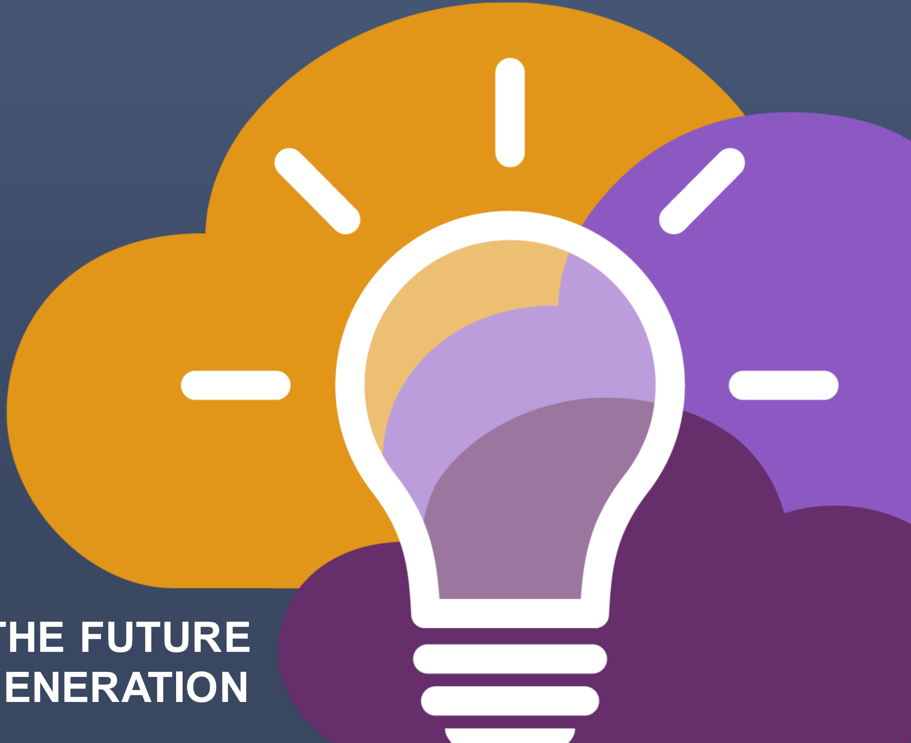
ScHARe

Making data AI-ready



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Working with Big Data

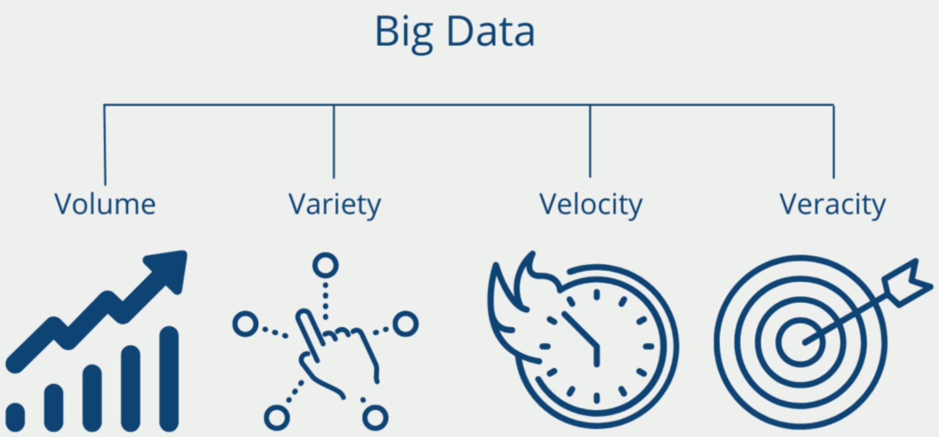
Extremely large datasets that are statistically analyzed to gain detailed insights, often **using AI** and substantial **computer-processing power**

Datasets can be **linked together (data integration)** to provide a comprehensive perspective for research knowledge generation (this includes data from RO1s, U54s, PARs, KO1s, etc.)

Data integrity (data quality) is the overarching completeness, accuracy, consistency, accessibility, and security of the data for its intended purpose.
This should always be assessed before using a dataset

FAIR data are data which meet machine-actionability principles of:

- Findability
- Accessibility
- Interoperability
- Reusability



Big Data is characterized by the 4 V's:

1. **Volume**: Enormous amounts of data
2. **Variety**: Diverse data types and structures
3. **Velocity**: High-speed data generation
4. **Veracity**: Challenges in ensuring data accuracy and reliability

