ScHARe

A Conceptual Model for Using AI to Advance Ethical Health Research

January 15, 2025

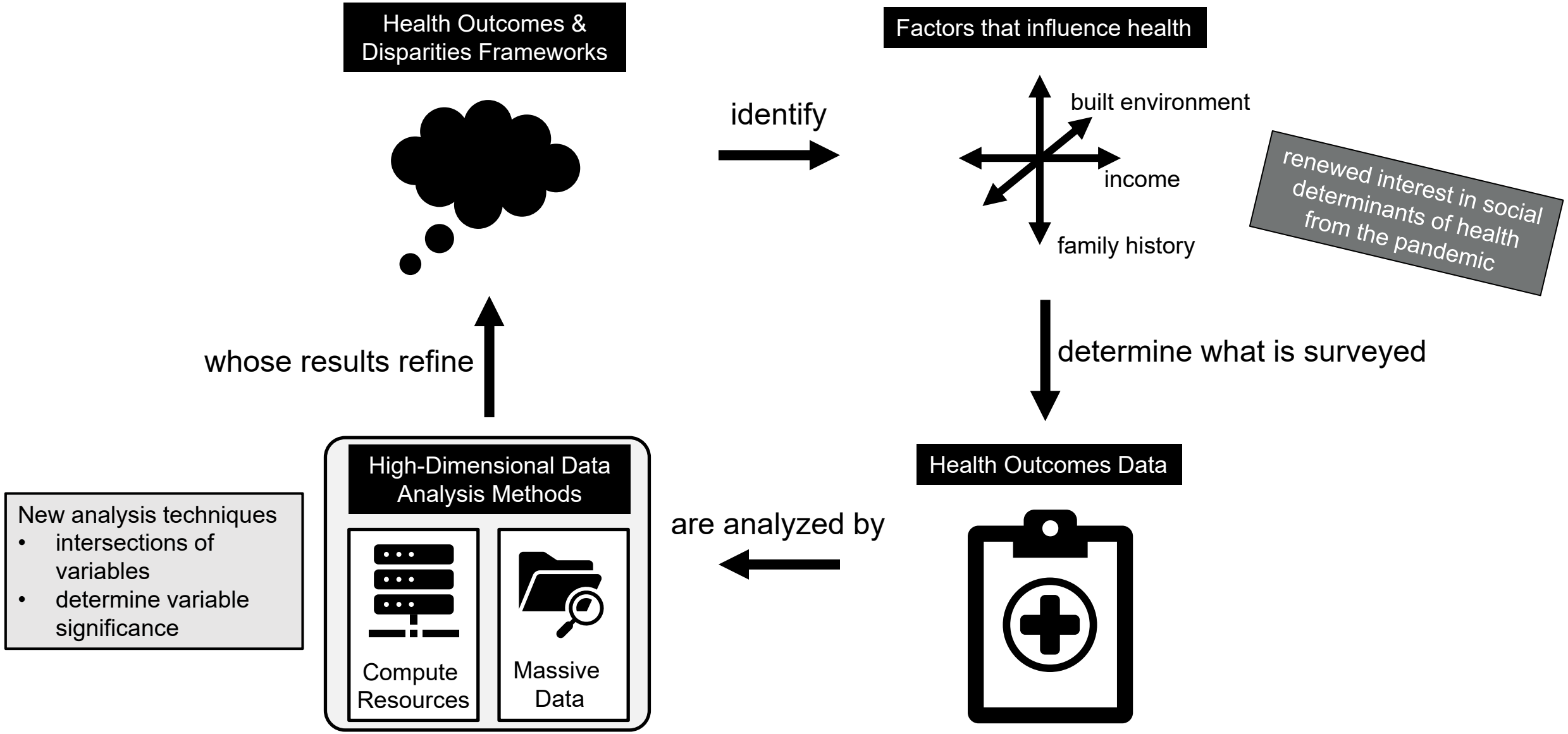
Deborah Guadalupe Duran, PhD · NIMHD
Elif Dede Yildirim, PhD · NIMHD
Mark Aronson, PhD · NIMHD

ScHARe

A Conceptual Model for Using AI

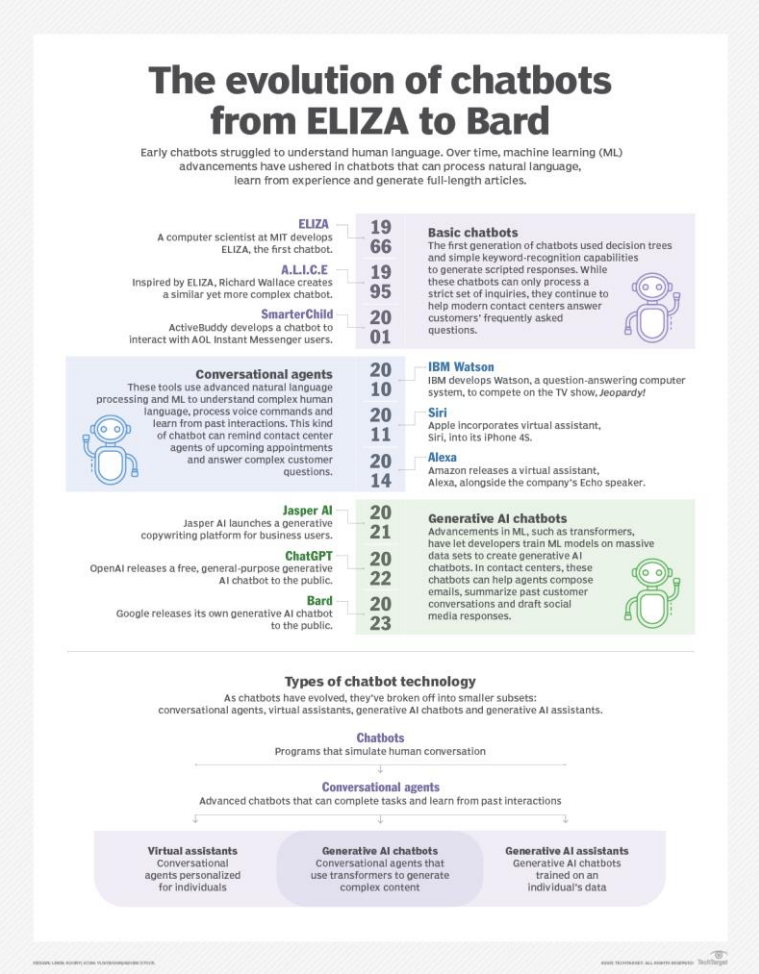
BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Data science is poised to accelerate the health outcomes research cycle

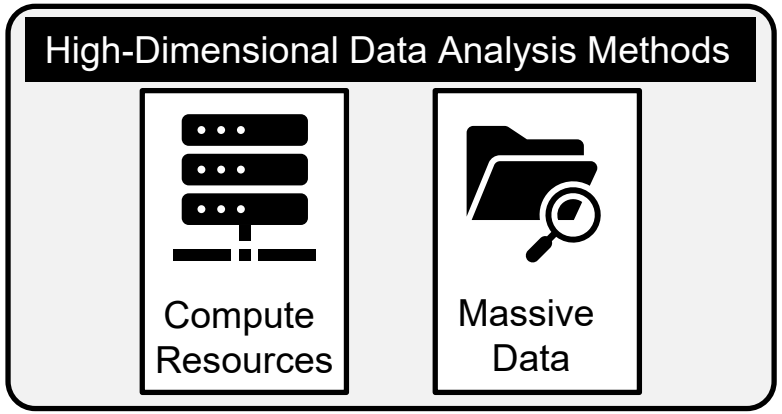


Recent focus on AI is new attention on a long-existing field

AI has, in some shape or form, been around since the 1960s



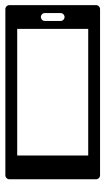
Fueling the current interest in AI is the intersection of two trends



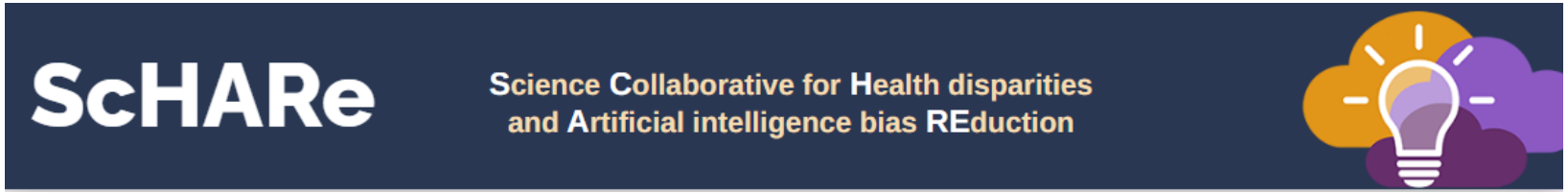
We are hearing about it more because it is being used in more places

chatbots

online customer service

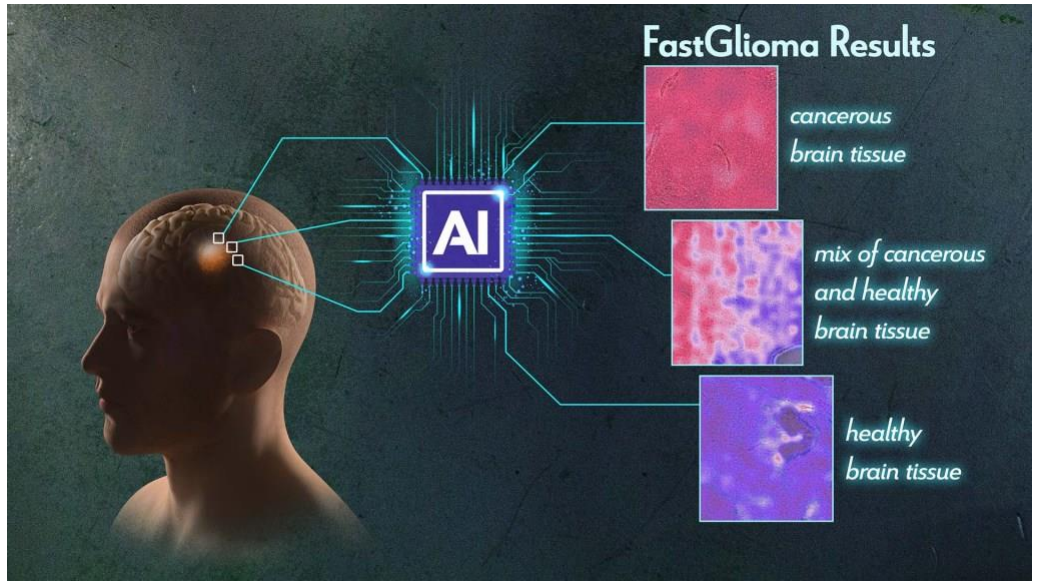


phone assistants



we are here to learn more and better understand the technology, its strengths and weaknesses, the areas it can impact and the concerns we need to address as we engage with it

