AI and Medical Data Types





AIMSS CURRICULUM 2021

Lecture A5 – April 3rd

Welcome!

Today I will talk about some complex health data-types and how AI can be applied to them.

Topics:

1. Electronic medical records (EMRs)
2. Medical Imaging and Video
3. EEG
4. Genetics, other -omics
5. Protein Folding
6. Drug Discovery

About the Presenter

- Sarah Davis
- AIMSS Co-President
- Computing Science M.Sc. Student
  - Alberta Machine Intelligence Institute
- B.Sc. Biological Sciences w/ Computing Science Minor
- Email: sdavis1@ualberta.ca

Disclaimers and Disclosures

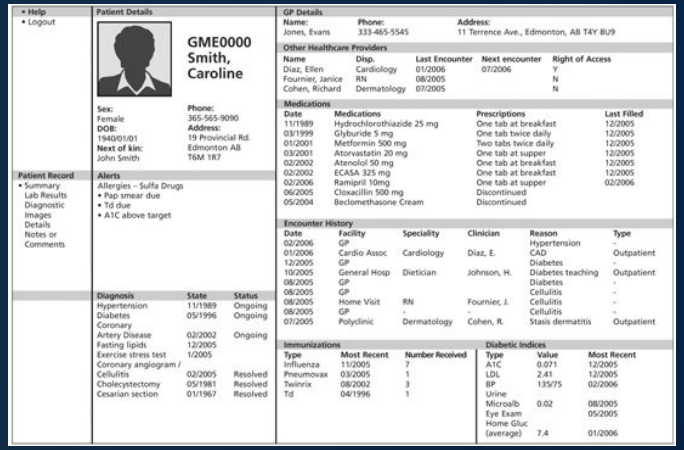
1. Western (specifically Canadian, Albertan) perspective when it comes to examples, history
2. I have worked in the analysis of (most of) these data types, no experience in generating them
3. Much of the information in this presentation was gathered from recent review articles (links to those and more can be found in the speaker notes)

ELECTRONIC MEDICAL RECORDS

AIMSS AI in Healthcare Curriculum

Lecture A5 — Medical Data Types

EMR Introduction



- Key idea: centralized and standardized health records
- Centralized medical records date back to 1907, system designed for the Mayo Clinic
- "Electronic" part integrated in the 1970s
  - First EHR system supported both nursing clinical documentation and physician computerized order entry

Core Functions of an EMR:

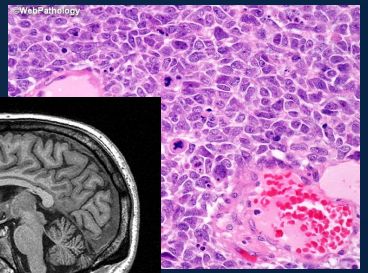
1. Health information and data
2. Results management
3. Order entry/management
4. Decision support
5. Electronic communication and connectivity
6. Reporting

What data does an EMR include?



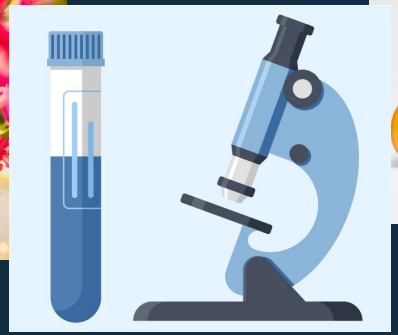




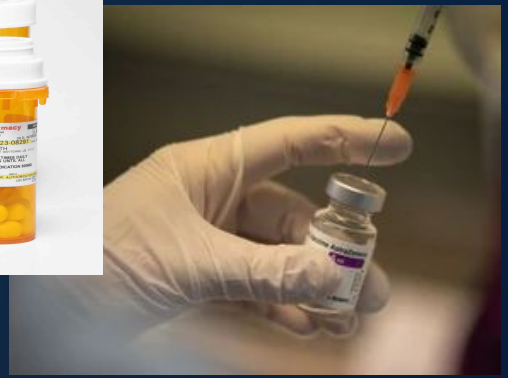












Advantages of EMR Data

- Both broad and deep
  - Contains thorough detail about an individual
  - Does this for (often) millions of patients!
- Deep learning often works best with this size of data
- This data can be used to learn both about individuals, populations, healthcare operations management, and more
- The data itself is a byproduct, we aren't spending extra money to generate it for research purposes



Drawbacks of EMR Data

- Completeness/quality/integrity issues
- Incredibly sensitive information
- Sheer scale
- Many different data types
- Longitudinal (how do we deal with differing hospital usage?)