Big data is difficult to process using traditional methods

Big Data: structured and unstructured data

Structured data is quantitative data that is organized and easily searchable

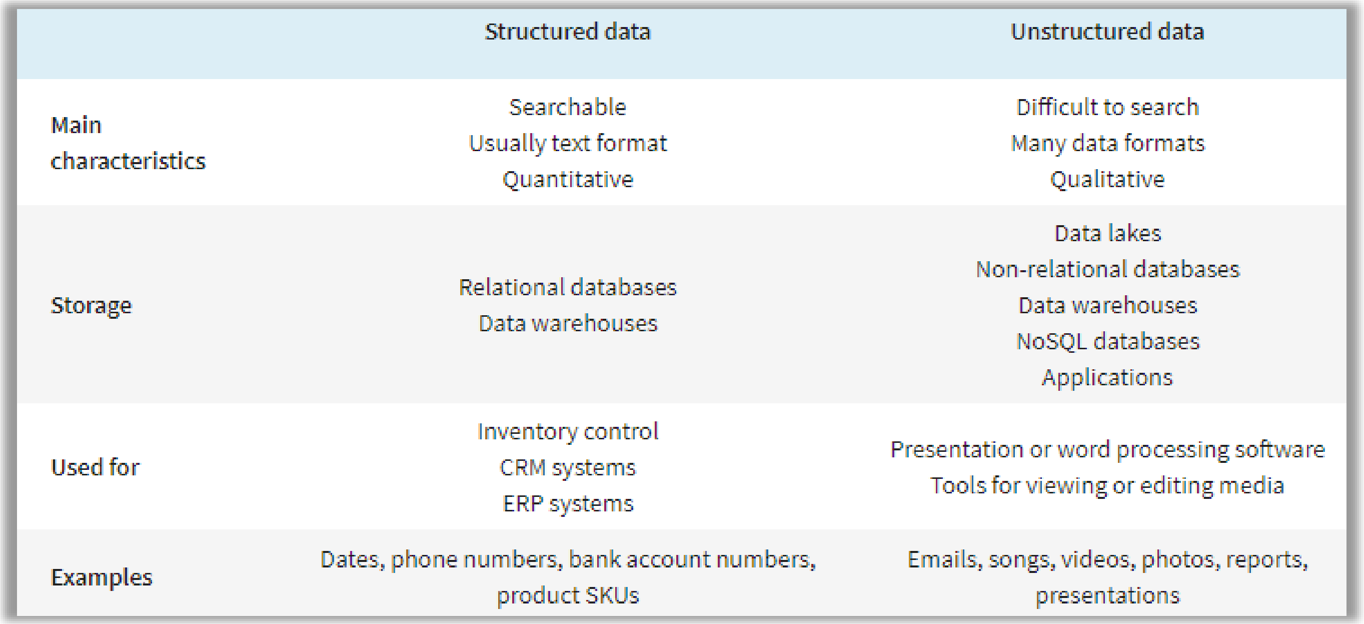
Some tools used to work with structured data include:

- OLAP
- MySQL
- PostgreSQL
- Oracle Database

Unstructured data is every other type of data that is not structured.

Some tools used to manage unstructured data include:

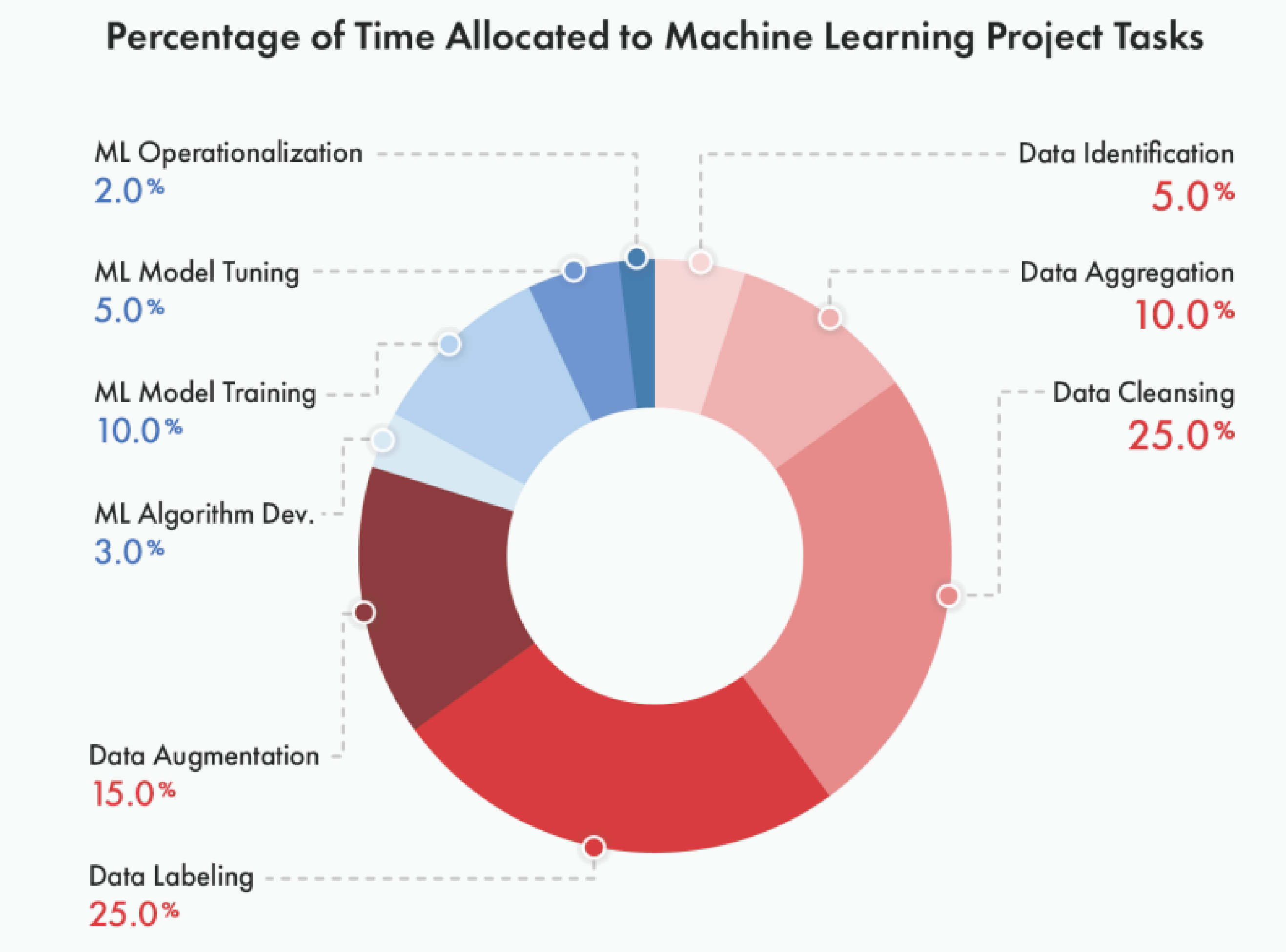
- MongoDB
- Hadoop
- Azure



	Structured data	Unstructured data
Main characteristics	Searchable Usually text format Quantitative	Difficult to search Many data formats Qualitative
Storage	Relational databases Data warehouses	Data lakes Non-relational databases Data warehouses NoSQL databases Applications
Used for	Inventory control CRM systems ERP systems	Presentation or word processing software Tools for viewing or editing media
Examples	Dates, phone numbers, bank account numbers, product SKUs	Emails, songs, videos, photos, reports, presentations