AI Advancements have fueled research advancements



Machine learning to compare state to state differences in access to opioid treatment

HSR Health Services Research

RESEARCH ARTICLE | Open Access

Using machine learning to advance disparities research: Subgroup analyses of access to opioid treatment

Yinfei Kong PhD, Jia Zhou PhD, Zemin Zheng PhD, Hortensia Amaro PhD, Erick G. Guerrero PhD

First published: 17 October 2021 | https://doi.org/10.1111/1475-6773.13896 | Citations: 4

Machine learning to build comprehensive view of drivers of health disparities

Health Affairs Scholar, 2024, 2(3), 1–7
https://doi.org/10.1093/haschl/qxae017
Advance access publication: February 14, 2024
Research Article

Health Affairs Scholar
EMERGING & GLOBAL HEALTH POLICY
OXFORD

American clusters: using machine learning to understand health and health care disparities in the United States

Diana M. Bowser[1,*], Kaili Mauricio[1], Brielle A. Ruscitti[2], William H. Crown[2]

[1]Connell School of Nursing, Boston College, Chestnut Hill, MA 02467, United States
[2]Heller School for Social Policy and Management, Brandeis University, Waltham, MA 02454, United States
*Corresponding author: Connell School of Nursing, Boston College, Chestnut Hill, MA 02467. Email: bowserdi@bc.edu

Article

Foundation models for fast, label-free detection of glioma infiltration

https://doi.org/10.1038/s41586-024-08169-3
Received: 7 March 2024
Accepted: 8 October 2024
Published online: 13 November 2024
Open access

Akhil Kondepudi[1,2], Melike Pekmezci[3], Xinhai Hou[1,2], Katie Scotford[4], Cheng Jiang[1,2], Akshay Rao[1], Edward S. Harake[1], Asadur Chowdury[1], Wajd Al-Holou[5], Lin Wang[1], Aditya Pandey[5], Pedro R. Lowenstein[5], Maria G. Castro[5], Lisa Irina Koerner[6], Thomas Roetzer-Pejrimovsky[7,13], Georg Widhalm[6], Sandra Camelo-Piragua[8], Misha Movahed-Ezazi[9], Daniel A. Orringer[10], Honglak Lee[11], Christian Freudiger[12], Mitchel Berger[4], Shawn Hervey-Jumper[4] & Todd Hollon[1,5]

AI enables health outcomes research to operate at scale

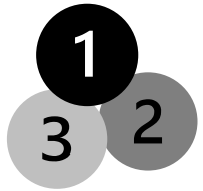
Sample Dataset

Participant ID	Race	Age	Health Outcome
12410294	North African	38	N
12487127	Alaskan Native	61	Y
12749182	White	54	Y

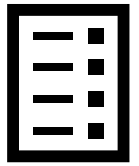
number of data features is (essentially) unlimited

number of data entries is (essentially) unlimited

Can consider many types of data



numeric



categorical

images

qualitative

Example: enables multi-level analysis

Community 1

Community factors
- transportation
- food availability
- access to healthcare

Individual 1

Individual factors
- demographic
- genetic
- physical activity

Individual 2

Individual 3

Community 2

Individual 4

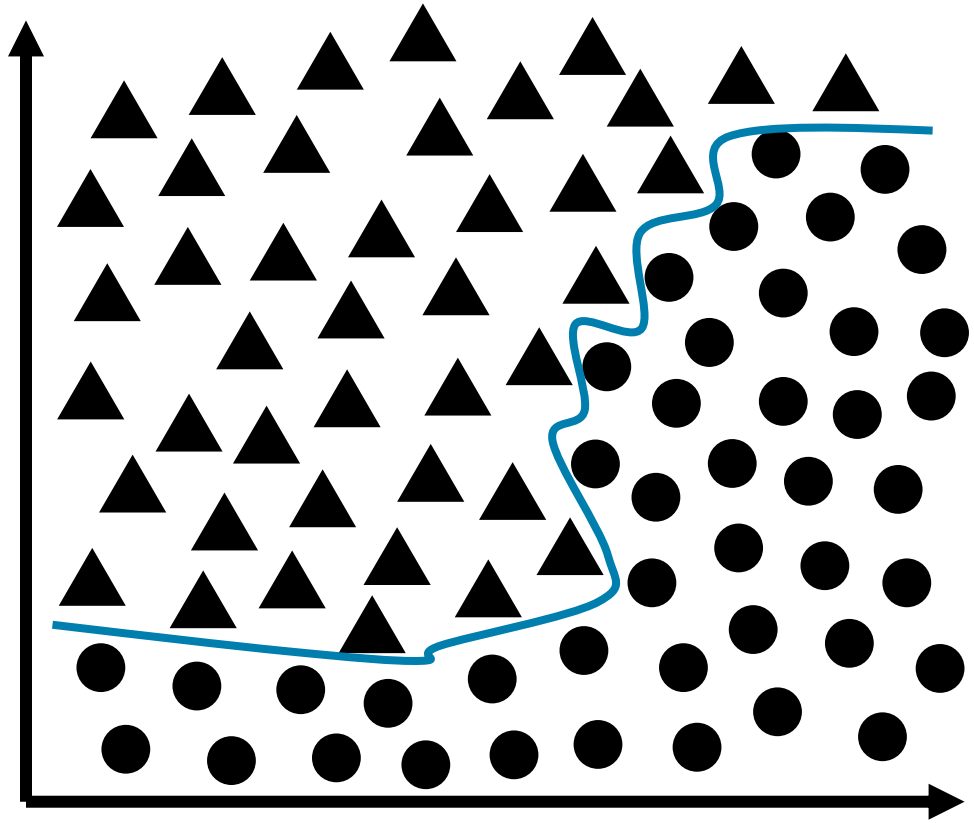
Individual 5

Individual 6

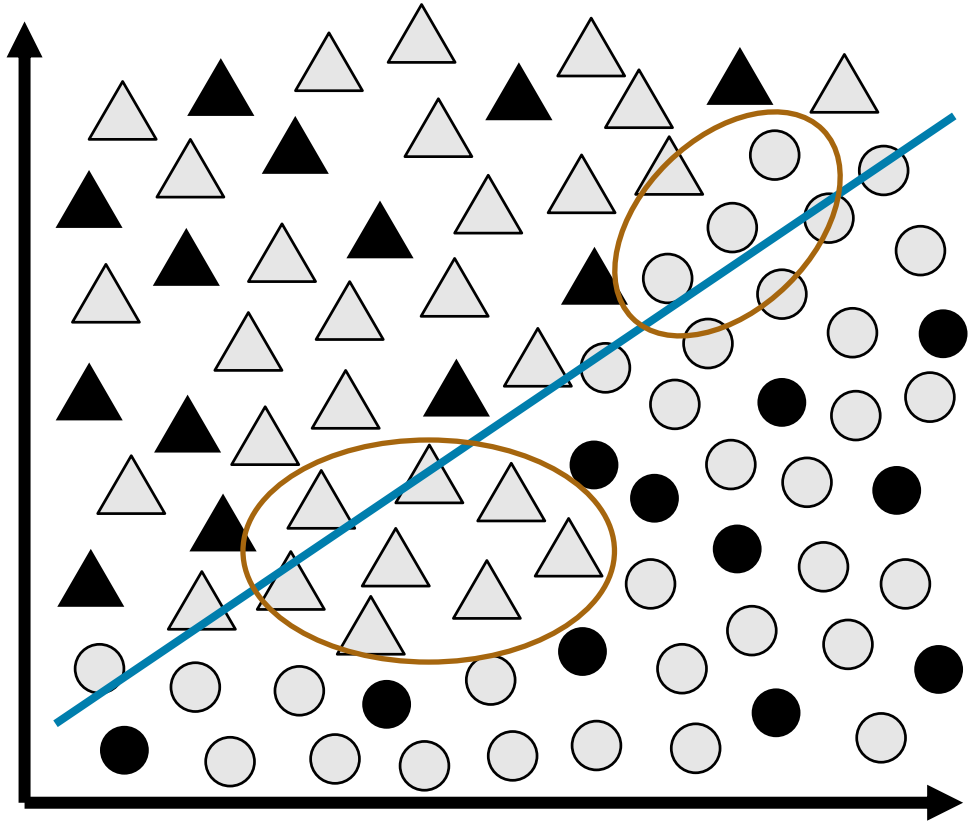
Machine learning enables us to create models that determine significant factors within and between different communities