Patient 1 = [🔶🟩]

Patient 2 = [🔺🔶🔶🔶🔵💜🔺🔶🔶⭐🟩🟩🟩🔶🟩⭐🟩🔵💜🔺🔶🔶⭐🔶
⭐🟩🟩🔶🔶⭐⭐🔶🔵💜🔺🔶⭐⭐🟩🔵💜🔺⭐🟩🟩🔺🔶🟩
🔵💜🔺🔶⭐🔵💜🔺🔶🟩🔵💜🔺🔶🔶⭐🟩🟩🟩🔶🟩⭐🟩🔵]

🔺 - Admitted
🔶 - Diagnosis
⭐ - Procedure
🟩 - Prescription
🔵 - Discharged
💜 - Time passes



How can AI help? (Some examples)

- Data entry and extraction
  - predicting and streamlining physician orders using recommender algorithms
- Optimizing care
  - identifying high-cost low-yield laboratory tests using multivariable prediction models
- Clinical outcome prediction
  - predicting patient-specific clinical events using recurrent neural networks



MEDICAL IMAGING AND VIDEO

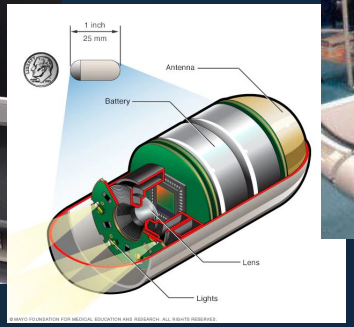
AIMSS AI in Healthcare Curriculum

Lecture A5 — Medical Data Types

Introduction

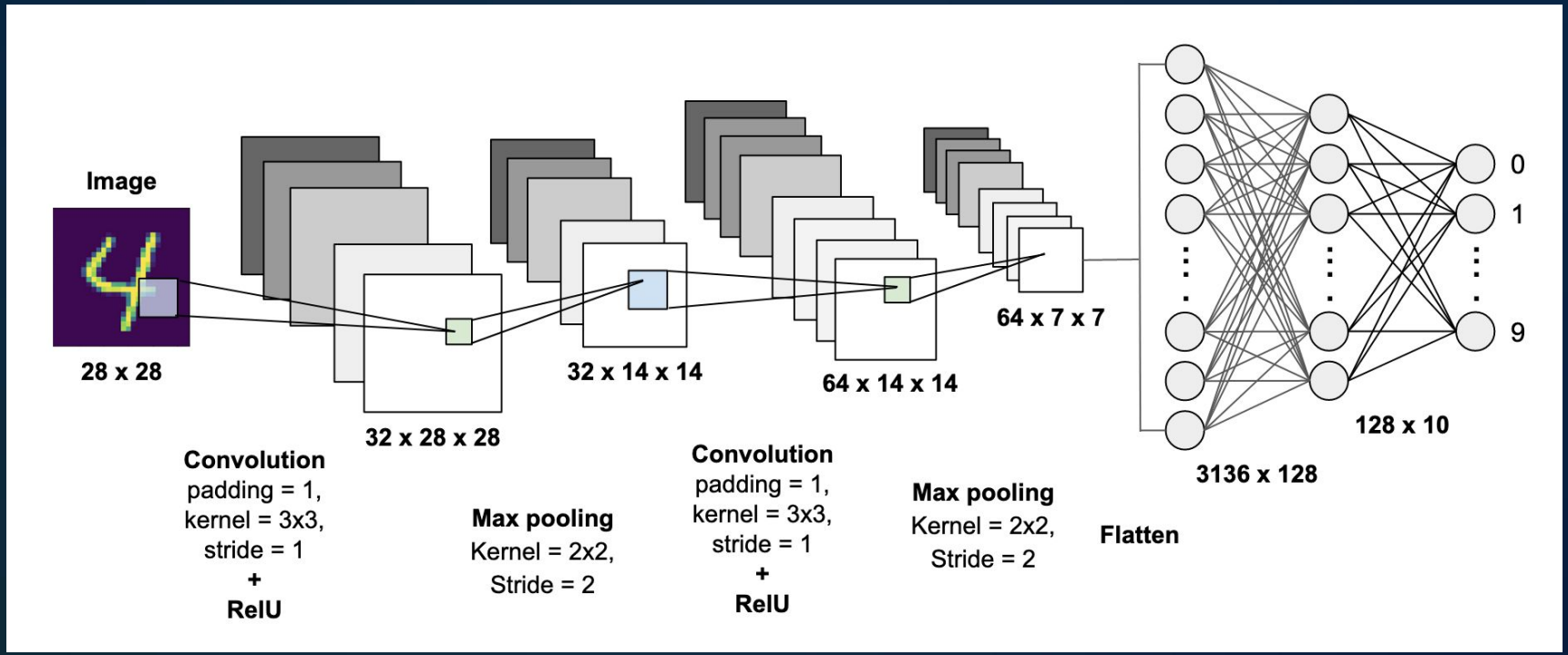
- With recent sensing technology, we have been able to capture minute medical detail (micro and macro)
- Important related concept: Convolutional Neural Network







