ScHARe

Science Collaborative for Health disparities and Artificial intelligence bias REduction

Artificial Intelligence in Data Science as a Broad New Horizon

A. AI Fundamentals

B. Computational Strategies: forms of AI that may not be conventionally referred to as machine learning… e.g., Generative AI & other forms of Deep Learning (DL)

Generative AI

Generative AI is a subset of DL models that generates content like text, images, or code based on provided input. Trained on vast data sets, these models detect patterns and create outputs without explicit instruction, using a mix of supervised and unsupervised learning.

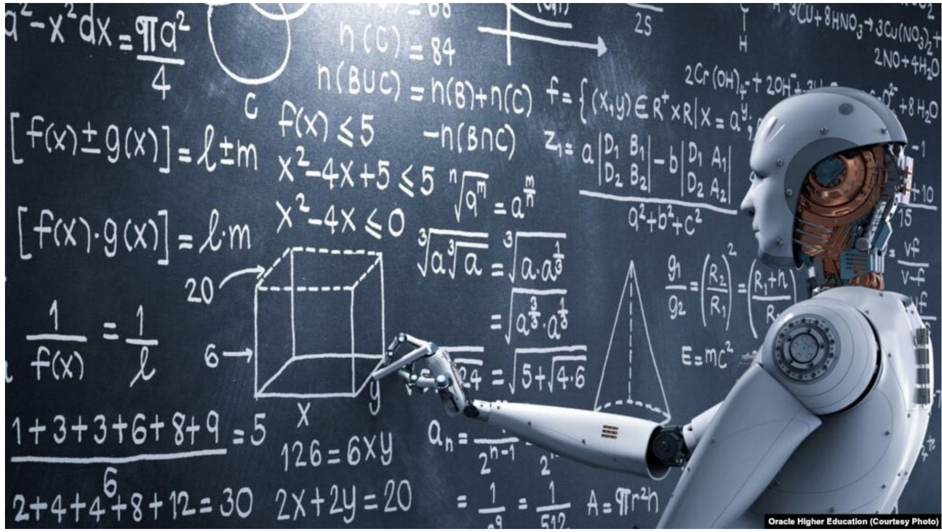
DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH National Institute of Diabetes and Digestive and Kidney Diseases

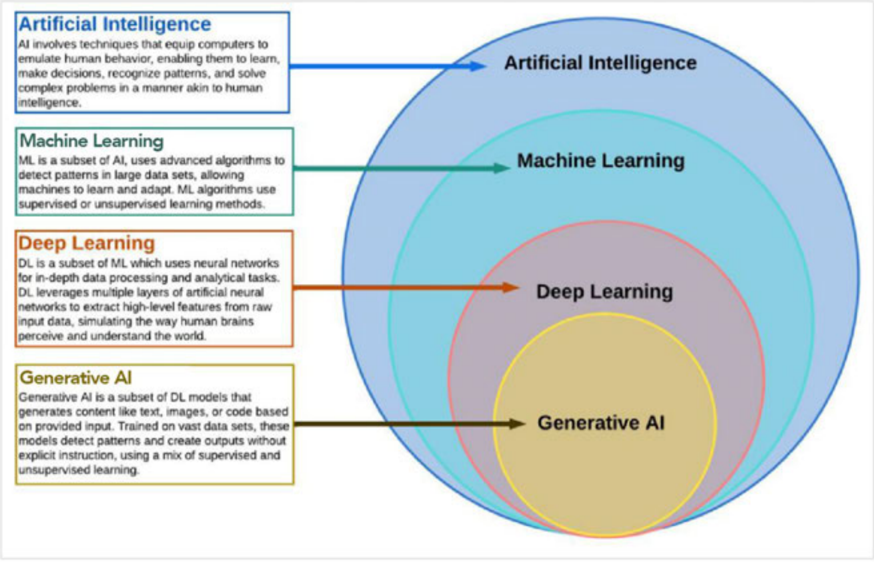
Artificial Intelligence Fundamentals: Definitions

- Definitions (& *distinctions* with specific subset of ‘machine learning’)
  - ***OURS:*** NIH Strategic Plan for Data Science (2018-2023*):
    - Artificial Intelligence: “the power of a machine to copy intelligent human behavior”
    - Machine Learning: “field of computer science that gives computers the ability to learn without being explicitly programmed by humans”

*NOTE: NIH Strategic Plan for Data Science **2023-2028** (in revision, [open for public comment]())



Oracle Higher Education



https://doi.org/10.3390/su151813484

Artificial Intelligence Fundamentals

- *Despite all the potential that AI has, and compelling performance shown… remain humble: per quote selected by an AIM-AHEAD leader*

“Say not, “I have found the truth,” but rather, “I have found a truth.”

— **Kahlil Gibran**

[sli.do questions]

Artificial Intelligence Computational Strategies

1. Natural Language Processing (NLP) for Text Mining:

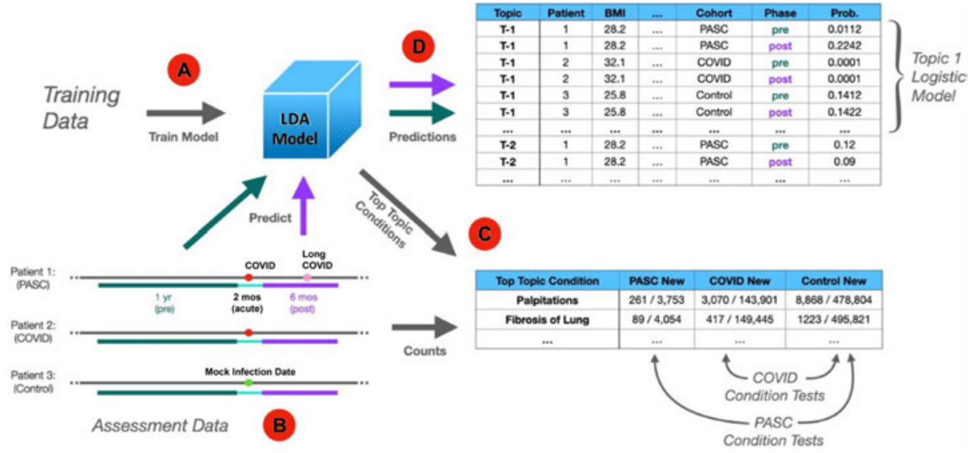
- a. Strategy: Extracting meaningful insights from large volumes of unstructured text data, such as medical literature, clinical notes, or patient narratives.
- b. Applications: Analyzing patient experiences, identifying disparities in healthcare narratives.
- c. Python Libraries: NLTK, SpaCy, gensim.
- d. Large Language Models: GPT, Llama