AI models find high dimensional patterns from data

Perfect Data = Perfect Model



Finite Data = Imperfect Model



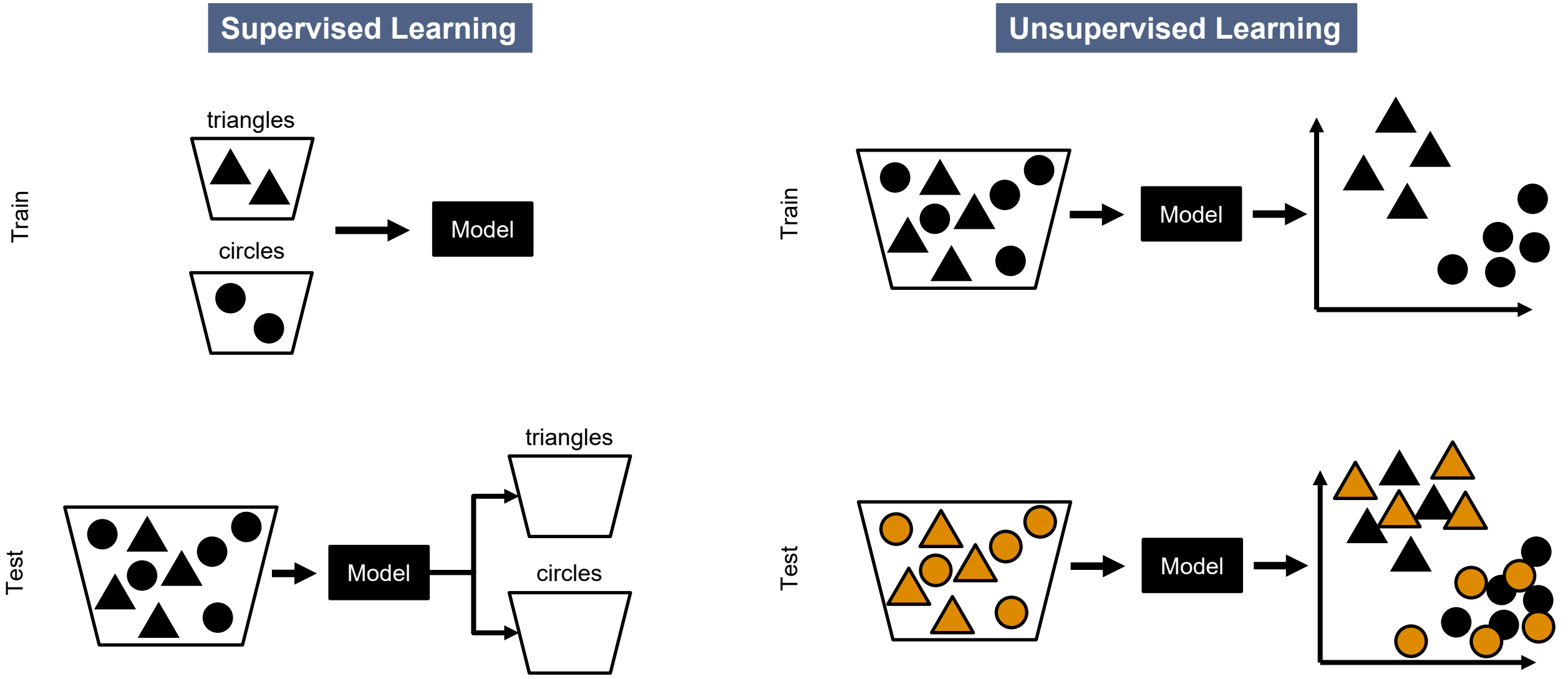
imperfect models lead to differential outcomes between groups (biases)

Slido Poll

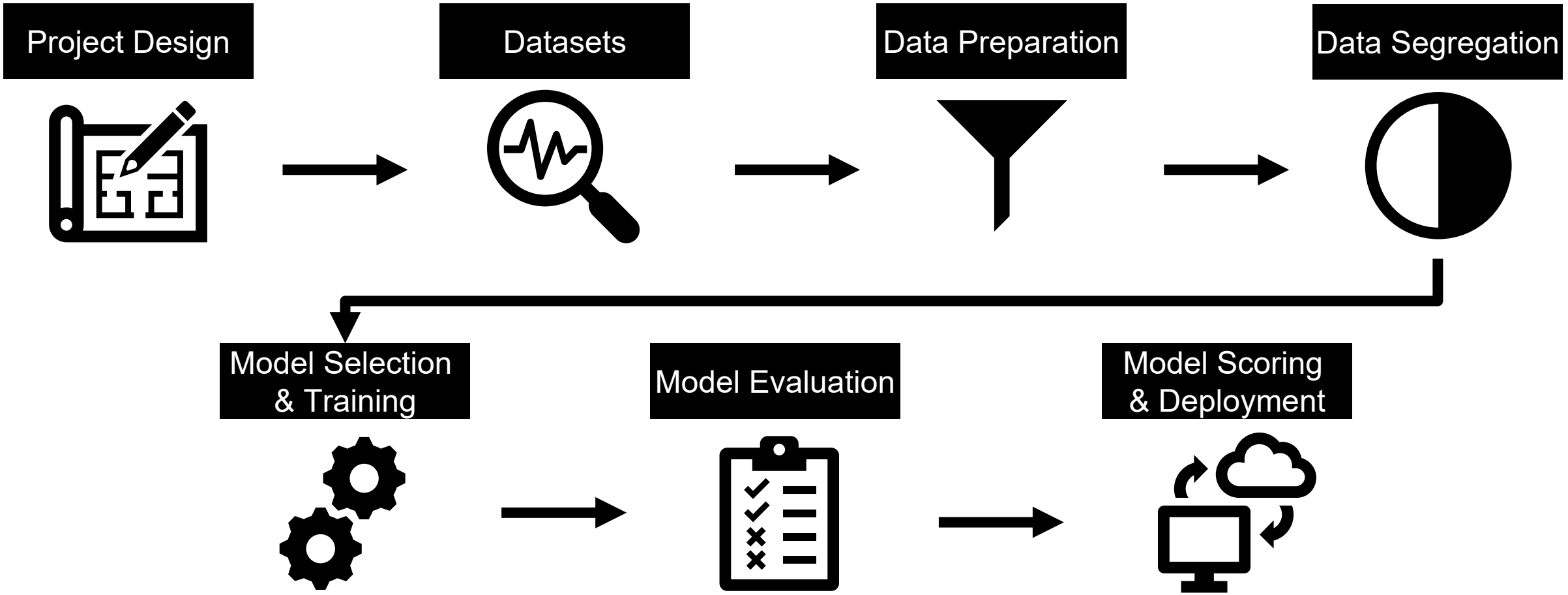
The current excitement about AI and Machine Learning has been brought on by the combination of which two trends?

a) New statistical analysis theories and cloud computing
b) Massive data and increased compute resources
c) Mass adoption of smartphones and new statistical analysis theories
d) Increased compute resources and novel health outcomes frameworks
e) Novel health outcomes frameworks and cloud computing

Supervised learning is a useful example case

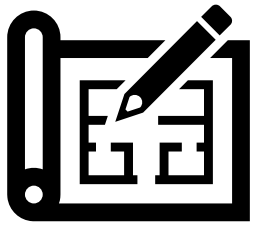


AI development pipeline



AI development pipeline

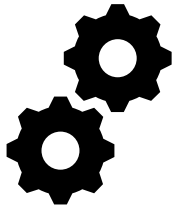
Project Design



What is it?

Crafting the research question and the data analysis strategy.

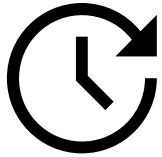
How to determine if an AI approach is appropriate



Is AI the right tool to answer the question?

Pattern Recognition

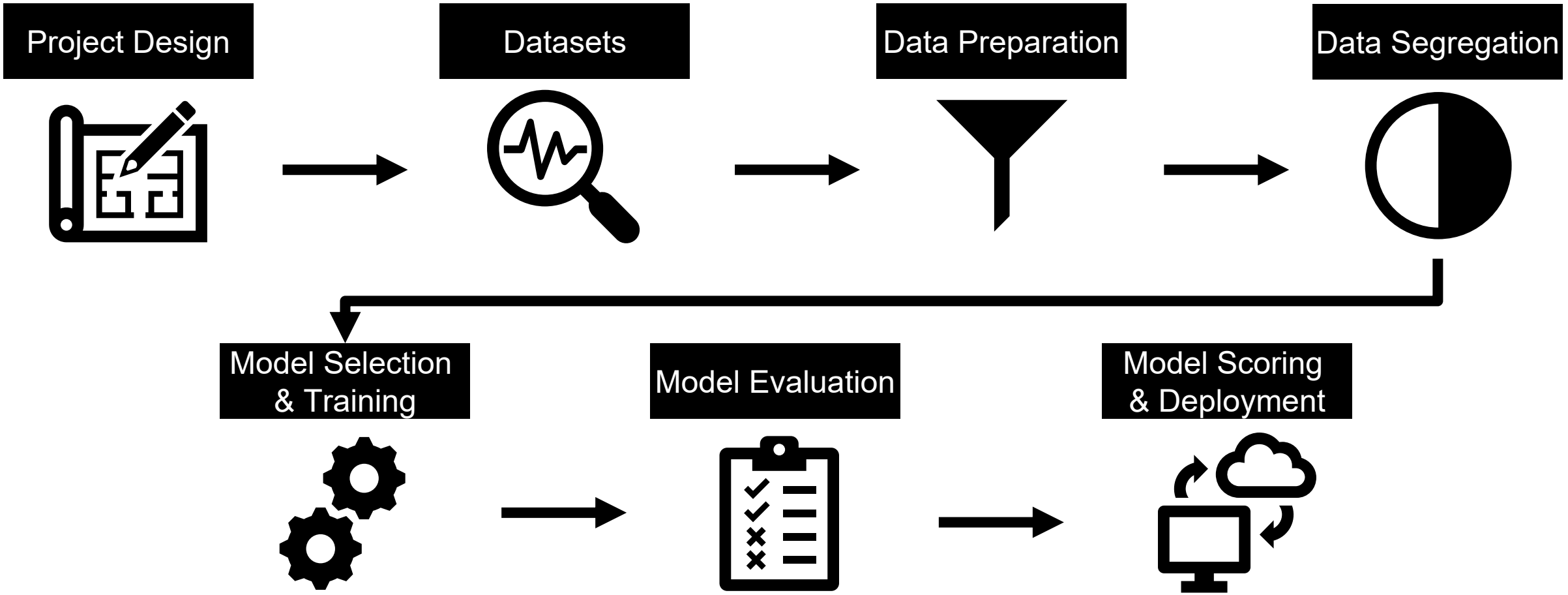
Prediction



Is the project design appropriate?

Is the data appropriate for the AI model to use?

AI development pipeline



Nonrepresentative data can lead to biases

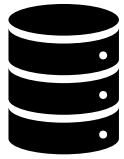
Datasets

What is it?

Acquiring data to be used for model development.