Big data is difficult to process using traditional methods

Preparing for AI



Clean data leads to more accurate, reliable, and effective AI models. Cleaning data is crucial for AI use because:

- **Accuracy:** Ensures the data is correct, improving the reliability of AI models.
- **Consistency:** Eliminates discrepancies, making the data uniform and easier to analyze.
- **Performance:** Reduces noise and irrelevant information, enhancing model efficiency.
- **Trustworthiness:** Increases confidence in the results produced by AI systems.
- **Compliance:** Helps in adhering to data quality standards and regulations.
- **Bias reduction:** Minimizes biases, leading to fairer outcomes.

Quality data can also be **aggregated with other quality data** for AI use

Good data starts with your dataset

Making datasets AI-ready: a multifaceted approach

Making datasets AI-ready involves ensuring they are suitable for use in machine learning and artificial intelligence applications.

Key aspects of AI-ready datasets:

- **Data quality:** Ensure data accuracy, completeness, and consistency. Address missing values, outliers, and inconsistencies that could impact model performance.
- **Data cleaning and pre-processing:** Apply techniques like normalization, scaling, and encoding to prepare the data for machine learning algorithms.
- **Feature engineering:** Create new features from existing data or transform existing features to improve model performance.
- **Documentation:** Provide clear and detailed documentation about the dataset, including variable definitions, data collection methods, and any transformations applied.

Why quality checks are essential for AI-ready data

Datasets are the lifeblood of AI models. Their quality directly impacts the performance, fairness, and reliability of the resulting models.

Poor quality data can lead to:

- **Biased models:** Unrepresentative or skewed data can lead to models that perpetuate existing biases and produce discriminatory outcomes.
- **Inaccurate results:** Inconsistent or erroneous data can cause models to learn incorrect patterns and generate unreliable predictions.
- **Wasted resources:** Training models on low-quality data is a waste of time, computational power, and financial resources.

Overview of quality checks

Quality checks for AI-ready datasets encompass various aspects, categorized into these key areas:

1. **Data completeness:**
   1. **Missing values:** Identifying and handling missing data points through imputation or removal.
   2. **Outliers:** Detecting and addressing unusual data points that might skew model training.

2. **Data consistency:**
   1. **Formatting:** Ensuring consistent data formats across the entire dataset.
   2. **Units and labels:** Maintaining consistency in units of measurement and data labeling.

3. **Data accuracy:**
   1. **Verification:** Cross-checking data with reliable sources to identify and correct errors.
   2. **Validation:** Comparing data against expected values or domain knowledge to ensure accuracy.

Overview of quality checks

4. **Data representativeness:**
    1. **Bias:** Analyzing the data for potential biases in sampling, labeling, or other aspects.
    2. **Generalizability:** Assessing whether the data adequately represents the target population for the intended AI application.

5. **Data documentation:**
    1. **Metadata:** Providing comprehensive information about the data, including its origin, collection methods, and usage guidelines.
    2. **Version control:** Maintaining clear versioning of the data to track changes and ensure consistency.

Poll

What is the primary purpose of verifying data against reliable sources?

a) To identify missing values

b) To ensure data accuracy

c) To check for outliers

d) To maintain data consistency

Checklist for quality checks

Data completeness:

Check for missing values and implement appropriate handling strategies.Identify and address outliers.

Data consistency:

Ensure consistent formatting throughout the dataset.

Verify consistency in units and labels.

Data accuracy:

Perform data verification against reliable sources.

Validate data against expected values or domain knowledge.

Data representativeness:

Analyze the data for potential biases.

Assess the generalizability of the data to the target population.

Data documentation:

Create comprehensive metadata describing the data.

Implement version control mechanisms.

Importance of completeness and data dictionaries for AI-ready datasets

Two critical aspects of ensuring datasets are AI-ready are completeness and data dictionaries. Let's explore why each is crucial:

1. Completeness:

A complete dataset refers to one with minimal missing values or outliers that could significantly impact the training and performance of AI models. Missing data can lead to:

- **Biased models:** if specific data points are consistently missing, the model might learn skewed patterns and produce unfair results.
- **Inaccurate predictions:** missing data can hinder the model's ability to capture the full picture and lead to unreliable outputs.
- **Inefficient training:** training models on incomplete data can be computationally expensive and inefficient, yielding suboptimal results.

Importance of completeness and data dictionaries for AI-ready datasets

2. Data dictionaries:

Data dictionaries act as the instruction manuals for your dataset, providing crucial information about each variable. They define:

- **Variable names:** clear and consistent names that facilitate understanding and avoid confusion.
- **Data types:** specifying the format of data (e.g., Numerical, categorical, text) ensures proper interpretation by the model.
- **Descriptions:** explanations of the meaning and potential values of each variable, promoting clarity and reducing ambiguity.
- **Units of measurement:** standardizing units (e.g., Meters, kilometers) ensures consistent interpretation and analysis.