Artificial Intelligence Computational Strategies

1. Natural Language Processing (NLP) for Text Mining:

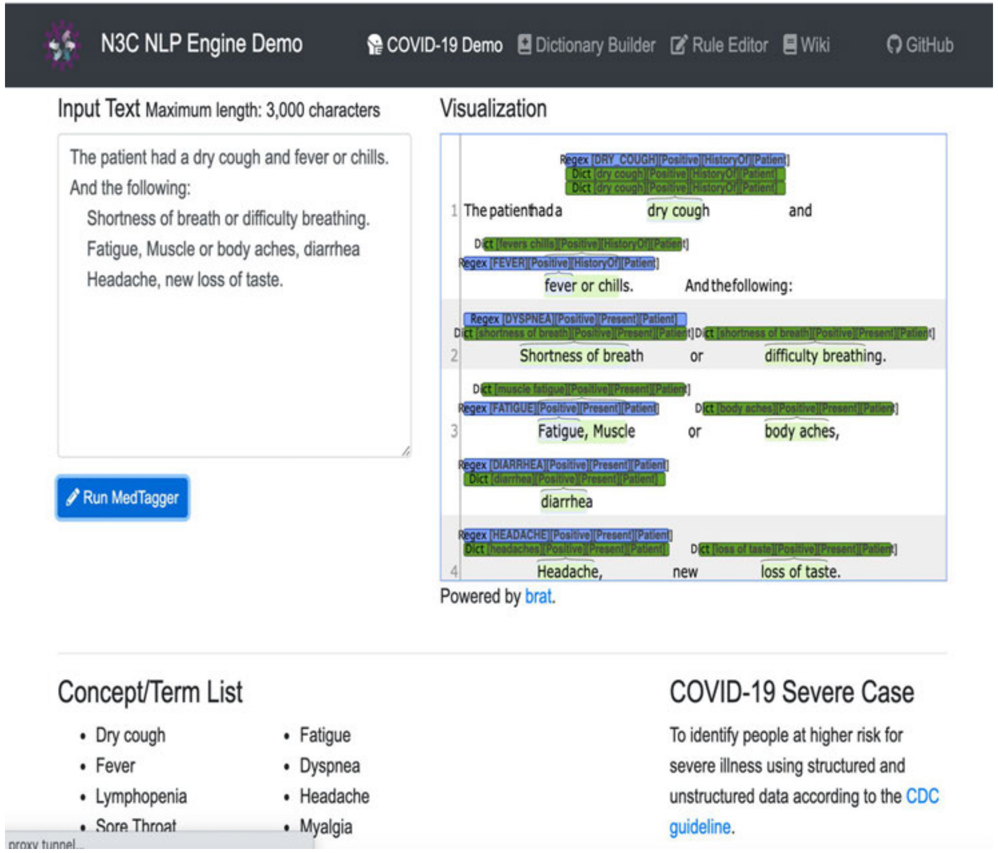
- b. Application **examples**:
  - Analyzing patient experiences
  - identifying disparities in healthcare narratives
  - *Classifying diagnostic coding of comorbidities*.

Finding Long-COVID: Temporal Topic Modeling of Electronic Health Records from the N3C and RECOVER Programs



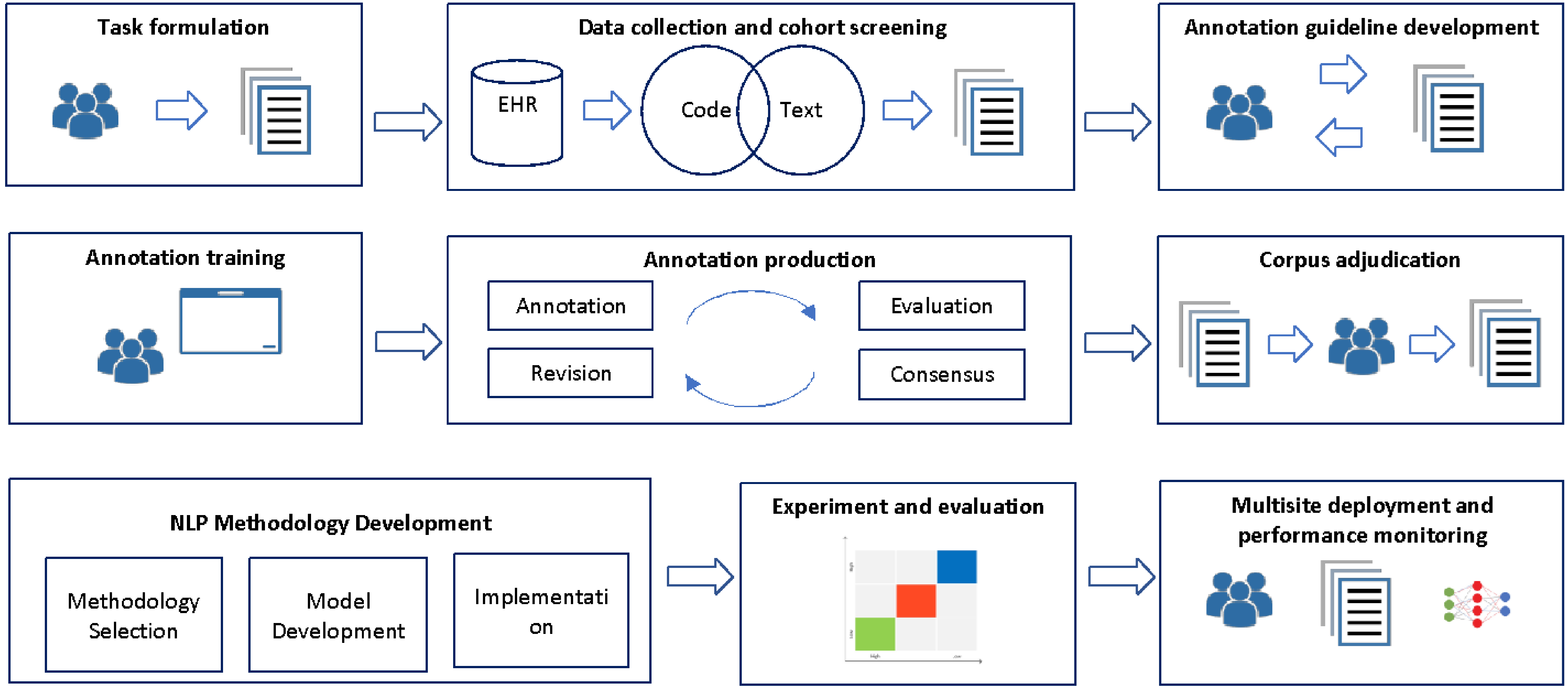
https://www.medrxiv.org/content/10.1101/2023.09.11.23295259v1.full-text

c/o Hongfang Liu: N3C NLP Engine …[in production]



A TRUST Process for NLP Model Development
Text Retrieval and Use towards Scientific rigor and Transparency