Convolutional Neural Networks



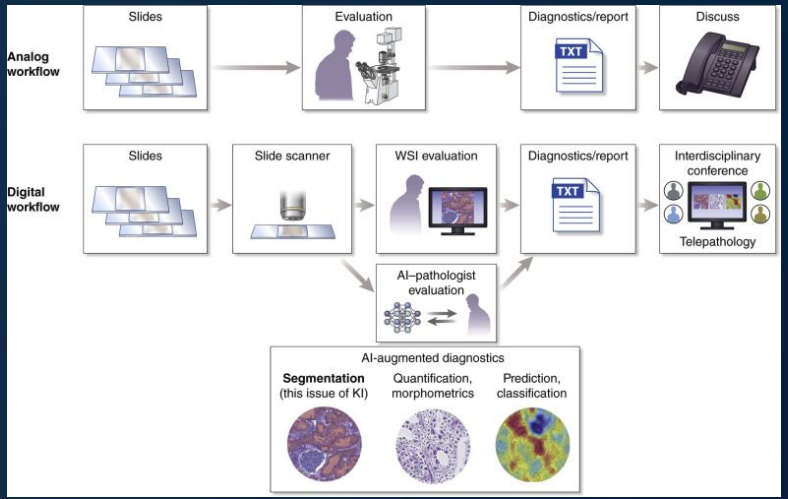
Medical Imaging

- Refers to the practice of using technology to view the human body to diagnose, monitor, or treat medical conditions
- Includes ultrasound, MRI, X-ray, CT, PET, etc. — each work best in certain body tissues
- Example AI applications:
  - Diagnosing macular degeneration from fundus images with deep CNNs
  - Classification of pulmonary tuberculosis from chest X-rays with CNNs
  - medo.ai



Histology

- Study of the microscopic anatomy of biological structures
- Example AI applications:
  - Automated brain histology classification using the Google Inception V3 CNN
  - Digital nephropathology by multistain segmentation using CNNs
  - pathai.com/what-we-do/



Medical Video



- Can be used in many different ways – from endoscopy to mental health interviews
- Video is just a sequence of images (frames)
- Example AI applications:
  - Video-based detection of the clinical depression in adolescents using Gabor wavelets at facial landmarks
  - Video-based AI for beat-to-beat assessment of cardiac function using CNNs trained on echocardiogram videos

AIMSS Summer Project

- Using deep-learning object detection techniques to track whereabouts of medical tools from video feed





Drawbacks

- For video and audio, some people may be uncomfortable having themselves and/or their voice captured
  - Could be solved by “streaming”
- Medical imaging equipment is often very expensive
- Creating labeled datasets is often difficult, as experts don’t always agree
  - Datasets are often small (solutions?)



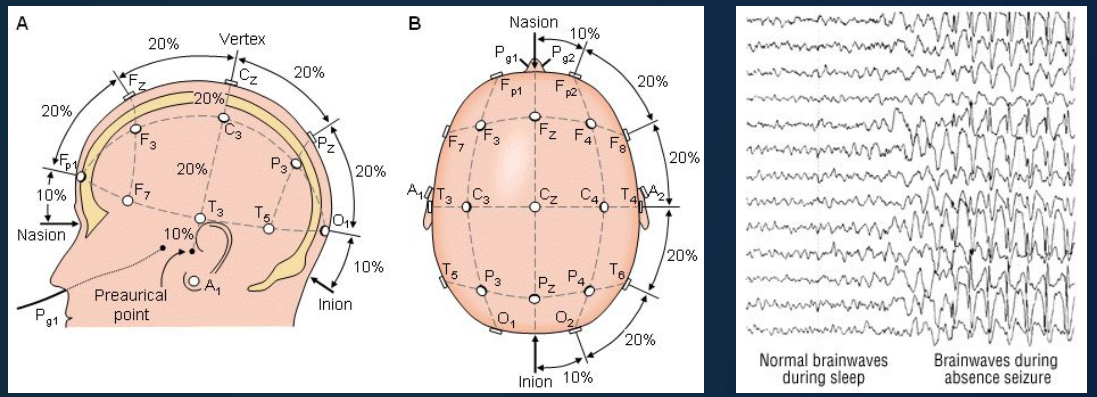
EEG

AIMSS AI in Healthcare Curriculum

Lecture A5 — Medical Data Types

EEG Data Introduction

- Electroencephalography (EEG) data measures brain waves
  - Voltage comes from ionic current within and between neurons
  - More activity → higher voltage
- ERP (Event Related Potentials): brain waves in response to particular stimulus



Advantages of EEG Readings

- Medical side: Helpful in diagnosing conditions such as epilepsy, dementia, parkinson's, head injuries, etc.
- Research side: used in psychology, sociology, linguistics, neuroscience, market research, biomedical engineering and more; used to learn about the brain itself
- Technology can be relatively inexpensive







Roll over image to zoom in

Muse 2: The Brain Sensing Headband

Visit the MUSE Store

4.5 ratings | 110 ratings | 22 answered questions

Amazon's Choice for "muse"

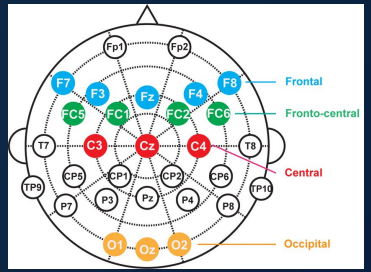
Price: CDN$ 299.99 prime FREE One-Day

Get a **$35 Amazon.ca Gift Card** instantly, plus up to 5% back for 6 months after approval for the Amazon.ca Rewards Mastercard.
Pay ~~$299.99~~ **$264.99** for this order after approval.