Addressing missing data: strategies for imputation

- Missing data is a common challenge in datasets, and how you handle it can significantly impact your research findings.
- Strategies for handling missing data:

  - **Deletion:** Remove rows or columns with a high percentage of missing values, but this can lead to information loss.
  - **Mean/median imputation:** Replace missing values with the mean or median of the respective variable.
  - **Model-based imputation:** Use statistical models to predict missing values based on other variables in the dataset.

Understanding and addressing missing data

Data Missingness Strategies: Understanding and Addressing Missing Data

Missing data, where values are absent from a dataset, is a prevalent challenge in various fields. It can significantly impact the results of data analysis and machine learning models. Fortunately, various strategies exist to address missing data

Understanding Missing Data:

Before delving into strategies, it's crucial to understand the **types of missing data**:

- **Missing Completely at Random (MCAR):** Missingness occurs randomly and is unrelated to any other variables in the dataset.
- **Missing at Random (MAR):** Missingness depends on observable variables in the dataset but not on the missing values themselves.
- **Missing Not at Random (MNAR):** Missingness is related to the missing values themselves, often due to unobserved factors.

Understanding and addressing missing data

Addressing Missing Data:

Several strategies can be employed to handle missing data, depending on the nature and extent of missingness:

1. **Deletion:**

- ❑ **Listwise deletion:** Removes entire rows with missing values, potentially reducing sample size and introducing bias if MCAR doesn't hold.
- ❑ **Pairwise deletion:** Removes only the data points with missing values for the variable being analyzed, potentially wasting information.

Understanding and addressing missing data

2. Imputation:

❑ **Mean/Median/Mode imputation:** Replaces missing values with the average, median, or most frequent value of the variable, respectively. Simple but may introduce bias, especially for skewed distributions.

❑ **Hot Deck imputation:** Replaces missing values with values from existing observations with similar characteristics, reducing bias but potentially introducing noise.

❑ **Model-based imputation:** Uses statistical models like regression or machine learning to predict missing values based on other variables, potentially more accurate but computationally expensive.

Poll

What strategies do you find most effective in handling missing values and outliers in datasets?

Dealing with proxies and small sample sizes: alternative approaches

- **Not all research questions may have readily available data for every variable**. In such cases, researchers might need to employ **proxy variables** or navigate situations with small sample sizes.

- Strategies for addressing proxies and small sample sizes:

    - **Proxy variables:** Carefully select proxy variables that are demonstrably linked to the desired variable, but be mindful of potential limitations and biases.

    - **Small sample size analysis:** Utilize appropriate statistical methods designed for small datasets, such as non-parametric tests or bootstrapping techniques.

Synthetic/AI Generated DATA

- Information that is **artificially generated** rather than produced by real-world events.
- Generated to meet **specific needs or certain conditions that may not be found in the original, real data**
- Typically created using **algorithms**, synthetic data can be deployed to **validate mathematical models** and to **train machine learning models**
- Often used for **underrepresented populations** in datasets

Digital Twins

Digital model of an intended or actual real-world physical product, system, or process (a physical twin) that serves as the **effectively indistinguishable digital counterpart** of it for practical purposes, such as simulation, integration, testing, monitoring and maintenance

The digital twin of a person, based on such computer simulations, could help drug developers design, test and monitor, and aid doctors in applying, the **safest and most effective treatments or therapies** that are specific and tailored to our genetics or biochemistry.

Not the answers to poor quality or missing data

Exploring the ethical considerations of using synthetic data

While synthetic data offers certain advantages, its use raises **ethical considerations** that researchers must address responsibly:

- **Transparency and disclosure:** Clearly communicate the use of synthetic data, including the number of actual people used to generate it, and its limitations to avoid misinterpretations.

- **Responsible use:** Ensure the synthetic data is used ethically and does not perpetuate harmful stereotypes or discriminatory practices.

- **Potential biases:** Be mindful of generalizability limitations and potential biases that might be introduced during the synthetic data generation process, and take steps to mitigate them.

Poll

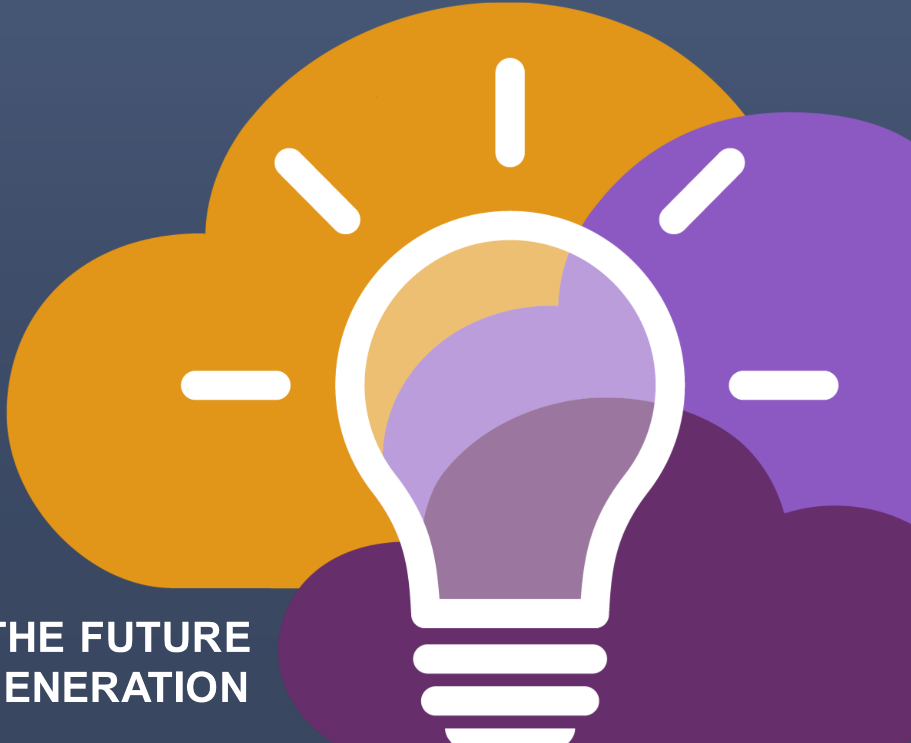
What other ethical considerations should researchers keep in mind when using synthetic data in their studies?

