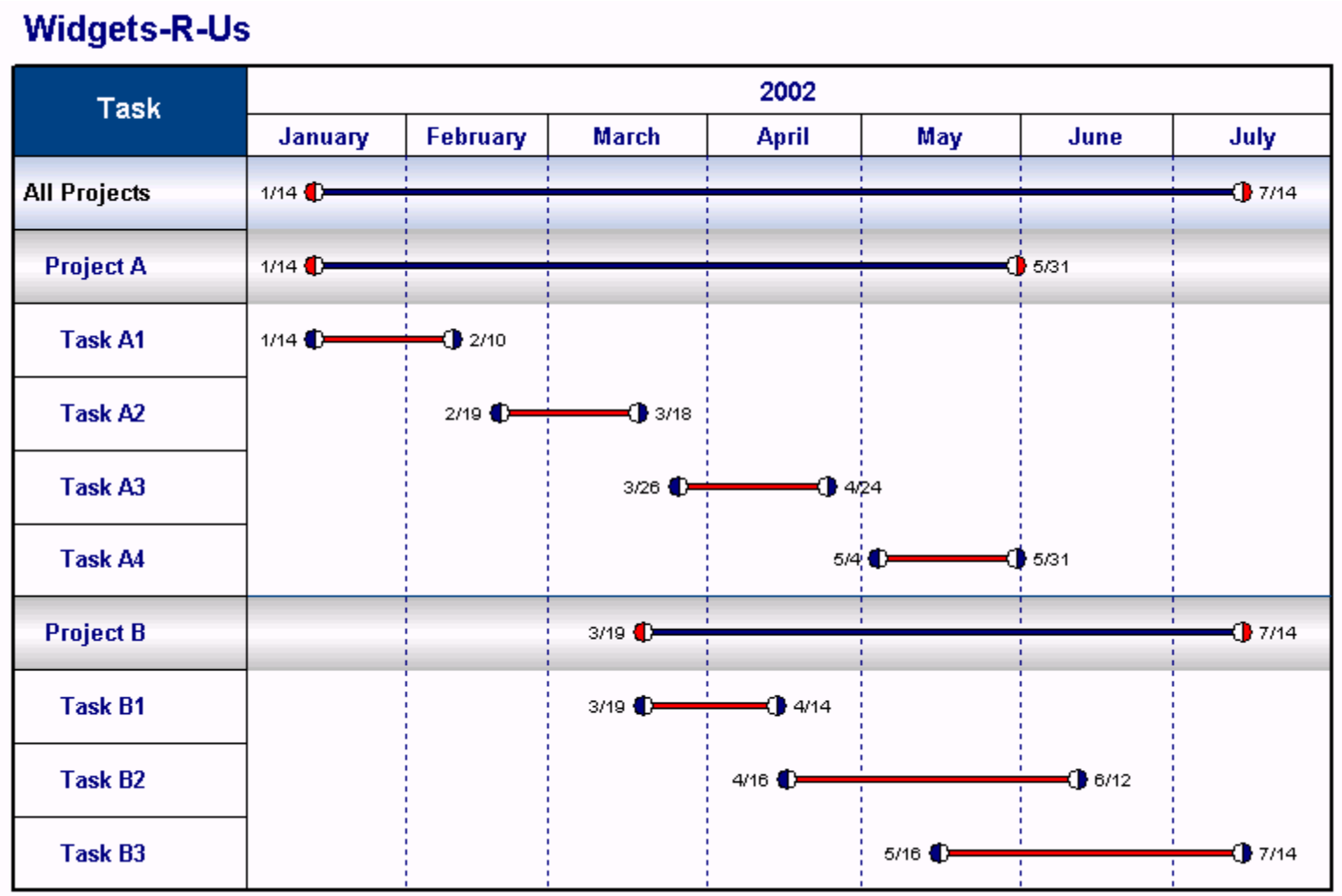
Title of Project

- Brief background and synopsis
- Project participants
- State of the art
- Key authors or papers
- Charts, figures to explain general concept

Title of Project (flow chart)



