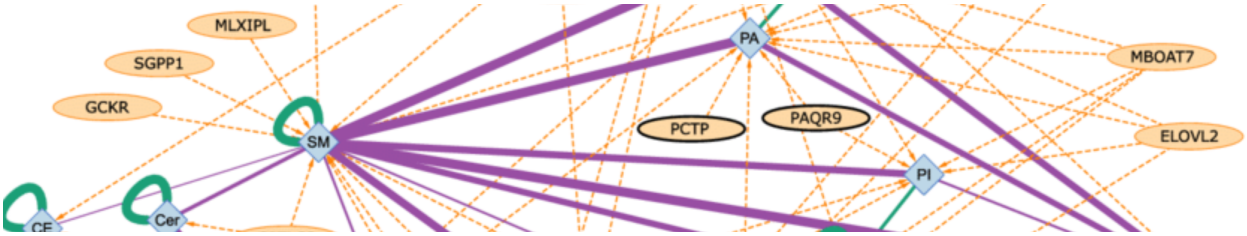
GAD: The Genetic Association Database

Prof. Rachlin

CS 3200: Database Design



DESCRIPTION:

In this assignment, we will explore GAD, the Genetic Association Database. [1] In the process we will re-discover some possible biological connections between seemingly disparate diseases.

While we sometimes think of genetic diseases as being associated with a single faulty gene, the truth is much more complex. Most diseases are "multi-genic" meaning that there are many genes which have been linked to the disease or disease phenotype. To say that there is a link or *association* between a gene and a disease means that some research study found a statistically significant connection between a *variation* of that gene, and the occurrence of a disease within a group of individuals. The association does not necessarily identify the underlying biological *cause* for the connection and the researchers who confirmed (or rejected) the association may have been focusing on a particular sub-population (Japanese, American Indian, etc.). Furthermore, the association may only suggest some increased *probability* of acquiring the disease. There may be other unknown connections with the environmental such as diet that are not fully understood.

GAD, the Genetic Association Database, is a simple catalog of research papers reporting an association (or lack of association) between genes and a disease. The gad table has the following columns:

gad_id – The table's primary key

association – Whether there is a positive association ('Y') or not ('N') or it is uncertain (' ')

phenotype – The name of the disease or condition

disease_class – The disease class (e.g., NEUROLOGICAL)

[1] Becker *et al.,* 2004. The Genetic Association Database. *Nature Genetics* **36**(5):431-2.

chromosome – The human chromosome where the gene can be found (1, 2, 3, …, 22, X, Y, M) *Careful! Chromosomes are not INT datatypes!*

chromosome_band – A descriptor for the chromosomal region

dna_start – The nucleotide position on the chromosome where the gene begins

dna_end – The nucleotide position on the chromosome the gene ends

gene – The official gene symbol

gene_name – The gene's full name

reference – The research paper where the association was reported

pubmed_id – The Pubmed ID (https://pubmed.ncbi.nlm.nih.gov)

year – Year of the publication

population – The population associated with the research study that reported the association.

The GAD table is imperfect. Sometimes the disease / phenotypes have inconsistent spelling or punctuation. Some columns contain *lists* of values rather than single values. (In other words, the data is not *normalized.*) In Data Science, sometimes we need to do the best we can with messy, incomplete, or inconsistent data.

INSTRUCTIONS:

Part A: Writing SQL Queries (80 Points)

Import the gad data (gad.csv) into a new database schema called **gad** having a single table, also called **gad**. Then open the attached START script and answer each of the first 12 questions with a SQL query. (Your answer to the 13th question can be typed into the script in the form of comments.

Part B: Research (20 Points)

Imagine you are a participant in a Northeastern-sponsored biomedical research conference: *Personalized Medicine in the 21st Century.* You've been invited to submit a poster consisting of a SINGLE POWERPOINT SLIDE. Your poster can be arranged however you like but should include:

a) A title + Your full name and Northeastern affiliation

b) A clear but broadly stated question along with a specific SQL query that you executed against the GAD dataset to investigate your question. YOU choose the question – try to make it interesting and engaging for fellow conference attendees![2]

c) One or more charts and graphs derived from your data. You may use ANY programming, visualization libraries, or charting tools you like to create your visualization including matplotlib, ggplot2, Microsoft Excel, python, java, R, etc.

d) A summary discussion of your results where you will *interpret* your visualization and draw any conclusions.

e) A properly formatted GAD citation.

I'll provide you with a PowerPoint slide template.

Your poster will be scored on the significance of the question, the clarity of your visualizations, and your written summary. The submissions will be compiled into the conference proceedings. Some authors may be offered the opportunity to discuss their results in class.

WHAT TO SUBMIT:

A .SQL script with the answers to each query for Part A. Please rename your script file: **cs3200_hw2_*yourname*.sql.** (No spaces.)

Your SINGLE-SLIDE POSTER in PDF format for Part B: Please rename your poster file **cs3200_hw2_poster_yourname.pdf.** (No spaces.) Posters submitted in the native PowerPoint (.pptx) format will not be accepted!!

[2] Science is sometimes like sales and marketing. You, the researcher, are trying to sell your insight, methodology, or discovery to your peers in the scientific community who buy and consume your ideas. Published research may involve a back-and-forth negotiation with a journal editor to convince them that your paper is even worthy of peer review. Like any marketing campaign, for your work to be widely accepted, it may need to be promoted at conferences and symposia, through social media, books, or magazines written for the public. Engaging in science can be both invigorating and brutal, but it is always a *profoundly human social endeavor*.