c/o Hongfang Liu

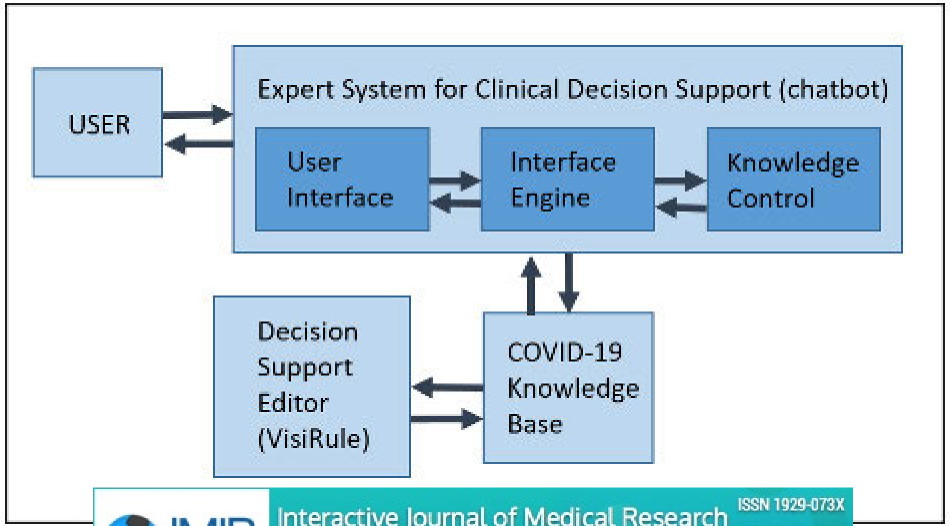


Artificial Intelligence Computational Strategies

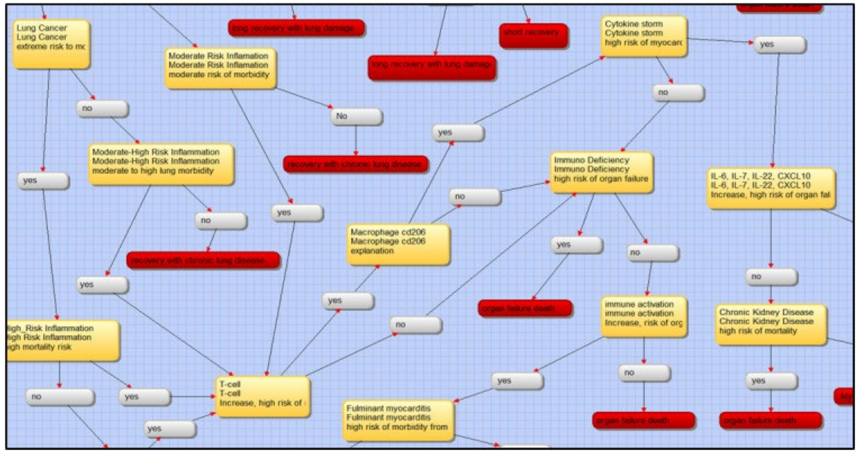
2. Expert Systems for Decision Support:

- a. Strategy: Building rule-based systems that emulate human expertise to assist healthcare professionals in decision-making.
- b. Applications: Diagnosis support, treatment planning. **Example @ right**
- c. Python Libraries: mainly 'rules engines' like [Experta](), c.2018 [PyKnow](), c.2010 [Pyke]()…

Using Decision Trees as an Expert System for Clinical Decision Support for COVID-19



JMIR Publications
Interactive Journal of Medical Research
ISSN 1929-073X
A general medical journal, focussing on innovation in health and medical research
About Search Archive Current Issue Submit Editorial Board

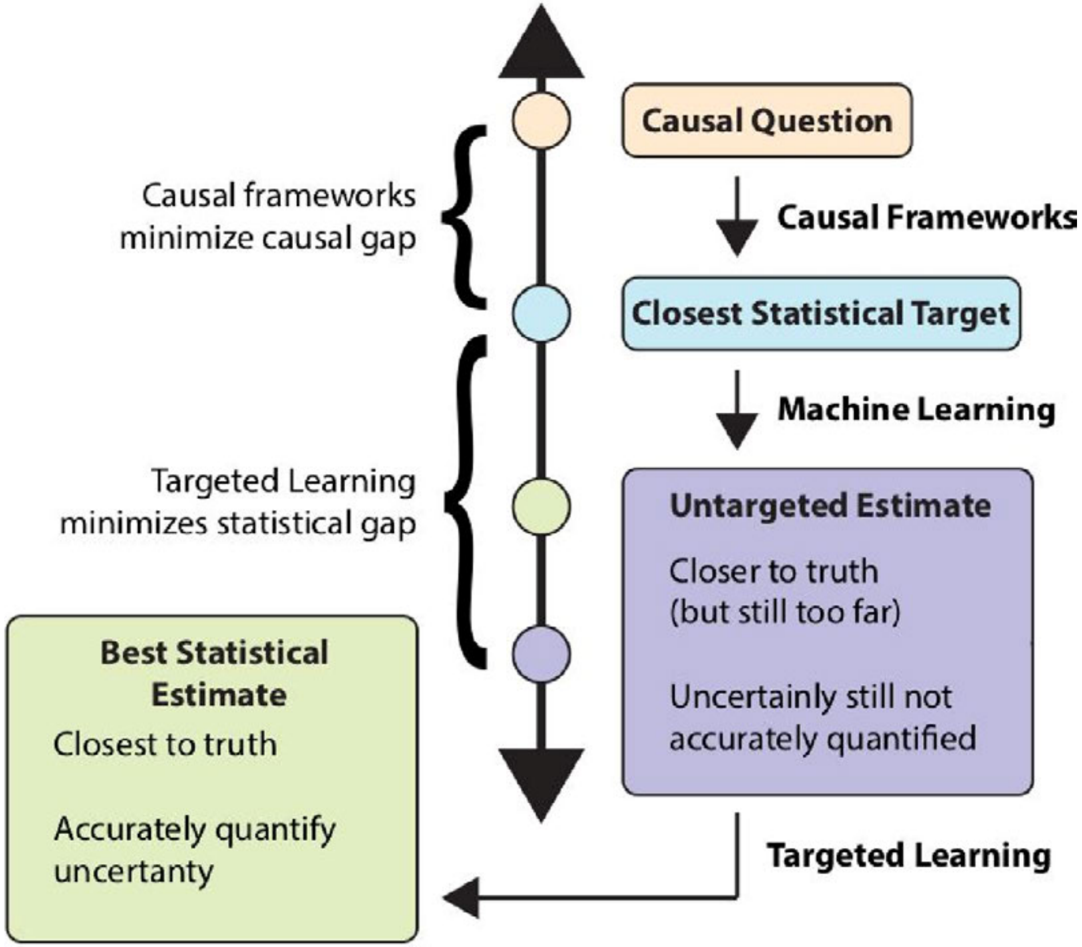


Visualized with graphical artificial intelligence software VisiRule

Artificial Intelligence Computational Strategies

3. Causal Inference Modeling using Machine Learning Algorithms:

- a. Strategy: Inferring causal relationships between variables in healthcare data to understand the impact of interventions or factors on health outcomes.
- b. Applications: Studying the effect of interventions on healthcare disparities, unclear if adequate portion of Big Tech investment.
- c. Python Libraries: CausalImpact, DoWhy, CausalLib (TMLE example doc), zEpid (TMLE doc), causal-curve, mossspider.