Sample Gantt Chart



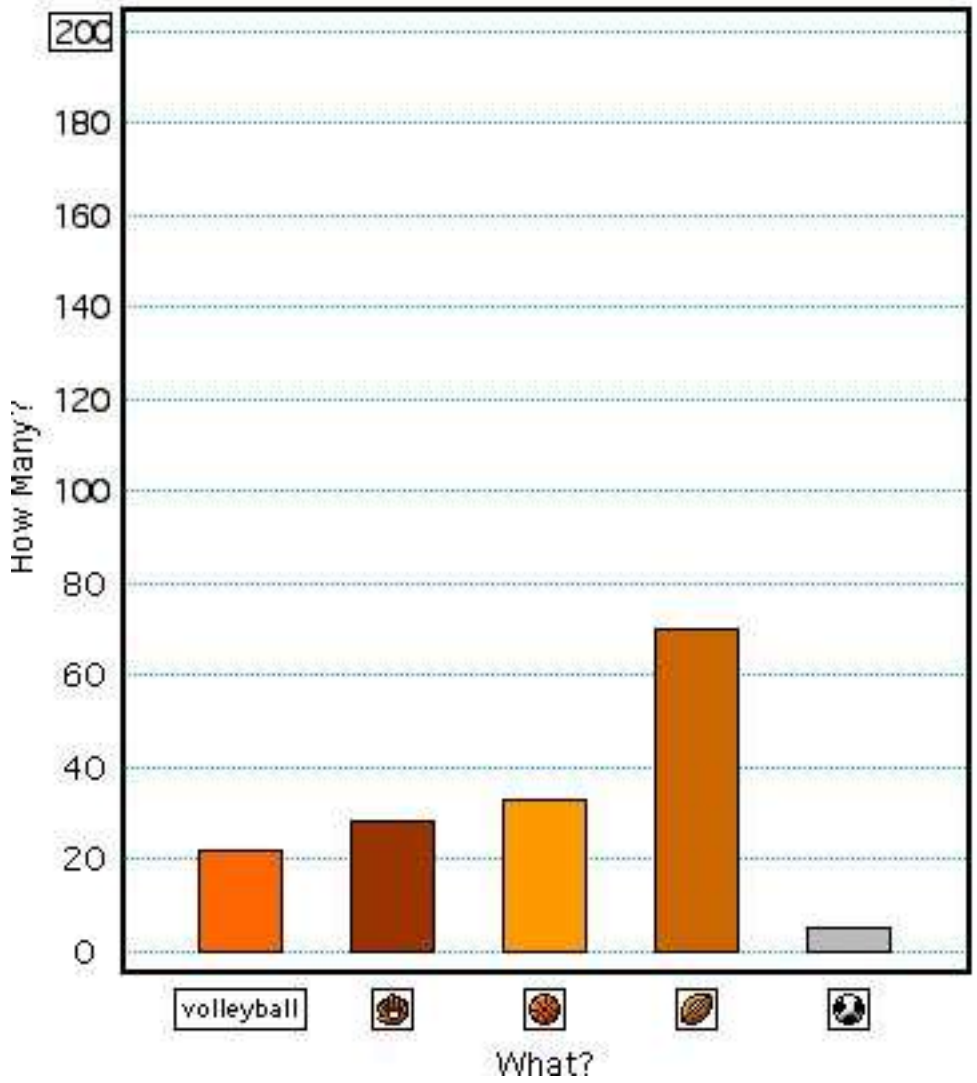
Title of Project

- **Summary of what you did last week**
- **Brief synopsis of progress or problems**