Poll

In your opinion, what are the biggest challenges researchers face in ensuring their datasets are truly 'AI-ready’ beyond the technical aspects?

ScHARe

Ethical and transparent Artificial Intelligence



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Ethical AI

It is crucial that **AI algorithms respect basic human values** and undertake their analysis and decision-making in a trustworthy manner.

Ethical AI builds tools that are faithful to values such as **accountability, privacy, safety, security, and transparency**.

Taken together with explainable AI, it is a way to **deploy AI in ways that further human values**.

Explainable AI (XAI)

One of the complaints about AI is the **lack of transparency** in how it operates. Many developers don’t reveal the data used or how various factors are weighted. Outsiders cannot tell how AI reached the decision that it did.

This lack of explainability can lead people to **not trust AI**.

XAI seeks to help **describe either the overall function of AI or the specific way it reaches decisions**, to make AI more understandable and trustworthy.

Artificial Intelligence Bias

Algorithms are widely used in healthcare- and policy-related decisions. However, many operate as “**black boxes**”, offering little opportunity for testing to identify biases.

Biases can result from:

- **social/cultural context not considered**
- **design limitations**
- **data missingness and quality problems**
- **algorithm development and model training**

If not identified, biased algorithms may result in decisions that lead to discrimination, unequitable healthcare, and/or health disparities.

Trust in AI

Caution Against:

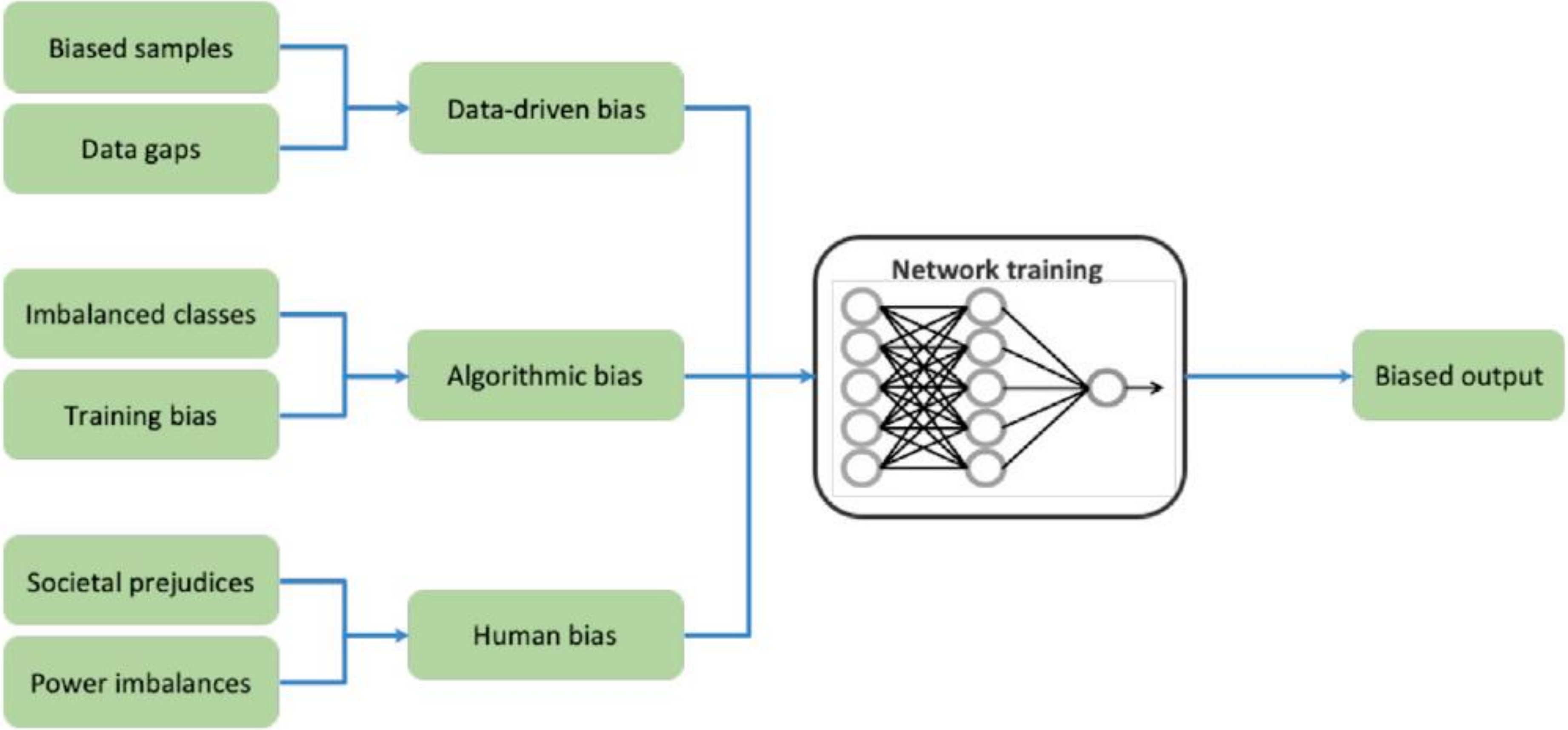
- **Epistemic trust**, which describes the willingness to accept new information from another person or entity as trustworthy, generalizable, and relevant.
- **Synthetic trust**, a misplaced belief in the model's capabilities and fairness.