Secure transaction | Returns Policy | Amazon Delivered

Drawbacks of this Data Type

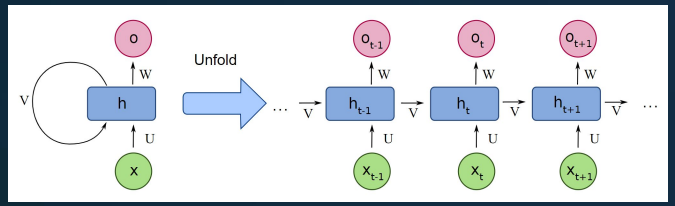
- EEG data is highly noisy
  - Requires a significant amount of preprocessing
  - Eye blinking, ambient radiation, technology quality...
- Hard to tell exactly where the activity is coming from

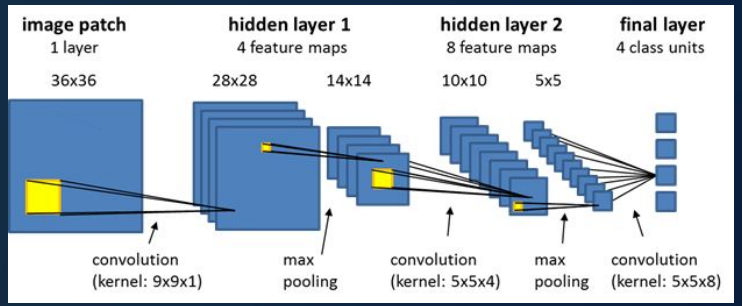


- Suffers from longitudinal data problem
  - How do we standardize the length for machine learning?

AI and EEG

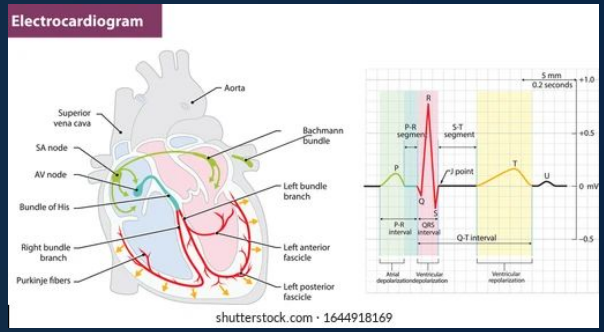
- Predicting learning difficulties based on EEG waveforms with a K-nearest neighbours approach
- Abnormal EEG reading identification using recurrent neural networks
- EEG decoding and visualization using convolutional neural networks



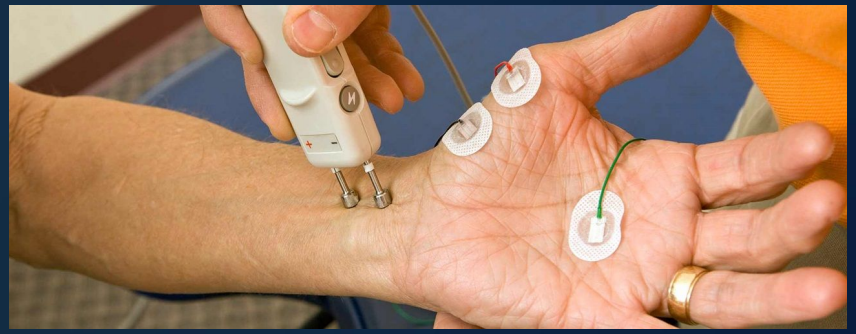


More Electrical Activity Measurements

- Electrocardiogram (ECG)
- Electromyography (EMG)
- Nerve conduction study (NCS)

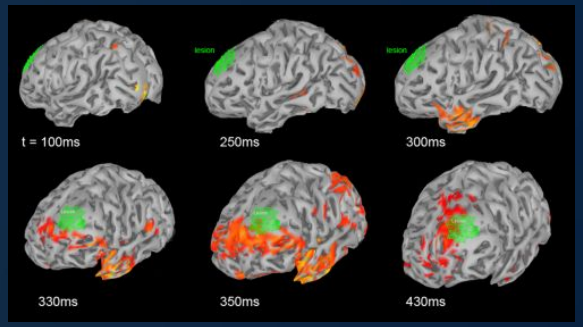


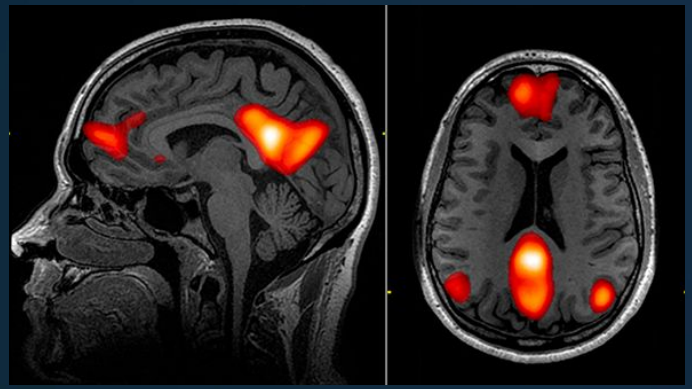


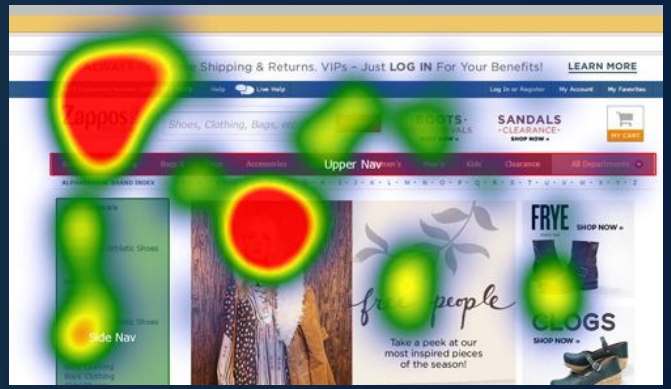


Other Way to Measure Brain Activity

- Magnetoencephalography (MEG)
- Functional magnetic resonance imaging (fMRI)
- Eye tracking*