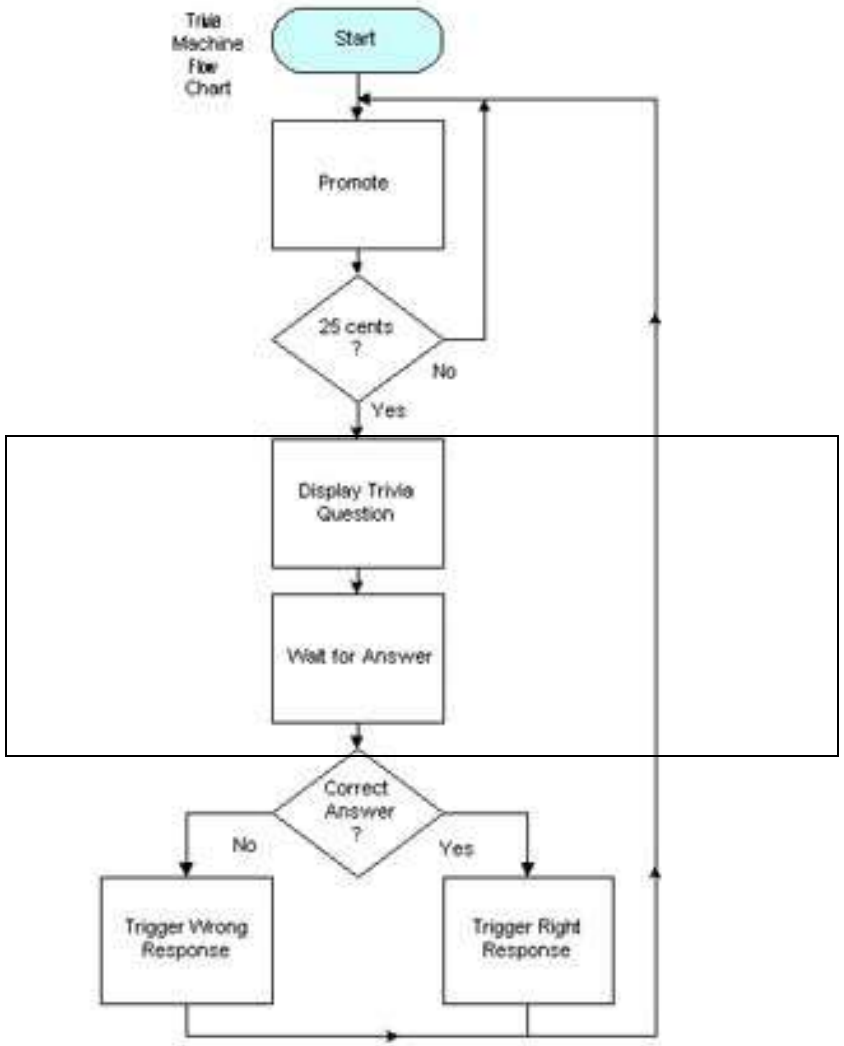
Title of Project

- Description of progress this week
- Discussion of successes
- Discussion of results, progresss

Graph (showing results or performance or program)



Flow Chart showing status



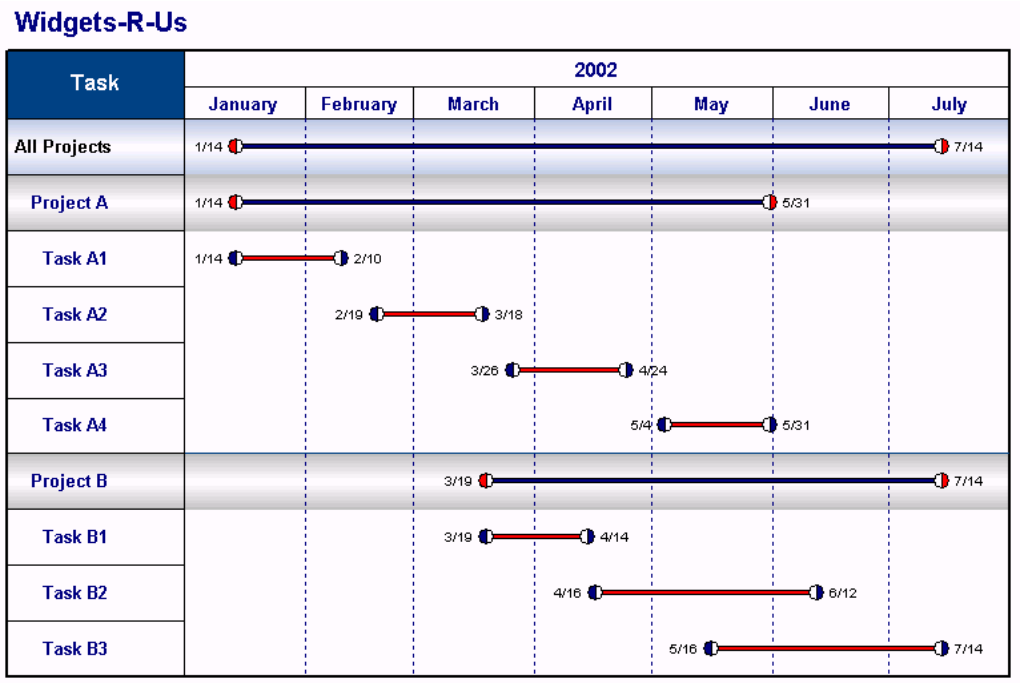
I’ m here

Title of Project

- Discussion of where you will go next week
- Changes in plan or direction
- Discussion of current or expected problems

Title of Project

- Conclusion, Discussion
- Gantt Chart reiterated



Bioin 401 HelpDesk

- **People who can help with some of your projects**
- **Located in Z8-30 (Zoology, 8th floor)**
- **Mark Berjanskii (mb1@ualberta.ca)**
- **Scott MacKay (xxx)**
- **Eponine Oler (eponine@ualberta.ca)**
- **Robyn Woudstra (xxx)**
- **Ray Kruger (xxx)**
- **Scott Han (xxx)**
- **Tanvir Sajed (xxx)**