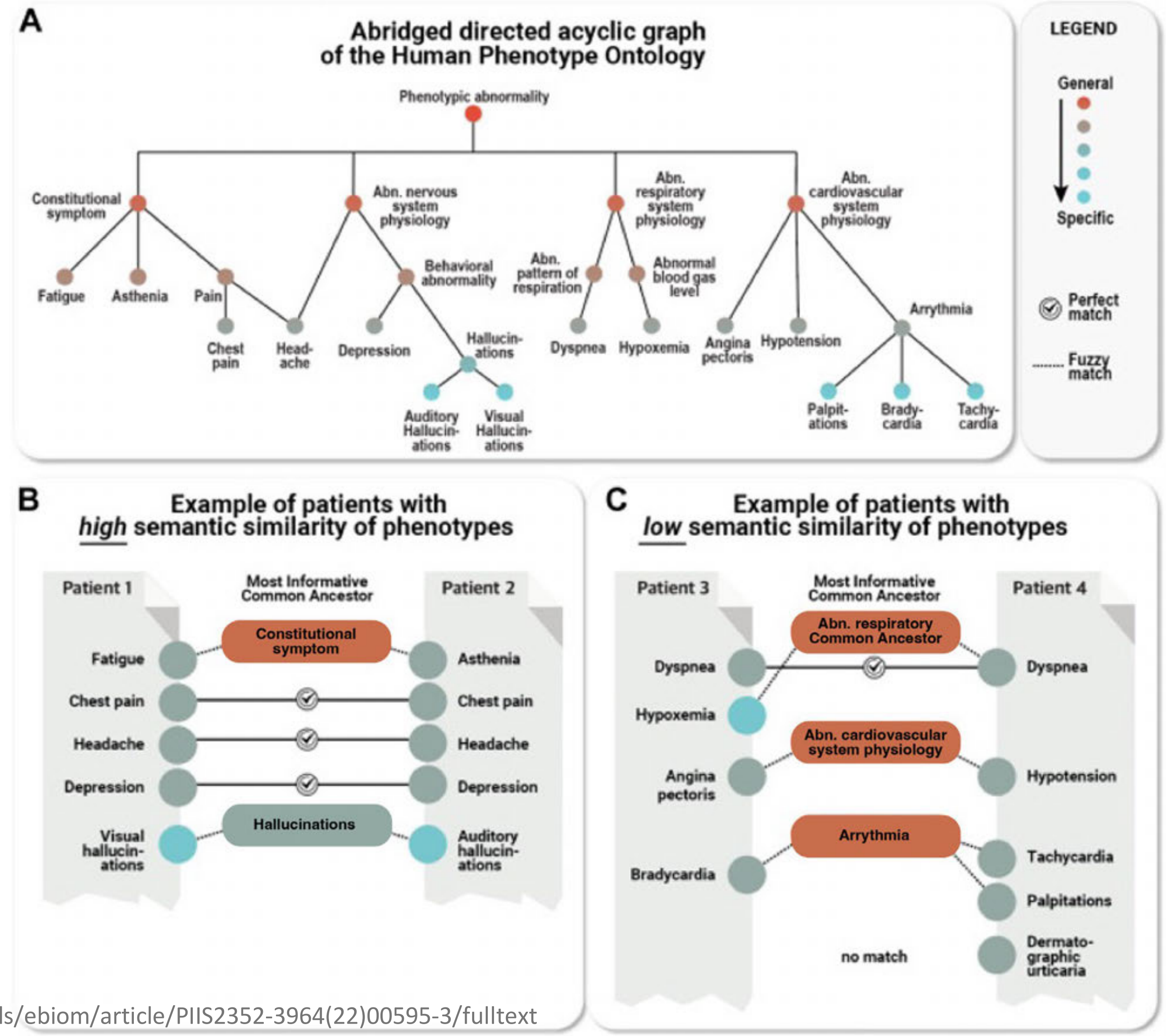
zEpid

Coyle, Jeremy R., Nima S. Hejazi, Ivana Malenica, Rachael V. Phillips, Benjamin F. Arnold, Andrew N. Mertens, Jade Benjamin-Chung, Weixin Cai, Sonali Dayal, John M. Colford, Alan E. Hubbard and Mark J. van der Laan. "Targeting Learning: Robust Statistics for Reproducible Research." *arXiv: Methodology* (2020): n. pag.

Artificial Intelligence Computational Strategies

4. Ontology and Knowledge Graphs:

- a. Strategy: Organizing and representing medical knowledge in structured formats to facilitate semantic understanding.
- b. Applications: Linking disparate healthcare data sources, enhancing interoperability.
- c. Python Libraries: [RDFlib](), [Owlready2](), [OntoGPT]()

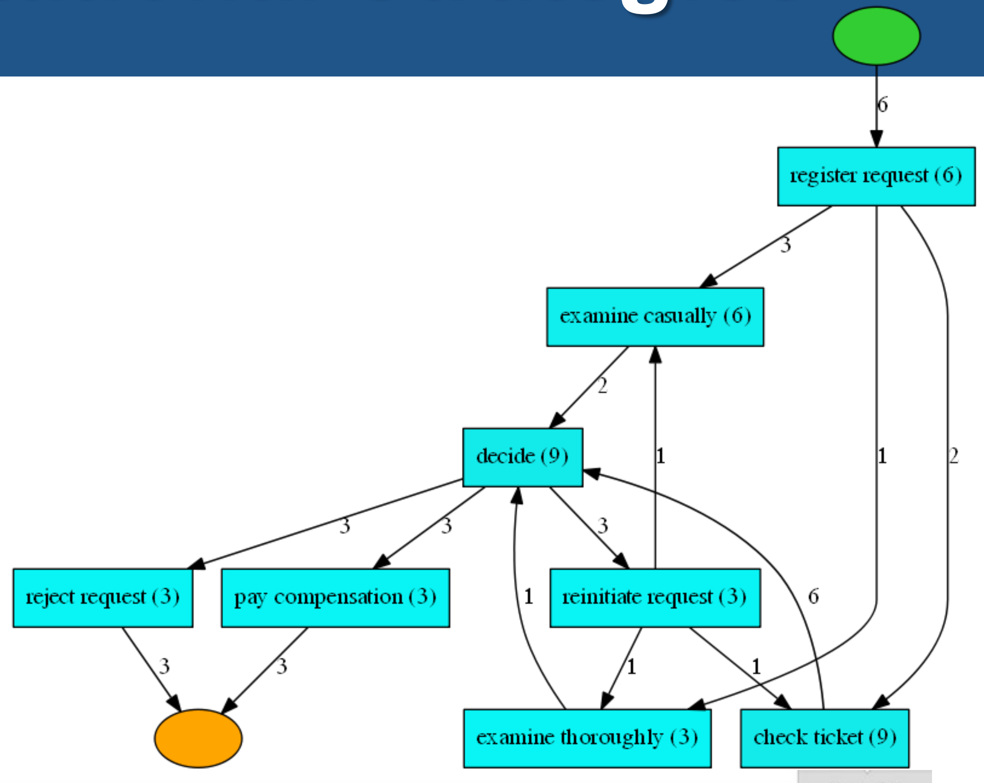


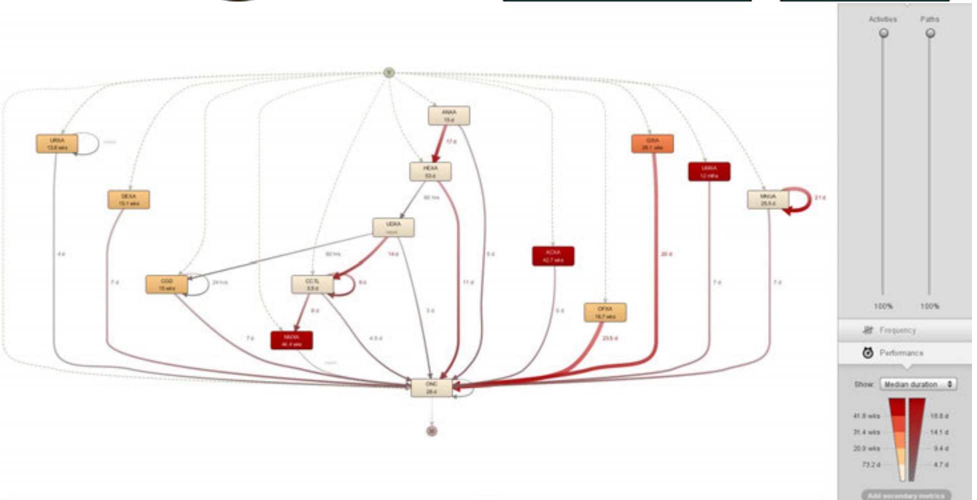
https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(22)00595-3/fulltext

Artificial Intelligence Computational Strategies

5. Process Mining:

- a. Strategy: Analyzing healthcare processes to understand workflow, identify bottlenecks, and optimize resource allocation.
- b. Applications: Improving efficiency in healthcare delivery.
- c. Python Libraries: [pm4py](), [ProM]().