Mistrust of AI

- Fear of misuses
- Fear because of harmful impacts of biases
- Lack of underrepresented populations/community trust

Algorithmic bias mechanisms

Bias can originate from unrepresentative/incomplete training data that reflects historical inequalities, or manifest at various points in the algorithm development process



Poll

Which of the following factors can contribute to bias in AI algorithms?

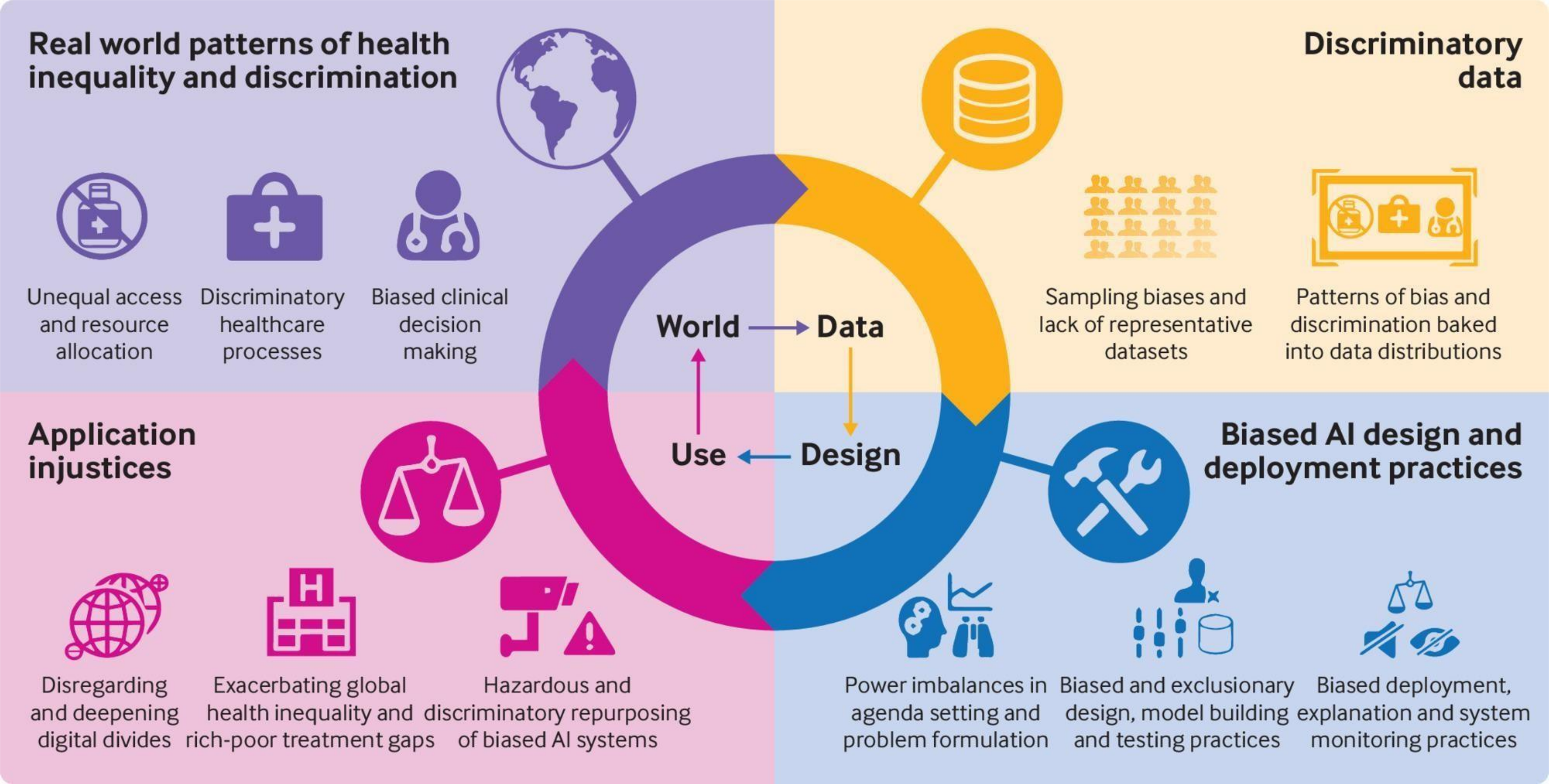
a) Data representativeness

b) Design limitations

c) Data documentation

d) Generalizability

The big picture



Real world patterns of health inequality and discrimination

Unequal access and resource allocation

Discriminatory healthcare processes

Biased clinical decision making

Discriminatory data

Sampling biases and lack of representative datasets

Patterns of bias and discrimination baked into data distributions

Application injustices

Disregarding and deepening digital divides

Exacerbating global health inequality and rich-poor treatment gaps

Hazardous and discriminatory repurposing of biased AI systems

Biased AI design and deployment practices

Power imbalances in agenda setting and problem formulation

Biased and exclusionary design, model building and testing practices

Biased deployment, explanation and system monitoring practices

Example: 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**.
- However, the algorithms used are basing most of their knowledge on a repository of **skin images from primarily fair-skinned populations.**
- **Bias emanates from unrepresentative training data that reflects historical inequalities:** decades of clinical research have focused primarily on people with light skin.
- The solution: **expand the archive to include as many skin types as possible**

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.