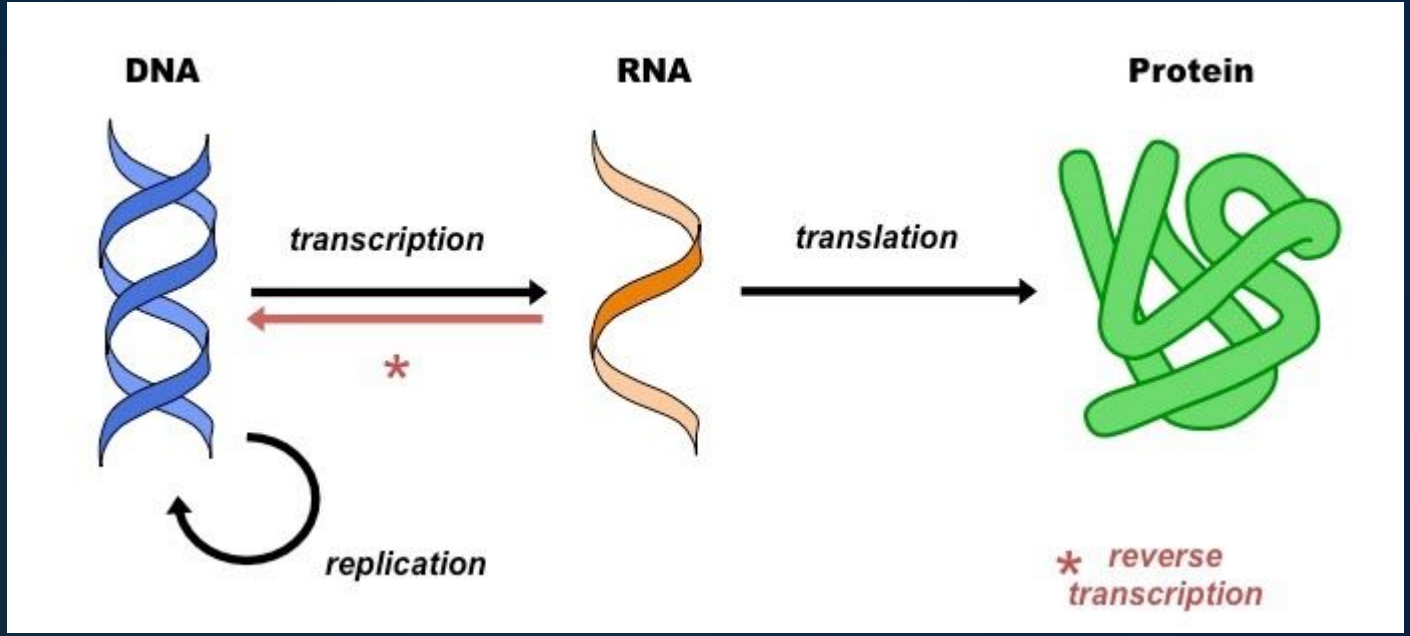


GENETICS AND OMICS

AIMSS AI in Healthcare Curriculum

Lecture A5 — Medical Data Types

Introduction



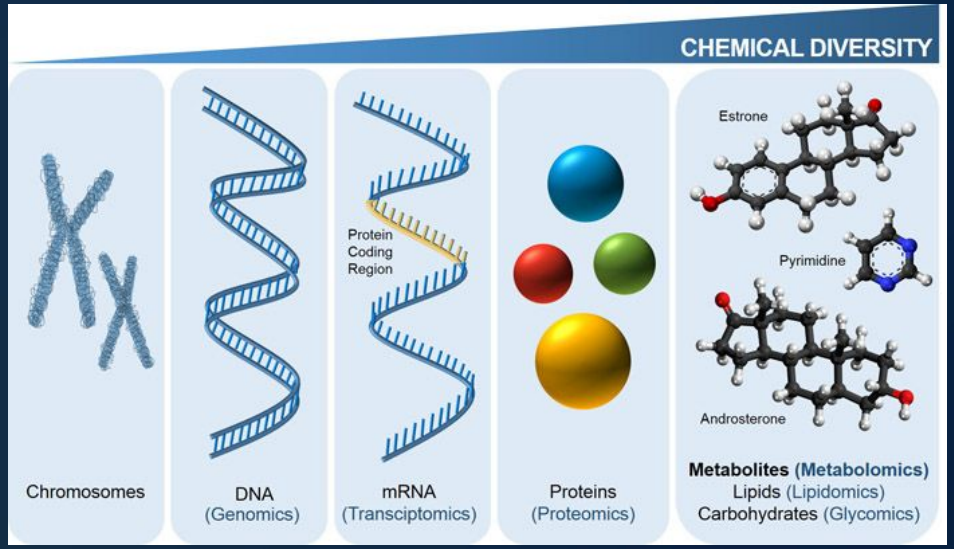
Introduction

- Genetics: *"scrutinizes the functioning and composition of the single gene"*
- Genomics: *"addresses all genes and their interrelationships in order to identify their combined influence on the growth and development of the organism"*



Other -Omics

- Transcriptomics
- Proteomics
- Metabolomics
- Lipidomics
- Epigenomics
- …



More useful → Harder to measure!