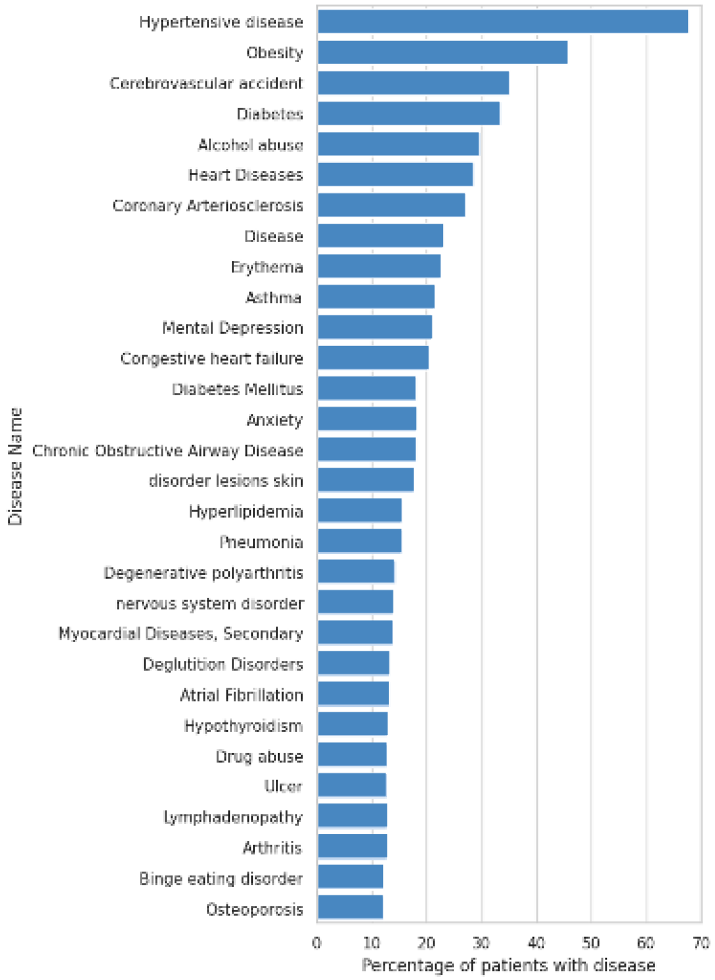




Artificial Intelligence Computational Strategies

6. Automated Coding and Classification:

- a. Strategy: Developing systems that automate the coding and classification of medical records for standardized reporting and analysis.
- b. Applications: Streamlining data coding processes, ensuring consistency.
- c. Python Libraries: [MedCAT](), [PyCaret]().



https://colab.research.google.com/github/CogStack/MedCATtutorials/blob/main/notebooks/introductory/Part_3_2_Extracting_Diseases_from_Electronic_Health_Records.ipynb#scrollTo=TupbSS6OVfgM

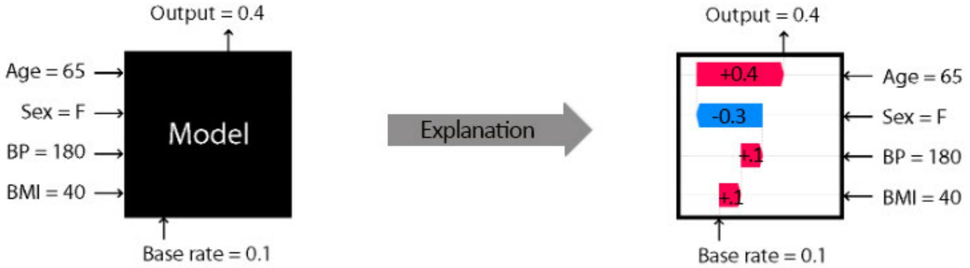
Artificial Intelligence Computational Strategies

7. Decision Support Systems with ***Explainability***:

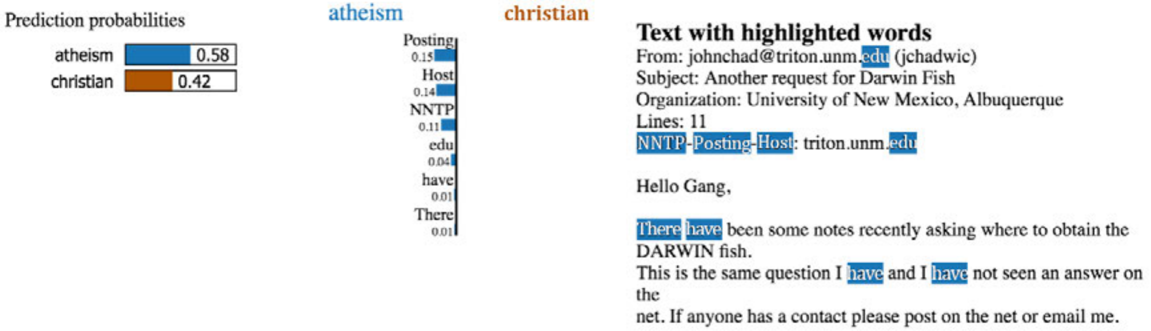
- a. Strategy: Creating AI systems that not only provide recommendations but also explain the reasoning behind the suggestions.
- b. Applications: Enhancing transparency and trust in decision support. **examples**
- c. Python Libraries: SHAP, Lime (Local Interpretable Model-Agnostic Explanations).

Examples quantify and visually show how specific features 'weigh in' on results…

SHAP



Local Interpretable Model-Agnostic Explanations



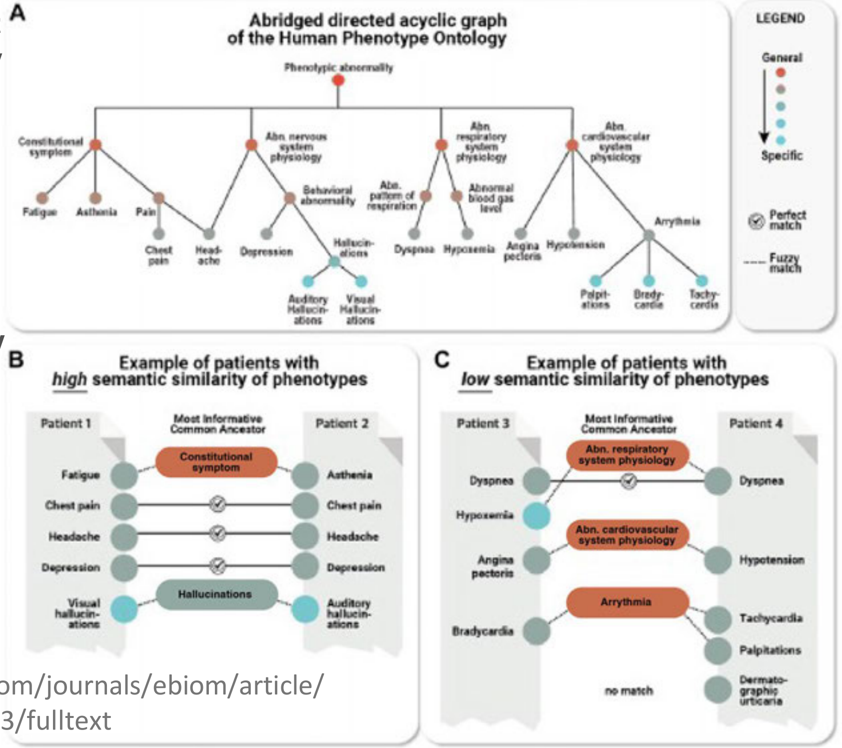
Artificial Intelligence Computational Strategies

Example: semantic-similarity-elicited long COVID types

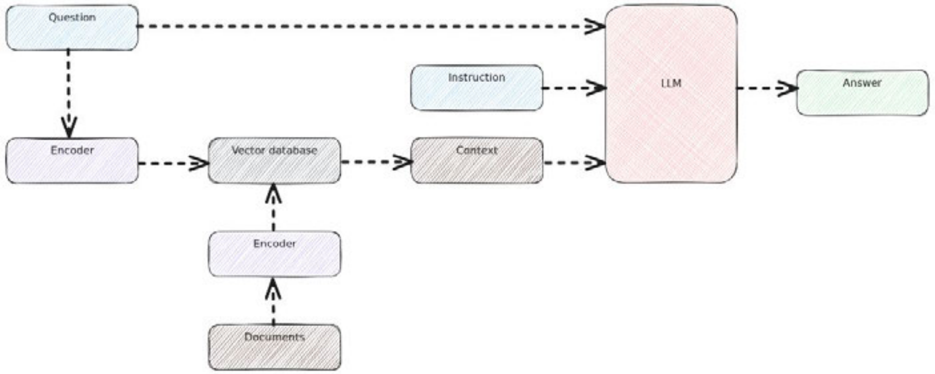
8. Semantic Analysis for Data Integration:

- a. Strategy: Applying semantic techniques to integrate heterogeneous healthcare data from various sources.
- b. Applications: Facilitating cross-domain data integration, enhancing data interoperability.
- c. Python Libraries: [RDFlib](), [Owlready2](), [OntoGPT]()
- d. other examples recently emerging:
  - OntoGPT-related [SPIRES]() - Semantic similarity
  - [Retrieval Augmented Generation]()
    - within Large Language Model Prompts (diagram @ right)

Calculating patient semantic similarity based on HPO phenotypes. A) HPO terms are arranged in a directed acyclic graph with specific terms -- excerpt of the entire ontology (15,247 terms) is shown. B) Example showing a pair of patients with relatively high phenotypic similarity



https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(22)00595-3/fulltext



Paper on race-bias in this space is, unfortunately, behind a paywall (even for NIH): https://ieeexplore.ieee.org/document/9669617

Artificial Intelligence Computational Strategies

Strategies Employed in Use Cases

National Institute of Diabetes and Digestive and Kidney Diseases

AI Computational Strategies

- We now engage participants to check our mutual understanding.

Artificial Intelligence: Fundamental Algorithms

Note: we provide link to asynchronous hands-on after ML portion…

We now quickly outline remaining number of algorithms primarily in use within AI/ML:

[*from 14 popular AI algorithms and their uses post*]

Popular deep learning algorithms

There are a number of very successful and widely adopted deep learning paradigms, the most recent being the transformer architecture behind today's generative AI models.

10 Convolutional neural networks

Convolutional neural networks (CNNs) are a type of deep neural network often used for machine vision. They have the desirable property of being position-independent. The understandable summary of a convolution layer when applied to images is that it slides over the image spatially, computing dot products; each unit in the layer shares one set of weights. A *convnet* typically uses multiple convolution layers, interspersed with activation functions. CNNs can also have pooling and fully connected layers, although there is a trend toward getting rid of these types of layers.

11 Recurrent neural networks

While convolutional neural networks do a good job of analyzing images, they don't really have a mechanism that accounts for time series and sequences, as they are strictly feed-forward networks. Recurrent neural networks (RNNs), another kind of deep neural network, explicitly include feedback loops, which effectively gives them some memory and dynamic temporal behavior and allows them to handle sequences, such as speech. That doesn't mean that CNNs are useless for natural language processing; it does mean that RNNs can model time-based information that escapes CNNs. And it doesn't mean that RNNs can *only* process sequences. RNNs and their derivatives have a variety of application areas, including language translation, speech recognition and synthesis, robot control, time series prediction and anomaly detection, and handwriting recognition. While in theory an ordinary RNN can carry information over an indefinite number of steps, in practice it generally can't go many steps without losing the context. One of the causes of the problem is that the gradient of the network tends to vanish over many steps, which interferes with the ability of a gradient-based optimizer such as stochastic gradient descent (SGD) to converge.

Artificial Intelligence: Fundamental Algorithms

- *Note: we are including these passages only to expose you to terms…*

12 Long short-term memory Long short-term memory networks (LSTMs) were explicitly designed to avoid the vanishing gradient problem and allow for long-term dependencies. The design of an LSTM adds some complexity compared to the cell design of an RNN, but works much better for long sequences. In LSTMs, the network is capable of forgetting (gating) previous information as well as remembering it, in both cases by altering weights. This effectively gives an LSTM both long-term and short-term memory, and solves the vanishing gradient problem. LSTMs can deal with sequences of hundreds of past inputs.

13 Transformers Transformers are neural networks that solely use *attention* mechanisms, dispensing with recurrence and convolutions entirely. Transformers were invented at Google. Attention units (and transformers) are part of Google's BERT (Bidirectional Encoder Representations from Transformers) algorithm and OpenAI's GPT-2 algorithm (transformer model with unsupervised pre-training) for natural language processing. Transformers continue to be integral to the neural architecture of the latest large language models, such as ChatGPT/Bing Chat (based on GPT-3.5 or GPT-4) and Bard (based on LaMDA, which stands for Language Model for Dialogue Applications). Attention units are not terribly sensitive to how close two words in a sentence appear, unlike RNNs; that makes them good at tasks that RNNs don't do well, such as identifying antecedents of pronouns that may be separated from the referent pronouns by several sentences. Attention units are good at looking at a context larger than just the last few words preceding the current word.

14 Q-learning Q-learning is a model-free, value-based, off-policy algorithm for reinforcement learning that will find the best series of actions based on the current state. The "Q" stands for quality. Quality represents how valuable the action is in maximizing future rewards. Q-learning is essentially learning by experience. Q-learning is often combined with deep neural networks. It's used with convolutional neural networks trained to extract features from video frames, for example for teaching a computer to play video games or for learning robotic control. AlphaGo and AlphaZero are famous successful game-playing programs from Google DeepMind that were trained with reinforcement learning combined with deep neural networks. As we've seen, there are many kinds of machine learning problems, and many algorithms for each kind of problem. These range in complexity from linear regression for numeric prediction to convolutional neural networks for image processing, transformer-based models for generative AI, and reinforcement learning for game-playing and robotics.

Artificial Intelligence: Pros & Cons

Per Think-a-thon Planning outline:

- *Strengths:*
  - *Flexible to multiple data modalities and – with ENOUGH data – quite robust,*
  - *Some aspects are ‘explainable’ through additional ‘extra’ steps*
- *Weaknesses:*
  - *NOT interpretable,*
  - *assumption-dense, yet assumptions typically NOT transparently assessed*
  - *often very dependent upon the tacit decisions made by those applying AI*