Data can be…

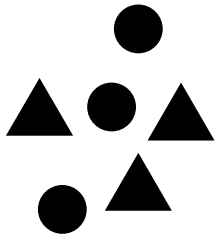
collected firsthand



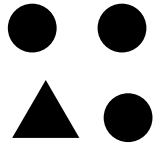
reused from existing databases

Where can biases arise?

the world

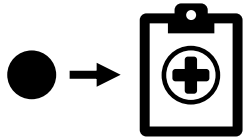


your data

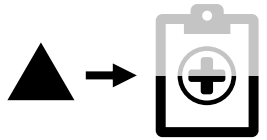


Nonrepresentative data (misbalanced sample sizes)

some groups have complete information

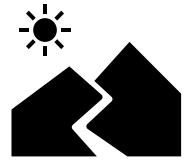


others do not



Disproportionately Missing data

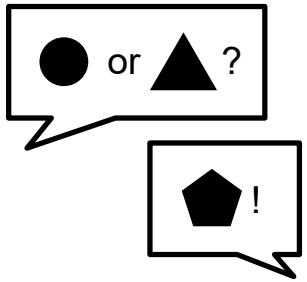
reality



metric

Misleading metrics

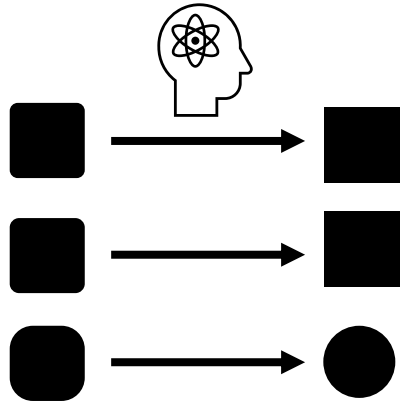
absent from health records



nonstandard answers

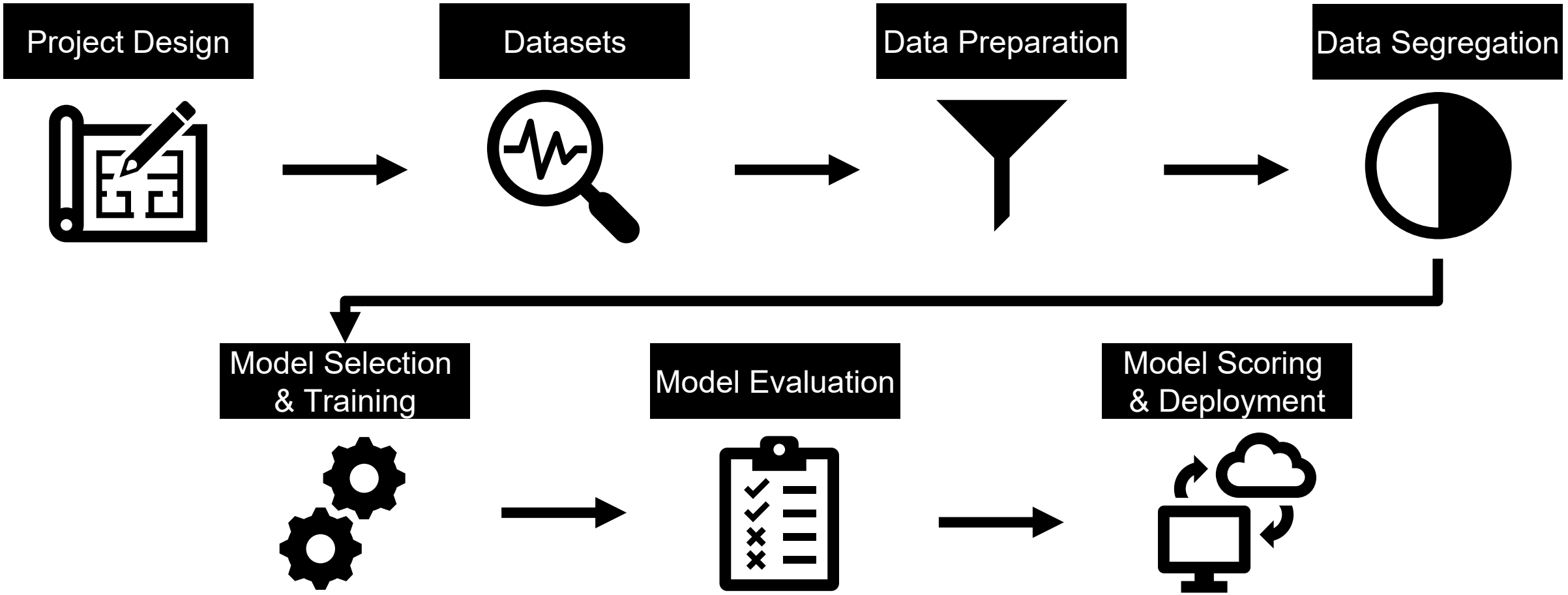
unstructured data

Difficult-to-capture Data (e.g. social determinants of health)



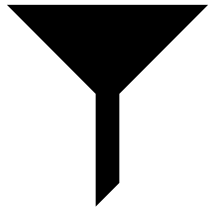
Inconsistent Semantics & Incorrect Data Labels

AI development pipeline



Data for AI model training must be prepared

Data Preparation

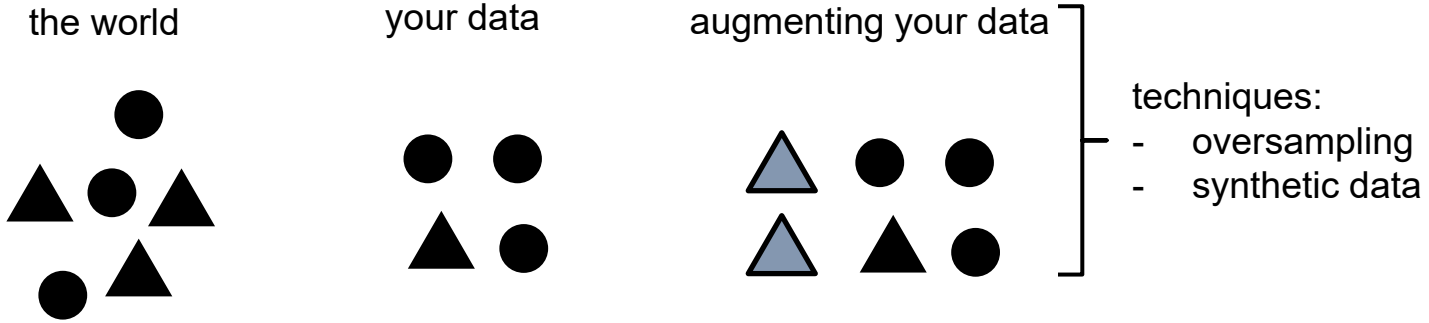


What is it?

Getting data ready to be used for model training and testing. Often includes “cleaning” and “transforming” steps.

This is hard and there are specialized people who do this: data managers, data wranglers, data “mungers”

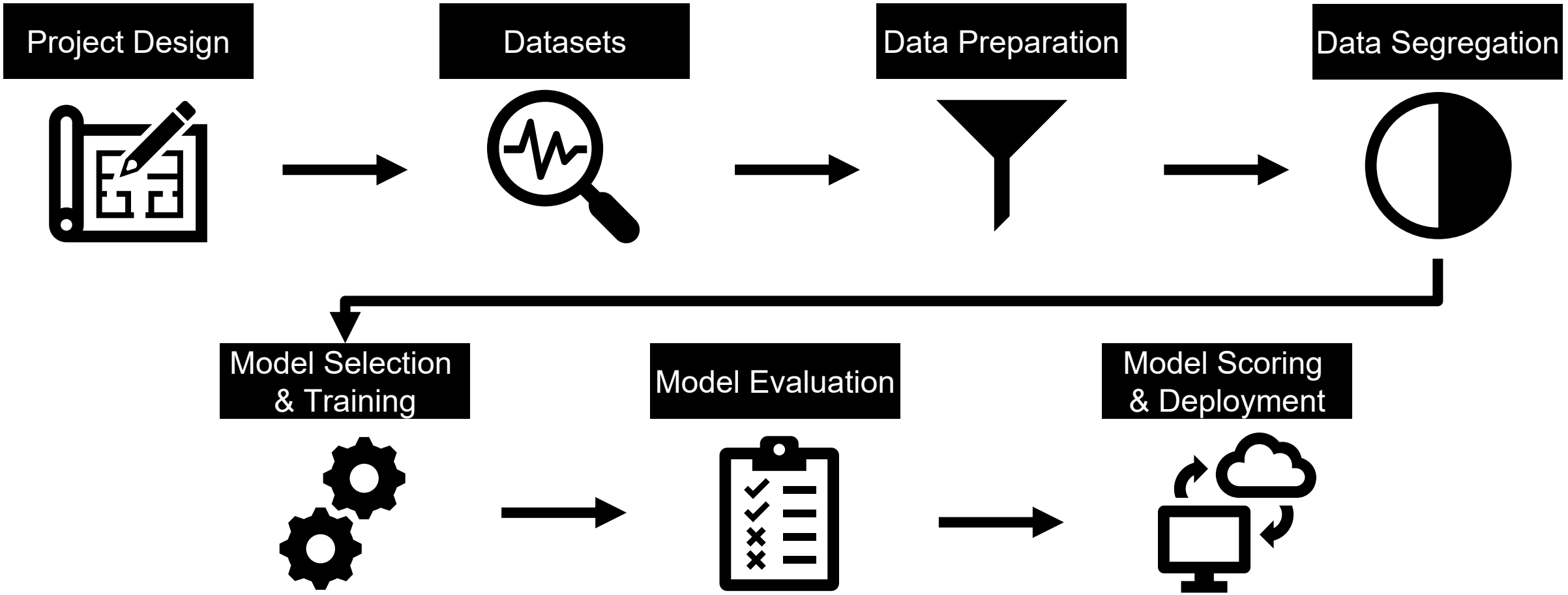
Where can biases arise?