Advantages of this Data Type

- Fully understanding genetics (and omics) promises a significant change in how we do medicine
  - We have much to gain in the realm of personalized health

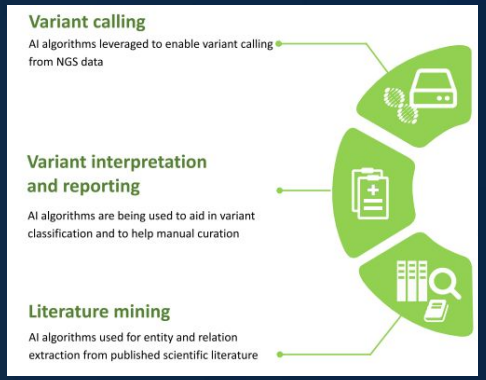


Drawbacks of this Data Type

- Sizable technology barrier currently
  - Leads to data quality problems! (e.g. noise)
  - Worse in the measurement of “useful” data
  - Worse in affordable technology
- Huge number of features (approx. 20k genes…)
- Batch effects
- Reproducibility level is low
- Not everything is genetic! Nature vs Nurture



AI and Omics



- Applications of AI in Cancer Genomics:
  - Variant calling using CNNs, random forest regression
  - Variant interpretation using deep neural networks
  - Literature mining using entity extraction (from NLP)
- Predicting endometriosis from transcriptomic and methylomic data with decision trees, support vector machines, and random forest classification
- Predicting acute kidney injury in coronary angiography recipients and finding proteomic biomarkers using LASSO regression

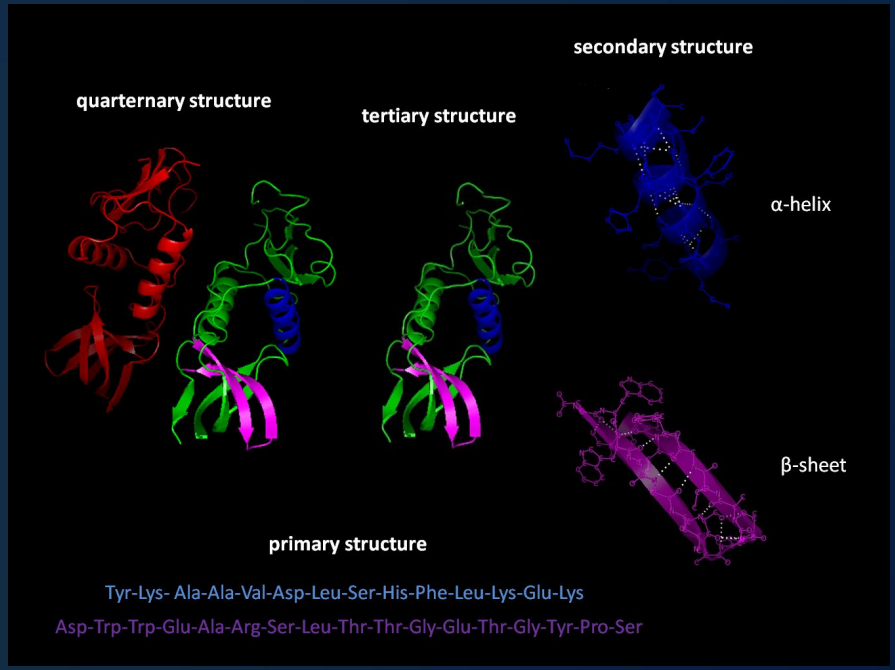
PROTEIN FOLDING

AIMSS AI in Healthcare Curriculum

Lecture A5 — Medical Data Types

Introduction

- Protein folding happens all the time in your body
- Problem: how do we infer higher-order protein structures from lower-order protein structures?
  - i.e. skip a step



Advantages

- Enormous downstream potential in drug discovery (next section)
- Model genetic variation into a protein structure
  - Gain insight into biological effects of altering sequence
  - Implications in genetic engineering
- Applications in other fields too (such as computer graphics)