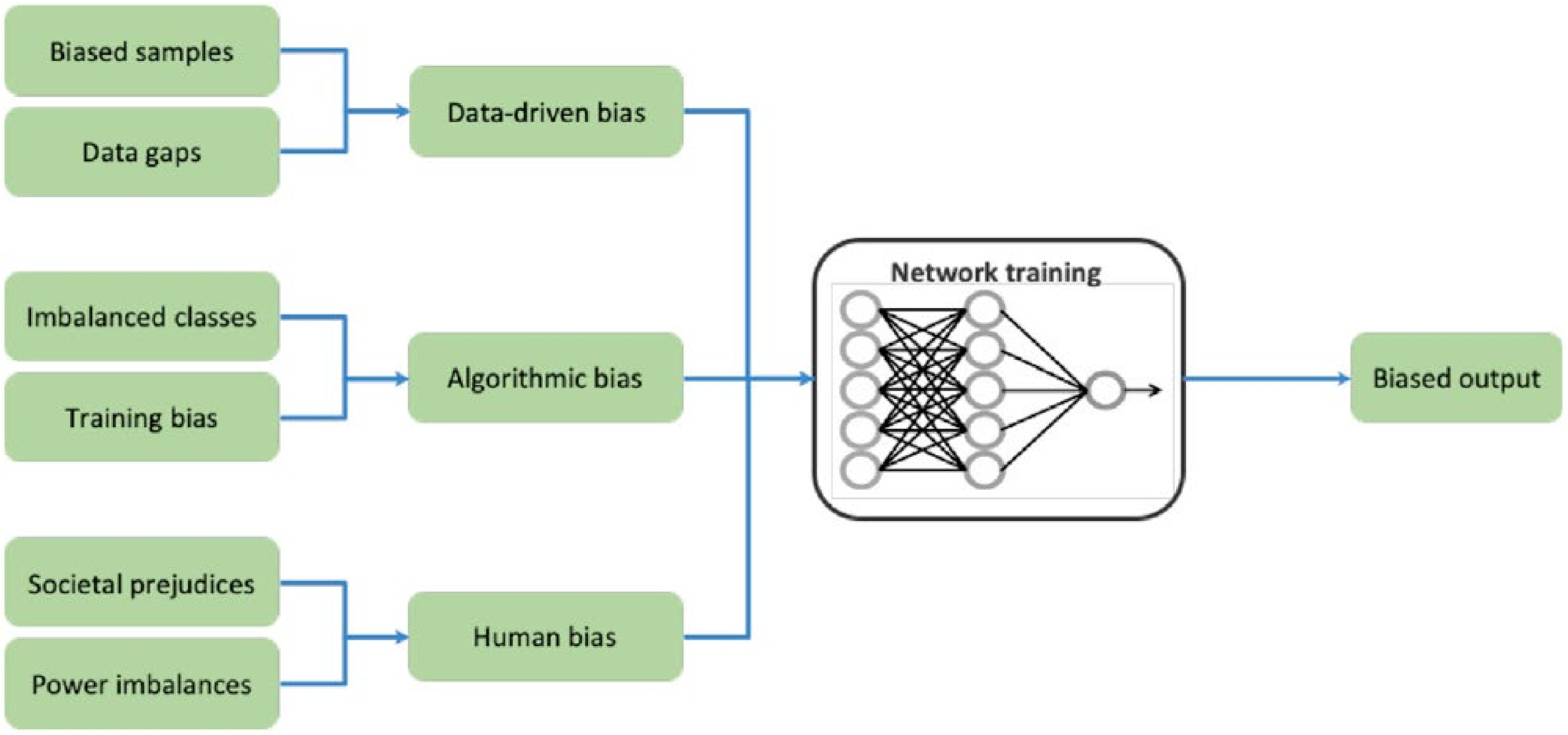
ScHARe

AI Bias

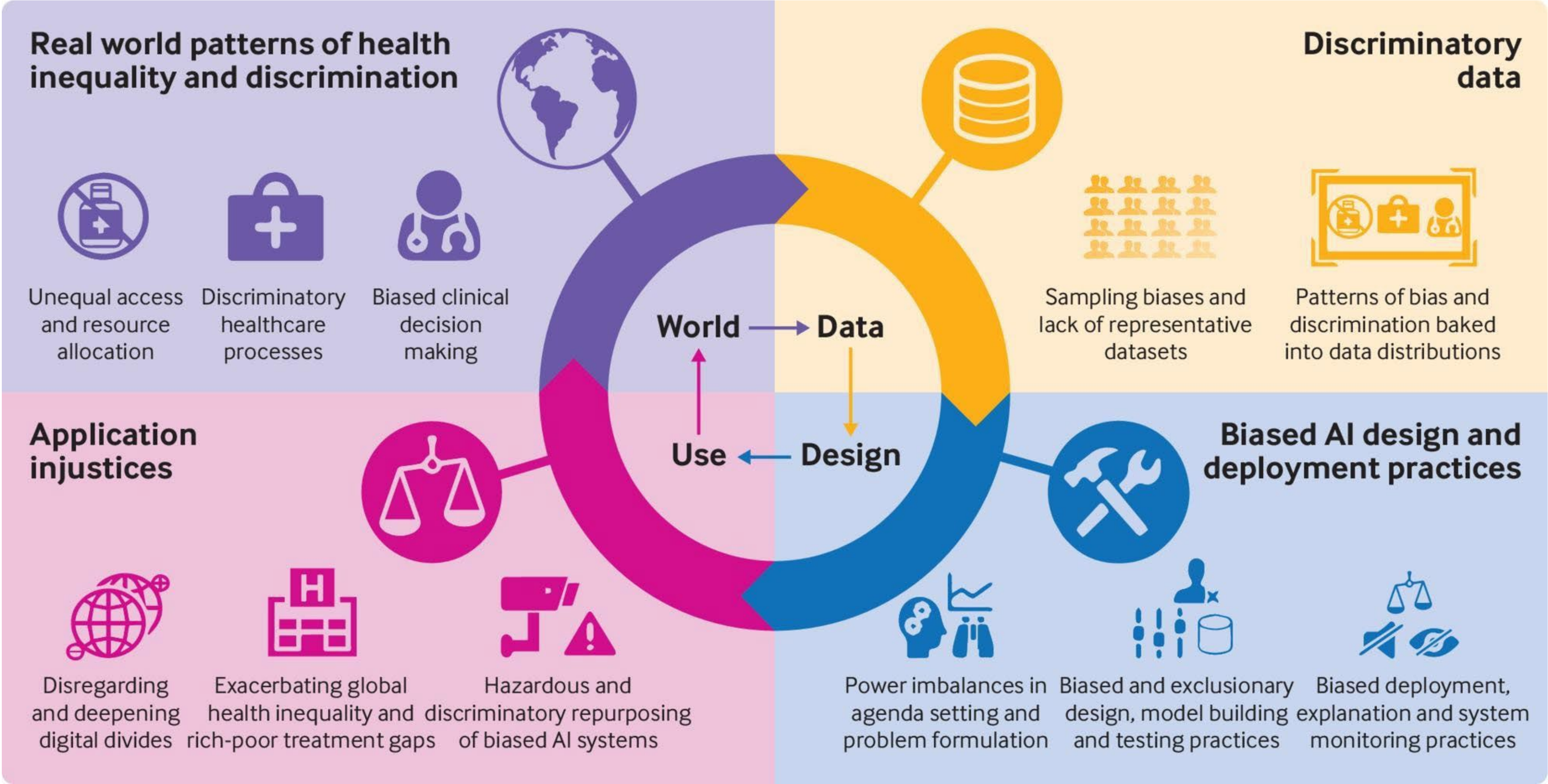
AI bias

- Algorithms are using Big Data to **influence decisions affecting people's health.**
- **Training data** that specifies what the correct outputs are for some people/objects **is used to learn a model** which is then applied to other people/objects to make predictions about the correct outputs for them
- Algorithms run the **risk of replicating and amplifying human biases** affecting protected groups, leading to outcomes systematically less favorable to them
- **Bias can originate from unrepresentative/incomplete training data** that reflects historical inequalities, or manifest at various points in the algorithm development process

Algorithmic racial bias mechanisms



The big picture



Example 1: Algorithm favors healthier white patients over sicker black patients

The issue

An algorithm used to predict which patients would benefit from extra medical care **flagged healthier white patients as more at risk than sicker black patients**

- An analysis on 3.7 million patients found that **black patients ranked as equally as in need of extra care** as white patients collectively suffered from 48,772 additional chronic diseases
- The bias was discovered when researchers from a health system in Massachusetts found the **highest scores in their patient population concentrated in the most affluent suburbs of Boston**

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453. doi:10.1126/science.aax2342

Example 1: Algorithm favors healthier white patients over sicker black patients

The cause

- **The algorithm used a seemingly race-blind metric**: how much patients would cost the health-care system in the future

- **Cost isn’t a race-neutral measure of health-care need**: unequal access to care means that we spend less money caring for black patients than for white patients

The solution

- **Researchers tweaked the algorithm** to make predictions about their future health conditions

- The tweak increased the percentage of black patients receiving additional help from 17.7 to 46.5%

Example 2: Flawed racial adjustments in kidney function estimates

- **Race forms part of the algorithms used to assess kidney function through an eGFR equation** that uses serum creatinine measurement, age, sex, race, body weight

- The inclusion of a **coefficient for black patients** in the eGFR equation was based on small poor-quality studies. The more accurate **CKD-EPI equation** still contains a correction for black patients.