Compensating for groups with smaller representation

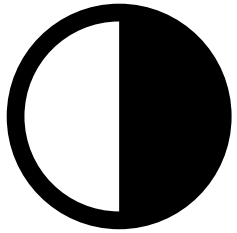
Filling in missing data

AI development pipeline

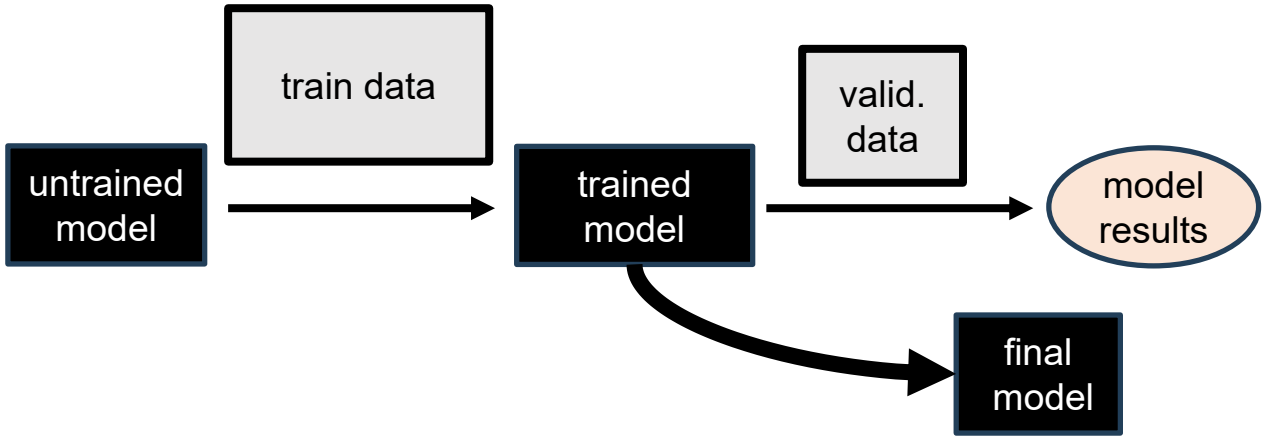


All data is split for model development and testing

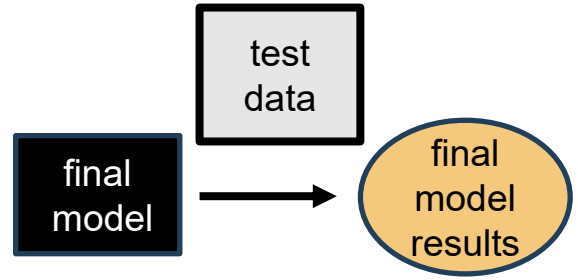
Data Segregation



development



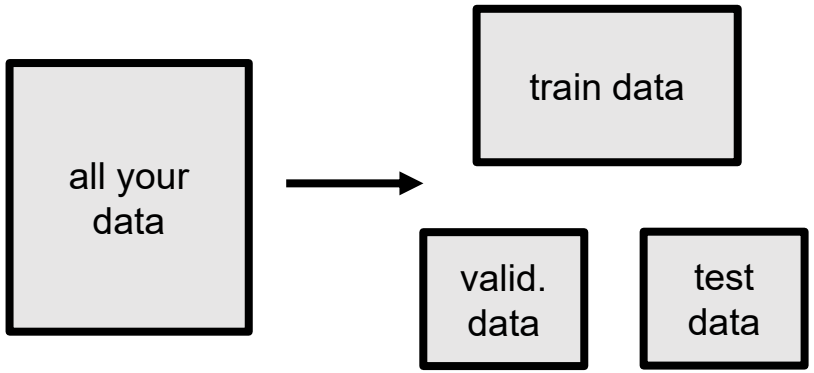
deployment



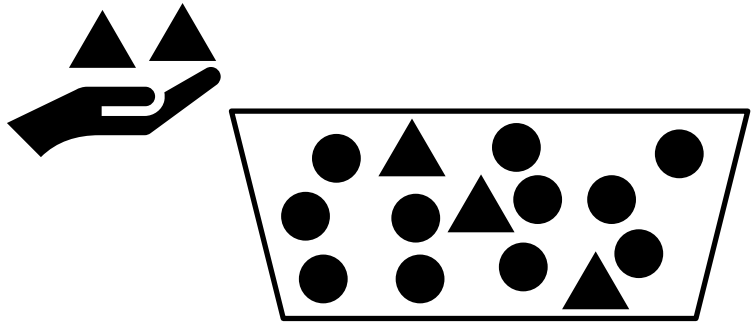
What is it?

Splitting up data into:

- data to be used to build the model (training data)
- data used to check the model building (validation data)
- data used to test the model (test data).



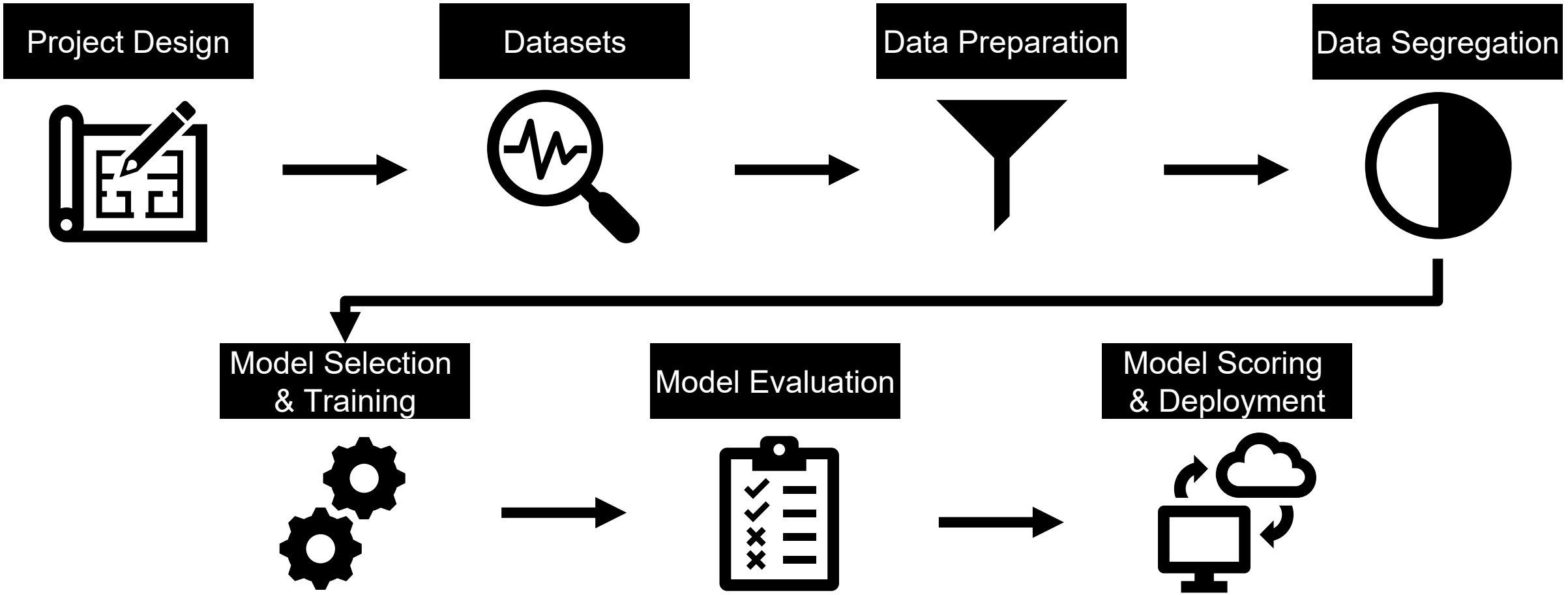
Where can biases arise?



Sampling strategies

Reweighing Strategies

AI development pipeline

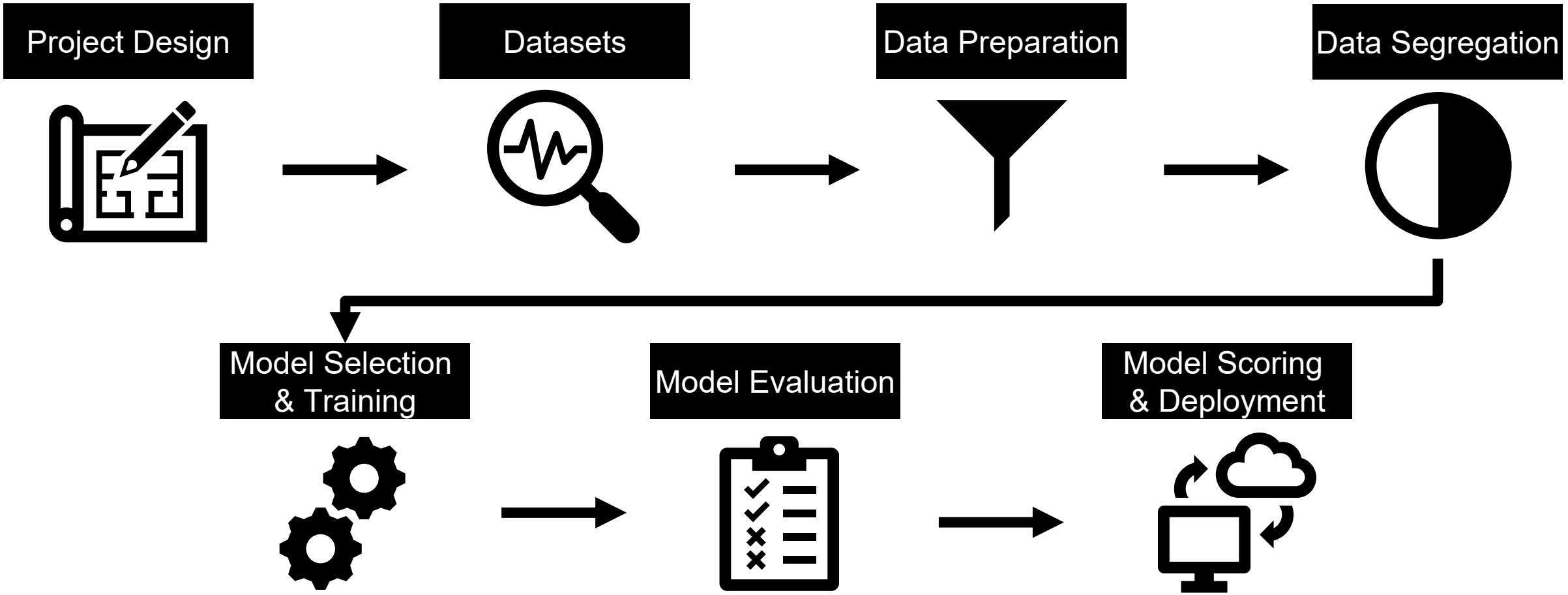


Slido Poll

Which of the following describe the purpose of validation data?

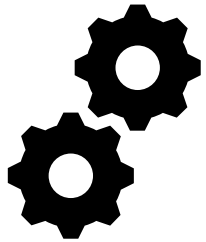
a) To train the model
b) To test the final performance of the model on unseen data
c) To evaluate the model during training
d) To store raw data collected before preprocessing
e) To monitor the compute resources during model training

AI development pipeline



AI Model training can introduce biases

Model Selection & Training



What is it?