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

U.S. lacks a comprehensive federal AI law

EU sets global standards with first major AI regulations

- **Europe became the first major world power to enact comprehensive AI regulations**, covering areas like transparency, use of AI in public spaces, and high-risk systems.
- **High-impact models with systemic risks** face stricter requirements, including model evaluation, risk mitigation, and incident reporting.
- Requires **models to comply with transparency obligations before they are put on the market**: drawing up documentation, complying with EU law and disseminating summaries about the content used for training.

Federal AI Governance Policy:

- The **White House**, **Congress**, and various federal agencies have been actively shaping AI governance.
- The **Federal Trade Commission**, the **Consumer Financial Protection Bureau**, and the **National Institute of Standards and Technology** have all contributed to AI-related initiatives and policies.
- Notably, existing laws do apply to AI technology, and the focus is on understanding how these laws intersect with AI rather than creating entirely new AI-specific legislation
- NIST – New guidance

Avoiding perpetuating bad AI: mitigating bias in datasets

Strategies to mitigate bias in datasets:

1. **Identify potential sources of bias:** Analyze data collection methods, sampling procedures, and variable selection for potential biases. Testing for biases in datasets and algorithmic models is **crucial for ensuring fairness and reliability** in data science.

2. **Utilize bias mitigation techniques:** Apply techniques like data balancing, weighting, or fairness-aware algorithms to mitigate bias in the data.

3. **Promote transparency and responsible AI practices:** Document the limitations of the data and potential biases to ensure responsible use of AI models trained on the dataset.

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

Poll

What techniques would you prioritize to address bias in datasets, and why?

Poll

Which technique involves increasing the representation of underrepresented groups in a dataset?