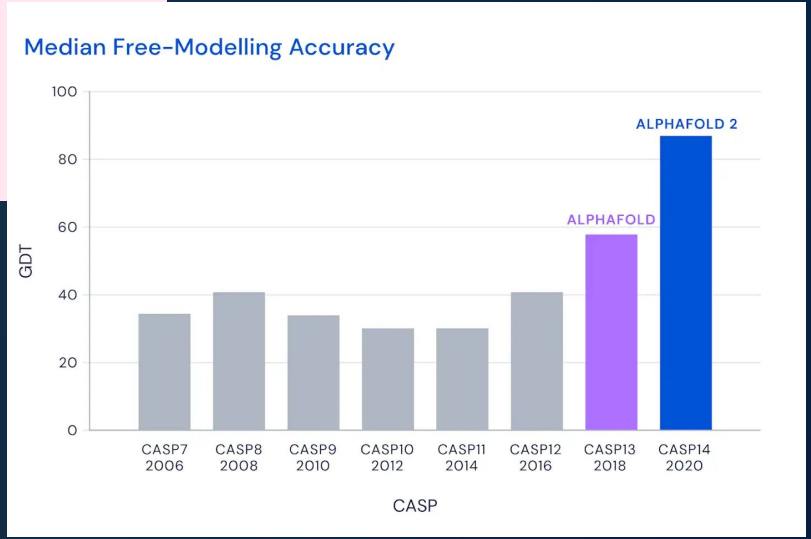


Drawbacks

- Protein folding is an extremely difficult problem
  - Some scholars have argued that this problem is NP-complete (thus unlikely to have an efficient solution)
- Labeled databases are small and very expensive to curate
  - Solving using crystallography and NMR requires specialized training, skill, and luck
- Protein folding depends on environment!



AlphaFold: a solution to a 50-year-old grand challenge in biology



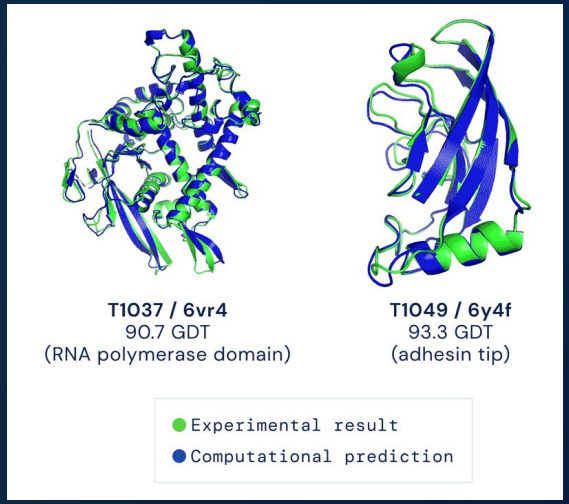
AlphaFold

- Google DeepMind's model for predicting protein structure
  - Part of the biannual CASP competition
- CASP uses the Global Distance Test (GDT) metric (1–100)
  - % amino acid residues within some threshold of actual position
  - GDT of 90%+ considered "functionally solved"
- AlphaFold had a median score of 92.4 GDT
  - Used an attention-based neural network system



AlphaFold

- Results are impressive! But…
- AlphaFold 2 did not actually "solve protein folding"
- Their median score was 92.4
  - Not all predictions above 90 GDT
- Does not work on all types of proteins
- Does not "explain" how proteins fold



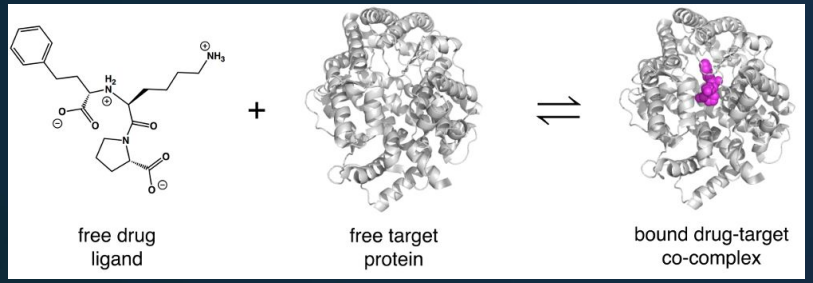
DRUG DISCOVERY

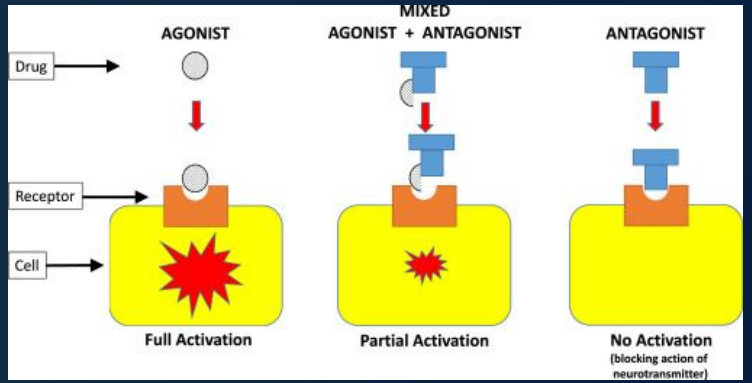
AIMSS AI in Healthcare Curriculum

Lecture A5 — Medical Data Types

Introduction

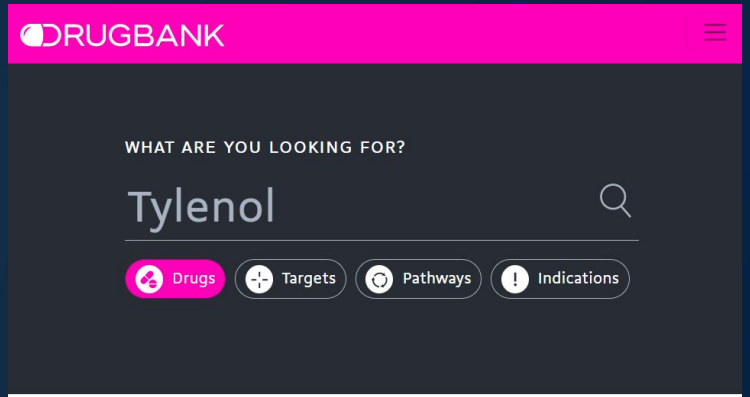
- Drugs work by interacting with receptors on cells in ways that mimic or inhibit natural biological molecules
- Drug discovery often involves large-scale screening assays where 10s or 100s of thousands of compounds are tested per day