Deciding which type of model to use and then training the model using the training data.

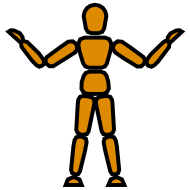
Common categories of models

- Classifiers vs regressors
- supervised vs unsupervised

Where can biases arise?

random forest

neural net

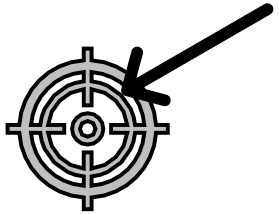


lasso regressor

Which model you pick

Training Parameters

“Hyperparameters”

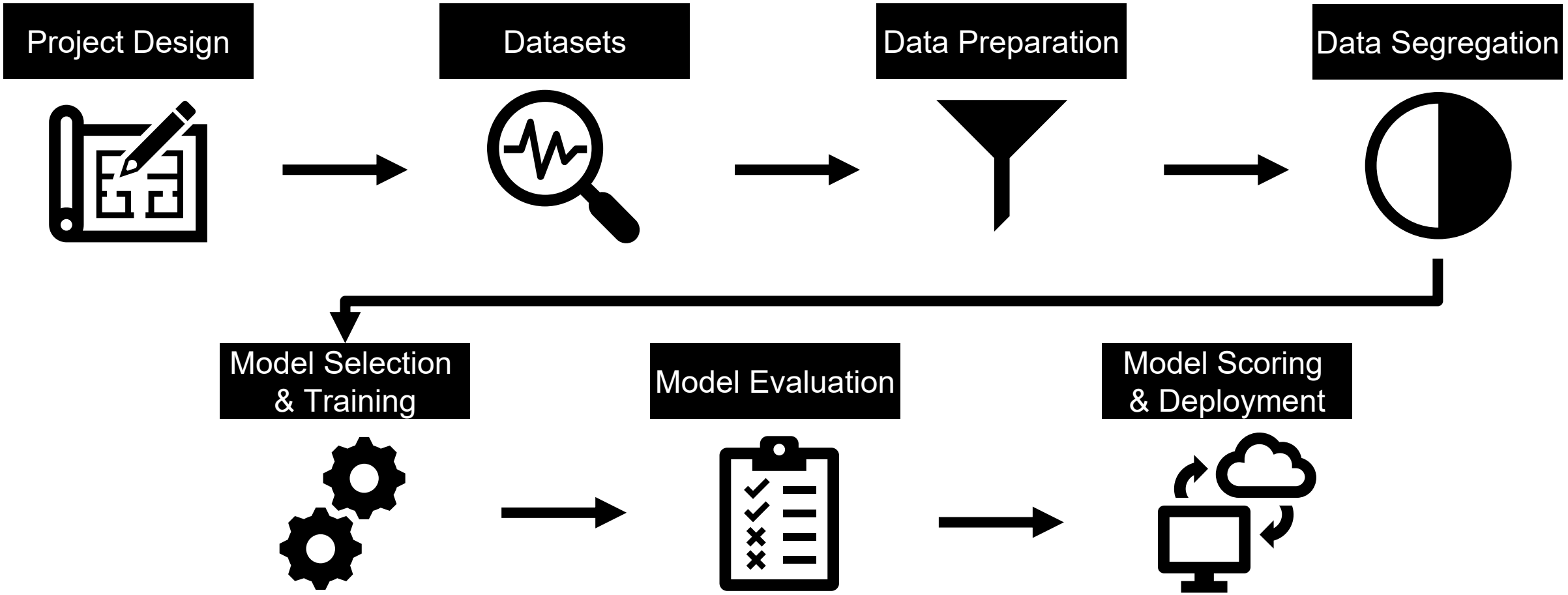


Error Calculation

“Loss functions”

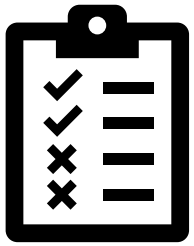
How it’s trained

AI development pipeline



Model evaluation is key for AI model development

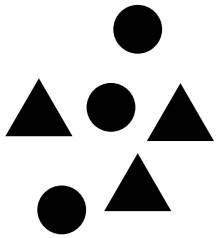
Model Evaluation

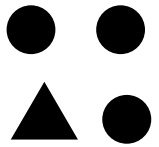


What is it?

Using various metrics to check the performance of different models and select the final parameters.

Where can biases arise?





Example metrics for model evaluation

using metrics that don't balance error can lead to biases

accuracy = $\frac{\#\ correct}{total}$ basic model is 75% accurate!

using metrics that balance errors can mitigate biases

F1 score and AUC – metrics that includes both correct and false positives and false negatives

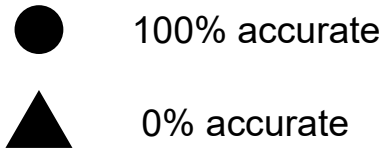
How the model performs across different parts of the data

using the same metric across all cohorts can lead to biases

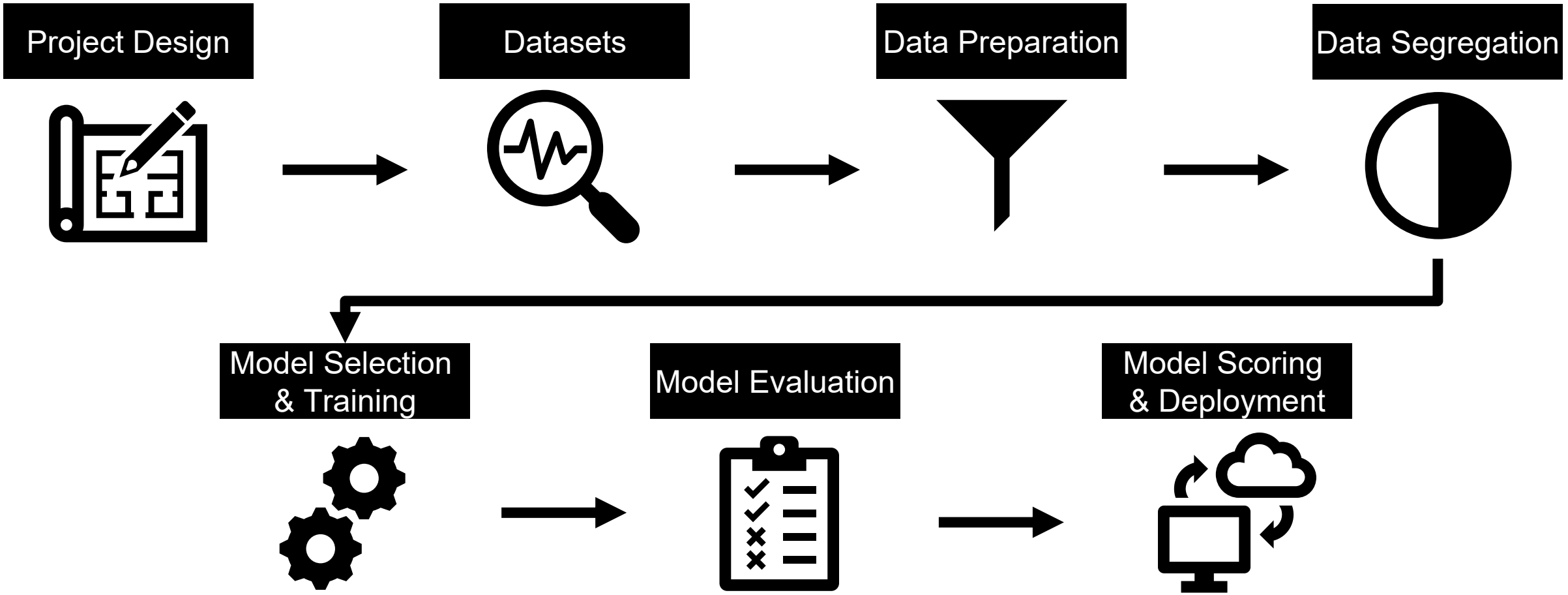
basic model → all data → 75% accurate

analyzing how the model works for each group can mitigate biases

basic model

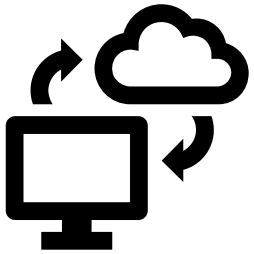


AI development pipeline



Deployment of AI requires proper use and trust

Model Scoring & Deployment

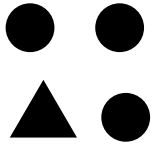


What is it?

The final scoring of the model performance using test data and the implementation of the model.

Where can biases arise?

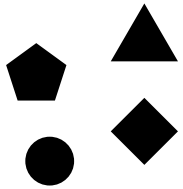
your data



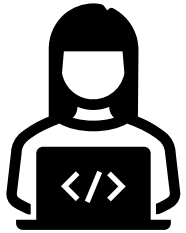
your model

basic model

deployment data

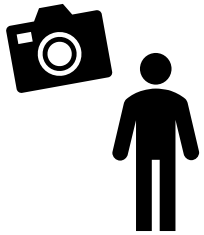


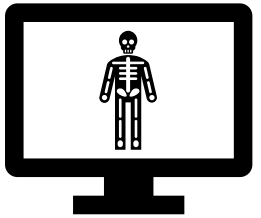
Using the model on data it was not trained on