a) Undersampling

b) Oversampling

c) Reweighting

d) Hypothesis testing

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

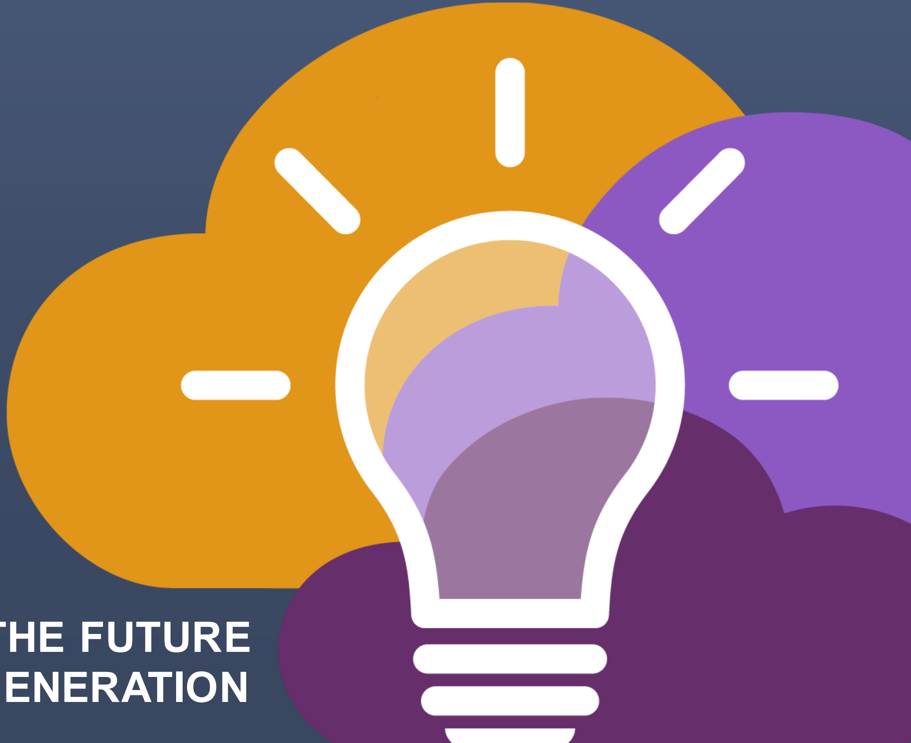
Testing for biases in algorithms

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

ScHARe

Computational strategies: traditional statistics



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Computational strategies

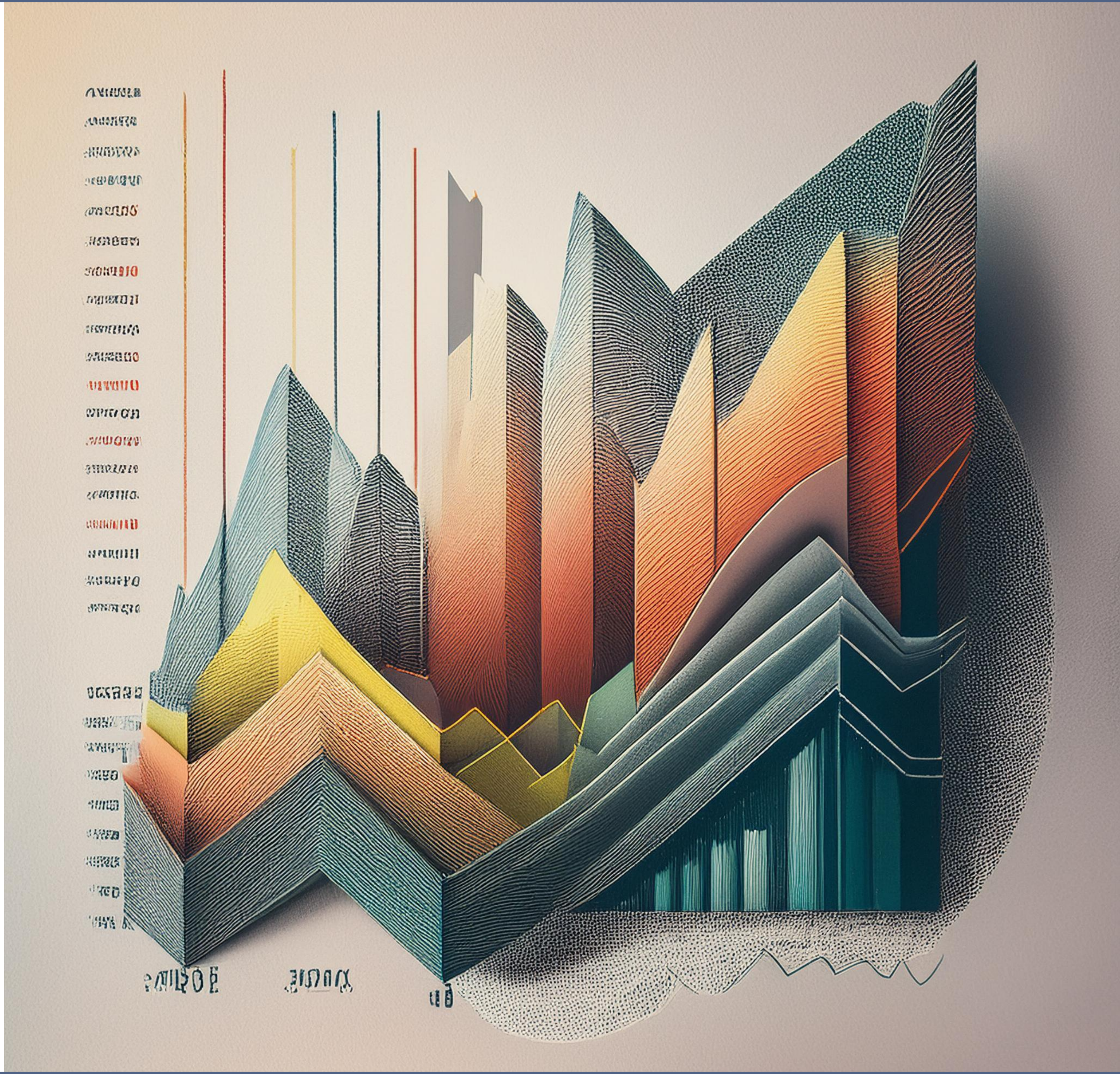
We will provide examples of **computational strategies used in healthcare disparities research**

Objectives:

- **Clarify the decision-making process for choosing between traditional statistics and Artificial Intelligence/Machine Learning**
- Explain the differences between these approaches and help you select the most suitable strategies for your analysis goals

Traditional statistics

- **Strengths:** robust, interpretable, well-established methodology
- **Weaknesses:** limited predictive power, assumption-dependent, often focused on hypothesis testing
- **Data types & use cases:** numerical data, identifying trends, correlations, causal relationships
- **Popular Python libraries:** NumPy, SciPy, Pandas



1. Descriptive statistics

- **Strategy:** Summarizing and describing key features of healthcare data, such as mean, median, standard deviation, and percentiles
- **Applications:** Understanding the central tendency and variability in healthcare variables
- **Python Libraries:** NumPy, pandas

2. Inferential statistics

- **Strategy:** Making predictions or inferences about a population based on a sample from that population
- **Applications:** Drawing conclusions about healthcare disparities from a subset of relevant data
- **Python Libraries:** SciPy, statsmodels

3. Hypothesis testing

- **Strategy:** Evaluating statistical significance to determine whether observed differences are likely to be real or due to chance
- **Applications:** Testing hypotheses about healthcare interventions or disparities
- **Python Libraries:** SciPy, statsmodels

4. Analysis of variance (ANOVA)

- **Strategy:** Assessing the statistical significance of differences among group means in healthcare data
- **Applications:** Comparing means across multiple categories to identify significant differences
- **Python Libraries:** SciPy, statsmodels

5. Chi-Square test

- **Strategy:** Assessing the association between categorical variables in healthcare datasets
- **Applications:** Examining relationships between demographic factors and health outcomes
- **Python Libraries:** SciPy, pandas

6. Regression analysis

- **Strategy:** Modeling the relationship between dependent and independent variables in healthcare data
- **Applications:** Predicting health outcomes based on various factors, identifying disparities
- **Python Libraries:** Statsmodels, scikit-learn

7. Survival analysis

- **Strategy:** Analyzing time-to-event data, such as the time until a patient experiences a particular health event
- **Applications:** Studying disparities in disease progression or survival rates
- **Python Libraries:** Lifelines, statsmodels

8. Correlation analysis

- **Strategy:** Examining the strength and direction of relationships between two continuous variables in healthcare datasets
- **Applications:** Assessing associations between risk factors and health outcomes
- **Python Libraries:** NumPy, pandas

9. Logistic regression

- **Strategy:** Modeling the probability of a binary outcome in healthcare data
- **Applications:** Analyzing factors influencing the likelihood of specific health events
- **Python Libraries:** Statsmodels, scikit-learn

10. Bayesian statistics

- **Strategy:** Updating beliefs about parameters based on new evidence in a probabilistic framework
- **Applications:** Incorporating prior knowledge into healthcare disparities research
- **Python Libraries:** PyMC3, Stan

11. Time series analysis

- **Strategy:** Analyzing temporal patterns and trends in healthcare data
- **Applications:** Studying disparities over time in health outcomes or interventions
- **Python Libraries:** Statsmodels, Pandas

Poll

What statistical method can assess the statistical significance of differences among group means in healthcare data?

a) Correlation analysis

b) Regression analysis

c) Analysis of variance (ANOVA)

d) Chi-square test