The issue

The CKD-EPI equation modifier **increases eGFR for black individuals by nearly 16%**, altering guideline-based diagnoses and referrals for care

Diao JA, Wu GJ, Taylor HA, et al. Clinical Implications of Removing Race From Estimates of Kidney Function. JAMA. 2021;325(2):184-186. doi:10.1001/jama.2020.22124

Example 2: Flawed racial adjustments in kidney function estimates

The cause

Including adjustment for race in these eGFR equations **ignores the substantial diversity within self-identified black patients and other racial or ethnic minority groups**.

The solution

- Healthcare organizations have started **removing the race-based adjustment from the eGFR equation**, reporting the "White/Other" value for all patients.
- This measure may **increase CKD diagnoses among black adults** and enhance access to specialist care, medical nutrition therapy, kidney disease education, and kidney transplantation.

Example 3: AI-driven dermatology leaves dark-skinned patients behind

- Machine Learning has been used to create **programs capable of distinguishing between images of benign and malignant moles** with accuracy similar to that of board-certified dermatologists.

- However, the algorithms used by most healthcare organizations are basing most of their knowledge on ISIC, an open-source repository of **skin images from primarily fair-skinned populations.**

Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatol. 2018;154(11):1247. doi:10.1001/jamadermatol.2018.2348

The issue

Lesions on patients of color are less likely to be diagnosed. The algorithms provide advancement for the Caucasian population, which already has the highest survival rate.

Example 3: AI-driven dermatology leaves dark-skinned patients behind

The cause

Bias emanates from unrepresentative training data that reflects historical inequalities: decades of clinical research have focused primarily on people with light skin.

The solution

- Researchers are taking measures to ensure a **more equitable demographic participation in clinical trials.**
- ISIC is looking to **expand its archive to include as many skin types as possible,** and has asked dermatologists to contribute photos of lesions on their patients with darker skin.

Testing for biases in datasets and algorithms

- Testing for biases in datasets and algorithmic models is **crucial for ensuring fairness and reliability** in data science.

- Here are general strategies and **techniques for testing biases**, categorized into datasets and algorithmic models.

Testing for biases in datasets

1. **Exploratory Data Analysis (EDA):**

    - **Explanation:** EDA involves visualizing and summarizing the main characteristics of the dataset using histograms, box plots, and summary statistics. The goal is to understand the data distribution
    - **Importance:** EDA helps identify outliers, imbalances, and biases
    - **Example:** If EDA reveals a dataset on job applicants is heavily skewed towards a specific gender, it might indicate a bias in the sampling process
    - **Python Libraries:** Pandas, Matplotlib, Seaborn

Testing for biases in datasets

2. **Demographic Analysis (DA):**

   - **Explanation:** Break down the dataset based on demographic attributes (e.g., age, gender, ethnicity) and analyze the distribution within each group
   - **Importance:** DA can identify imbalances/over-representations in specific groups
   - **Example:** In a healthcare dataset, if one demographic group is over-represented, it may lead to biased predictions
   - **Python Libraries:** Pandas, Matplotlib, Seaborn

Testing for biases in datasets

3. **Data Stratification:**

   - **Explanation:** Divide the dataset into subgroups based on relevant features and analyze each subgroup independently
   - **Importance:** This helps detect biases that may exist disproportionately in specific subgroups
   - **Example:** In a credit scoring dataset, stratifying by income levels can reveal biases in credit approval rates
   - **Python Libraries:** Pandas

Testing for biases in datasets

4. **Bias Detection Tools:**

   - **Explanation:** Use tools like IBM's AI Fairness 360 or Google's What-If Tool that offer automated metrics for assessing bias in datasets and models
   - **Importance:** Automated tools efficiently identify subtle biases and provide quantitative measures, facilitating a systematic approach to bias detection
   - **Examples:**
     - AI Fairness 360 provides a set of algorithms to evaluate fairness across various demographic groups
     - Google's What-If Tool allows interactive exploration of model predictions and visualization of outcomes across different subsets of data
   - **Tools:** AI Fairness 360, What-If Tool

Fixing biases in datasets

Several techniques can be employed to address bias in datasets:

- **Oversampling** involves increasing the representation of underrepresented groups in the dataset, ensuring a more balanced distribution
- **Undersampling** reduces overrepresented groups
- **Using synthetic data** generation introduces artificially generated data points to mitigate imbalances
- **Reweighting** or adjusting the importance of specific instances during model training helps address bias
- Regularly **updating and expanding datasets** with diverse, representative samples further contribute to minimizing bias

Testing for biases in algorithms

1. **Performance Metrics Disaggregation:**

   - **Explanation:** Evaluate model performance metrics (e.g., accuracy, precision) separately for different subgroups defined by sensitive attributes
   - **Importance:** Disparities in performance metrics across groups may indicate bias
   - **Example:** Testing a healthcare algorithm disaggregating accuracy by racial groups reveals slightly lower accuracy for Black patients. Fixes: root cause analysis and algorithm adjustments
   - **Python Libraries:** Scikit-learn

Testing for biases in algorithms

2. **Confusion Matrix Analysis:**

   - **Explanation:** Analyze the confusion matrix (a table that summarizes the performance of a classification algorithm by comparing predicted and actual values) for different subgroups to identify disparities in model predictions, particularly for false positives and false negatives
   - **Importance:** Disparities in errors can pinpoint areas where bias may exist
   - **Example:** Analyzing a medical diagnosis algorithm using a confusion matrix to evaluate the model's effectiveness in making medical diagnoses. Differences in false positives between genders might indicate bias. Fix: adjusting decision thresholds, retraining with balanced data, consulting domain experts
   - **Python Libraries:** Scikit-learn

Testing for biases in algorithms

3. **Fairness Indicators:**

   - **Explanation:** Integrate fairness indicators (measures that assess whether a model's predictions treat different groups equitably) into the model evaluation process to identify bias
   - **Importance:** Fairness indicators provide a structured approach to measure bias
   - **Example:** Using Google's TensorFlow Fairness Indicators to compare prediction accuracies of a healthcare decision support algorithm across different racial groups. Fixes: retraining the algorithm with balanced data, adjusting decision thresholds
   - **Python Libraries:** TensorFlow Fairness Indicators

Testing for biases in algorithms

4. **Sensitivity Analysis:**

   - **Explanation:** Assess how changes in input features impact model predictions. This involves tweaking one feature at a time and observing the model's response
   - **Importance:** It helps identify features that disproportionately influence the model, potentially leading to biases
   - **Example:** In a healthcare decision support algorithm predicting diabetes risk, assessing how variations in input variables (e.g., age, BMI) impact predictions for different racial groups. The analysis reveals that the algorithm disproportionately relies on a single variable affecting certain groups. Fixes: recalibrating the model to minimize the influence of that variable, retraining with a more diverse dataset
   - **Python Libraries:** Scikit-learn

Testing for biases in algorithms

5. **Counterfactual Analysis:**

   - **Explanation:** Counterfactual analysis involves exploring hypothetical scenarios by determining the minimal changes needed in input features to alter a model's prediction
   - **Importance:** It helps understand the model's decision boundaries and can highlight biases
   - **Example:** In a credit approval algorithm, if a loan application from a certain racial group is denied, the analysis involves identifying the minimal changes needed in the application features (income, credit score) for approval, shedding light on potential biases. Fixes: adjusting the decision thresholds, mitigating the impact of sensitive features, or retraining the model
   - **Python Libraries:** Alibi Counterfactual