basic model

How much users of the model **trust** the model

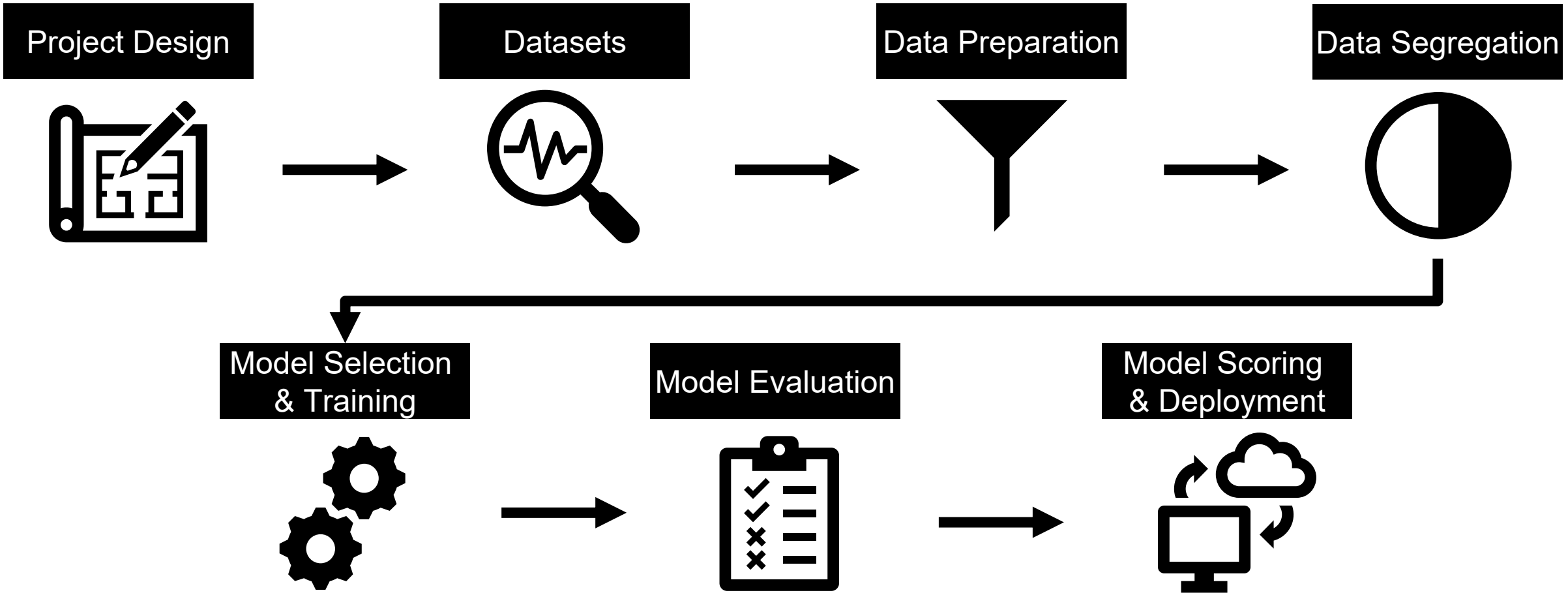




Transparency mitigates these biases by showing exactly what data the model was trained on (and therefore should be used on) and how it was developed

- promotes model understanding
- assists in determining model reuse
- requires documentation

AI development pipeline



Slido Poll

Why are model evaluation metrics important for mitigating model bias in AI systems?

a) They help ensure the model is trained on a larger dataset.

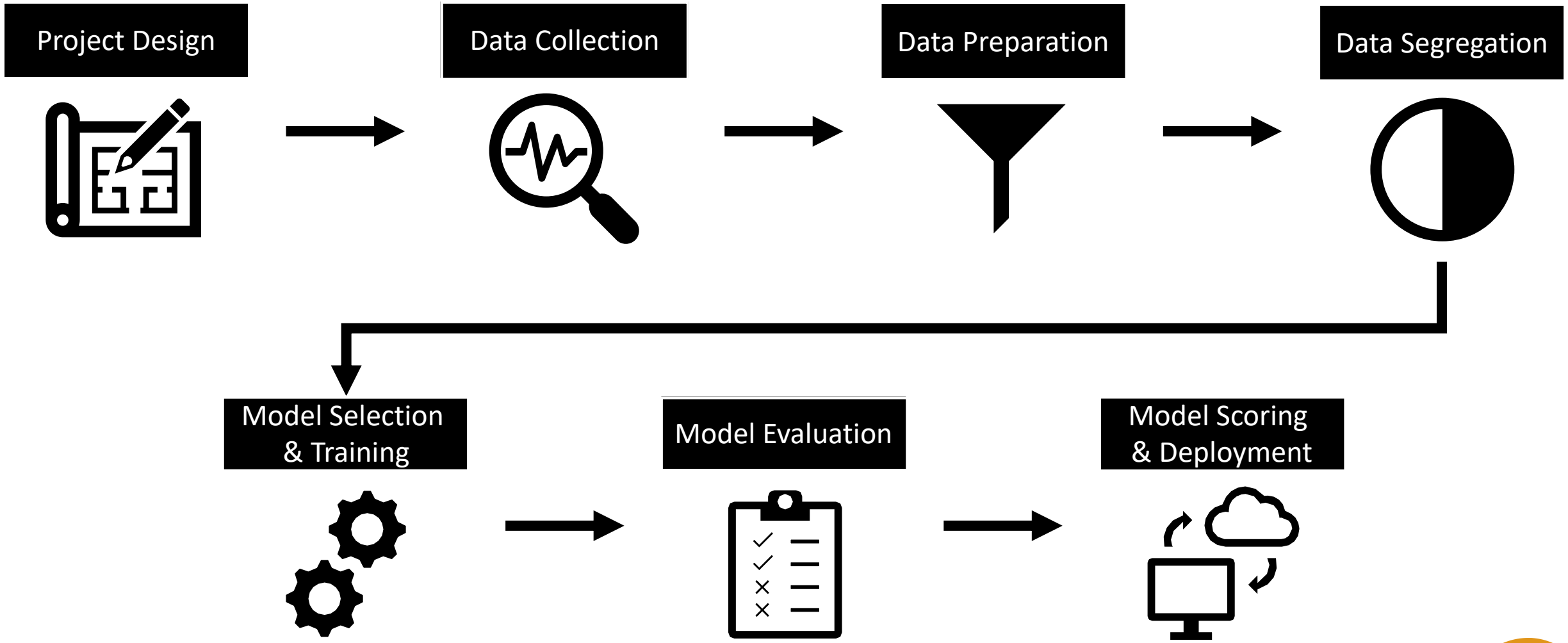
b) They identify disparities in model performance across different groups.

c) They automatically remove biased data from the training set.

d) They focus solely on improving the model's accuracy without considering fairness.

e) They ensure the model runs faster and uses fewer computational resources.

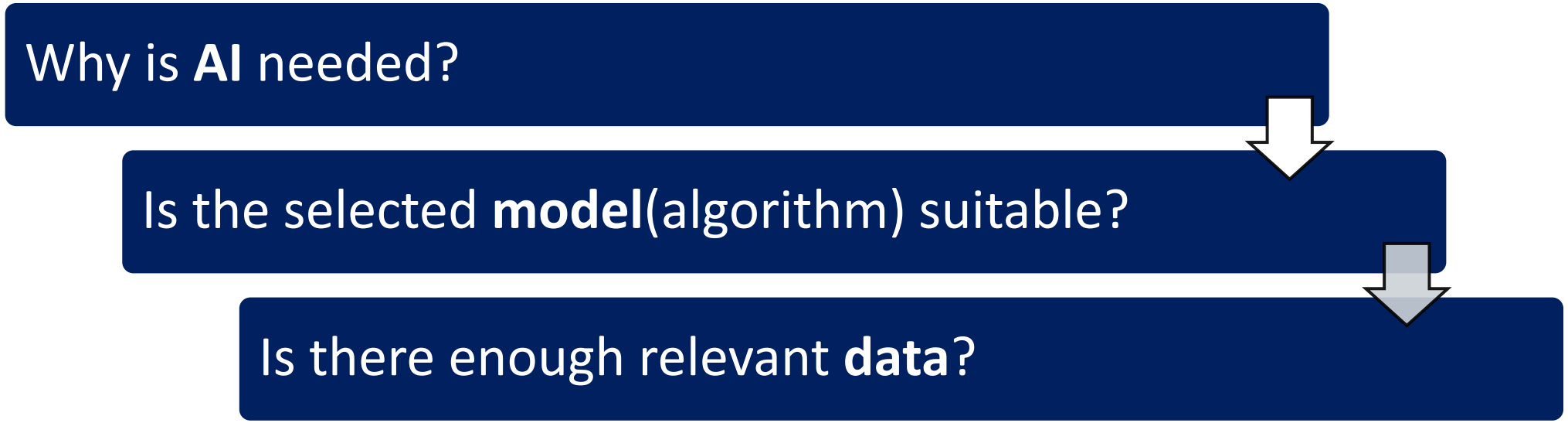
AI development pipeline



Project Design

Step 1. Project Design

Articulate the research problem and its significance



Involve the consumers/involved parties that will be affected by the AI system in the design process. Their insights can help identify potential biases and ensure that the model meets their needs and addresses their concerns.

Project Design

Why is AI needed?

What is the specific research question or problem that AI can address in this study?

Are you trying to uncover complex and interconnected relationships between variables?

Are you trying to reveal patterns using big data?

How does it differ from traditional methods?

What new insights or capabilities does it offer?

Project Design

Is the selected model suitable?

- Does the model align with the type of research question being asked?
- Does the model research questions match the available data characteristics?
- For the research question, does the model potentially lead to unexpected biases?
- Can the model handle the volume and dimensionality of the available data?
- Is the model robust/can produce reliable results given the limitations of the available data?

Project Design

Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records

Zheng et al. Translational Psychiatry (2020) 10:72
https://doi.org/10.1038/s41398-020-0684-2

Sample size = 236,347

Problem. Suicide is the tenth leading cause of death in the US, claiming the lives of more than 44,000 individuals in 2015. There is a need for an effective early warning system to identify those at high risk of attempting suicide.

Why AI? **Prediction accuracy** was limited when applied to a general population where the incidence of suicide attempts was extremely low. **Reasons** for a suicide attempt could be **complex** and associated to a **multi-level network** which limit revealing the different roles of **individual-level** and **population-level factors** in suicide attempts using univariate and multivariate analyses.