Data Sources and Databases

Database	Description
PubChem	PubChem is a database of chemical molecules which maintains three types of information namely, substance, compound and BioAssays.
ZINC	ZINC database contains 21 million compounds available for virtual screening. In this database various molecule features like molecular weight, logP etc are included.
ChEMBL	This database provides comprehensive information about 1 million bioactive (small drug-like molecules) compounds with 8200 drug targets.
NCI	NCI database had more than 275,000 small molecules structures, a very useful resource for researchers working in the filled of cancer/AIDS.
ChemDB	It is a databse of five million chemicals which contains information of chemicals that includes predicted or experimentally determined physicochemical properties, such as 3D structure, melting temperature and solubility.
ChemSpider	ChemSpider contains more than 28 million unique chemical entities aggregated from more than 400 diverse data sources.
BindingDB	It is a binding affinity database of small molecules which contains 910,836 binding data, for 6,263 protein targets and 378,980 small molecules.
PDB-Bind	It is a collection of binding affinities for protein-ligand complexes with known three-dimensional structures. It contains 5671 protein-ligand complexes.
PDBeChem	It provides comprehensive information of ligands, small molecules and monomers. Presently it consists 15502 ligands.
KEGG	It is a database resource that integrates genomic, chemical and systemic functional information.
HMDB	A database containing detailed information about small molecule metabolites found in the human body.
SMPDB	It is Small molecule Pathway database that containing more than 350 small-molecule pathways found in humans.
BIAdb	It is a comprehensive database of benzylisoquinoline alkaloids which contains information about 846 unique benzylisoquinoline alkaloids.
DrugBank	The database that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The database contains 6712 drug entries including 1448 FDA-approved small molecule drugs, 131 FDA-approved biotech (protein/peptide) drugs, 85 nutraceuticals and 5080 experimental drugs.
HIT	HIT is a comprehensive database for protein targets for FDA-approved drugs as well as the promising precursors. It currently contains about 1,301 known protein targets(221 proteins are described as direct targets).
SuperNatural	A freely available database of approximately 50,000 natural compounds.
NPACT	It contains experimentally validated plant-derived natural compounds exhibiting anti-cancerous activity. Currently it contains 1574 compound entries.
TTD	This database provide information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information and the corresponding drugs directed at each of these targets.
PharmaGKB	It is a pharmacogenomics knowledge resource that encompasses clinical information of drug molecules.
SuperDrug	This database contains approximately 2500 3D-structures of active ingredients of essential marketed drugs.

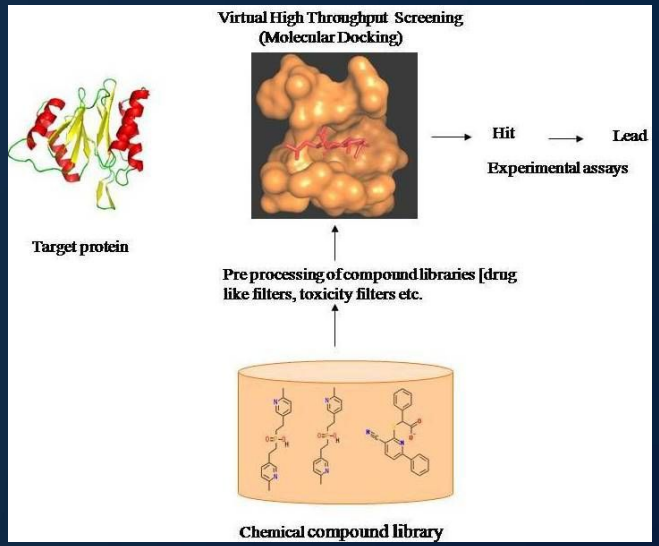


DRUGBANK

DrugBank is a pharmaceutical knowledge base that is enabling major advances across the data-driven medicine industry.

Virtual High-Throughput Screening

- Virtual High-Throughput Screening gained popularity in early 2000s
- Allows experimenters to discard very poor fits to save time and $$$
- Drawbacks:
  - Very sensitive to quality in compound libraries
  - Requires manually engineered features and parameters



AI and Drug Discovery

- Predicting physical properties of drugs:
  - Designing molecules with desired properties using deep neural networks and reinforcement learning
- Predictions of bioactivity
  - Matched molecular pair analysis using random forest, gradient boosting machines, and deep neural nets
- Predictions of toxicity
  - DeepTox: Toxicity prediction using deep neural nets



Questions?

Thank you!

Stay in Touch

Email: sdavis1@ualberta.ca
LinkedIn: linkedin.com/in/arahdavi
Twitter: @arahdavi

Check out linktr.ee/aimss for everything club related!