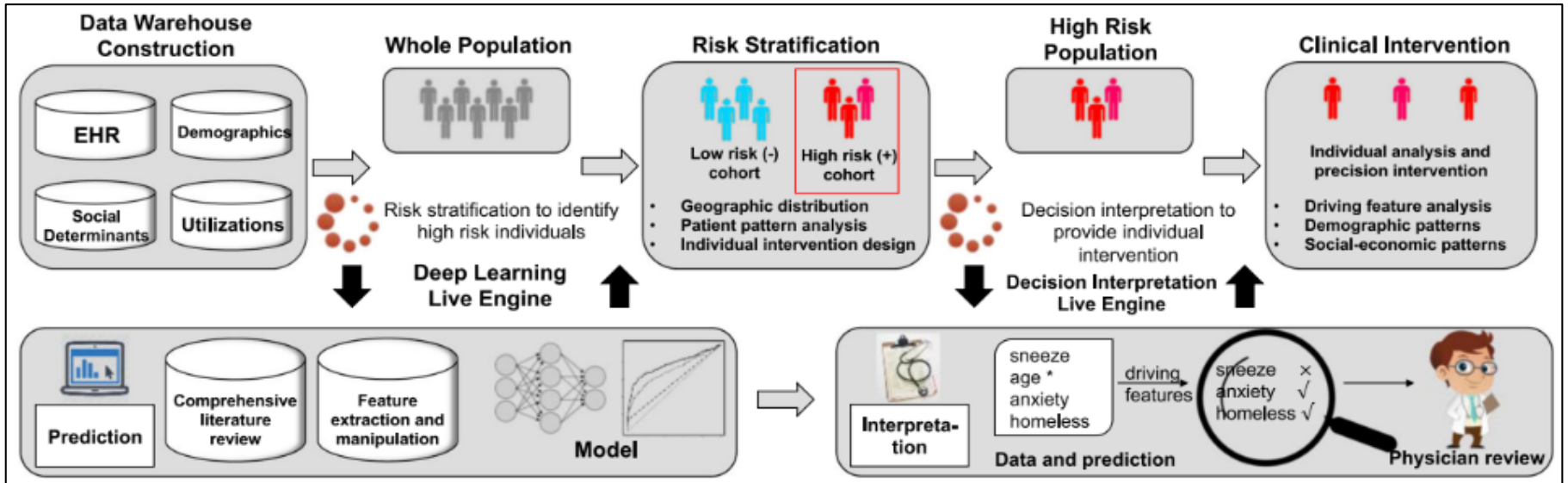


Fig. 1 Development of risk of suicide attempt early-warning system. The system is consist of the deep learning live engine and the decision interpretation live engine. The deep learning engine is design to provide a real-time risk stratification for the whole population, so that the high-risk population can be found in advance. The decision interpretation live engine is used to analyze the driving features of the high-risk population and help provide insight for individual intervention.

Selected Model. **Deep learning** algorithms are well suited to **uncover** and recognize (learn) the **hidden explanatory factors** of variation behind **complex data** and simultaneously maintain the utility of **generalization**.

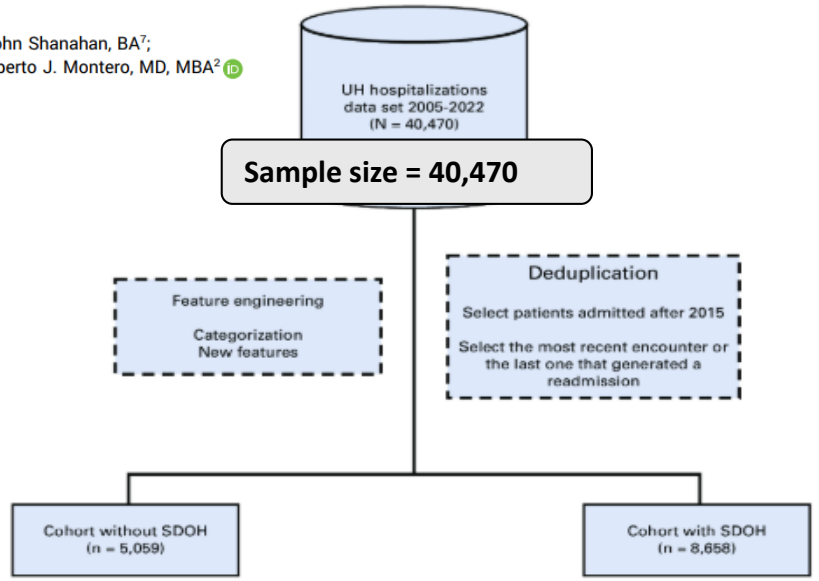
Project Design

Thirty-Day Unplanned Hospital Readmissions in Patients With Cancer and the Impact of Social Determinants of Health: A Machine Learning Approach

Nickolas Stabellini, BS[1,2,3,4]; Aziz Nazha, MD[5]; Nikita Agrawal, BA[6]; Merilys Huhn, BA[6]; John Shanahan, BA[7]; Nelson Hamerschlak, MD, PhD[8]; Kristin Waite, PhD[9]; Jill S. Barnholtz-Sloan, PhD[9,10]; and Alberto J. Montero, MD, MBA[2]

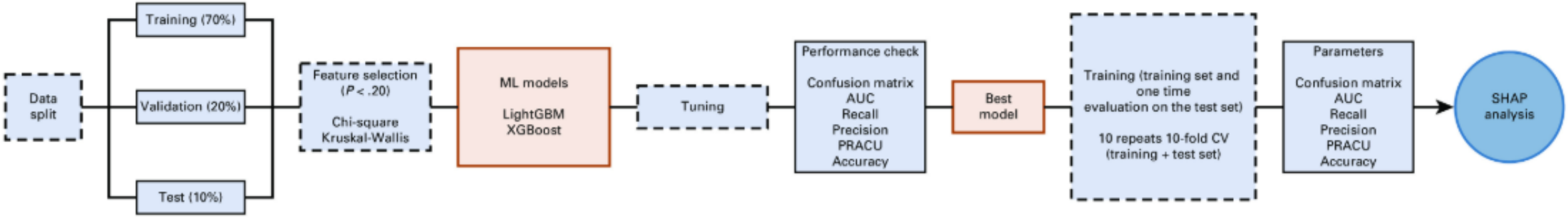
DOI https://doi.org/10.1200/CCI.22.00143

Why AI? Understanding **key drivers** of unplanned hospital **readmissions** in patients with cancer has been a topic of interest; however, the **multiple factors** involved have made applying traditional statistical methods difficult.



Problem. **Unplanned hospital readmissions** are approximately cost the US health care system **15-20 billion dollars** annually. **Preventing avoidable readmissions** has the potential to improve patient **quality of life** and help **reduce** hospital **costs**. ML–based models used to reduce hospital readmissions, but there is **a lack of** effective cancer-specific **models** that incorporate **social determinants of health.**

AI Pipeline



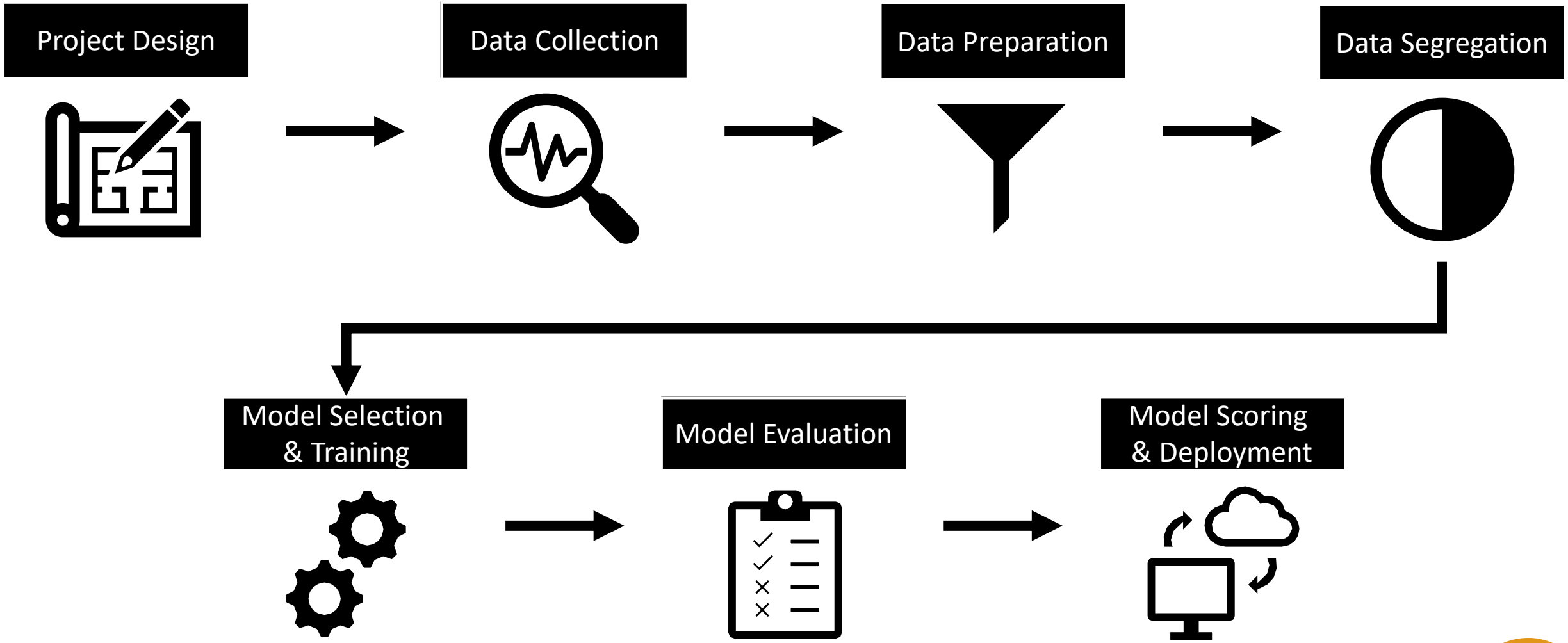
Selected Model. ML provides an alternative method for better understanding unplanned cancer readmissions, given its ability to **handle multiple variables** and to incorporate explainability via tree-based methods, which can offer crucial perspectives into clinical decision-making applications.

Project Design

Is there enough relevant data?

- What data sources are available for this research?
- What is the volume and format of the available data?
- Does the data contain the necessary features and labels?
- Is the data representative and inclusive of targeted audience?
- Is primary data collection required (in conjunction)?

AI development pipeline



Step 2. Data Collection

Data Collection

- ✓Leverage existing data repositories
- ✓Consider diverse range of data formats including, text, images, audio, relational databases
- ✓Prioritize data directly related to the problem being addressed

Electronic Health Records
Social Media
Insurance Claims
National Registries
Wearables
Census
Clinical Trials
Scientific Literature

Problem Specific Data Extraction

Problem Specific Data Collection

Data Design

Problem Specific Dataset

Problem Specific Dataset

Data Types

Data Collection

Unstructured Data

- Does not have a predefined format
- Requires more storage
- Text documents, images, audio, and video files

Structured Data

- Organized in a predefined format (rows, columns)
- Requires less storage
- Numbers, strings, dates

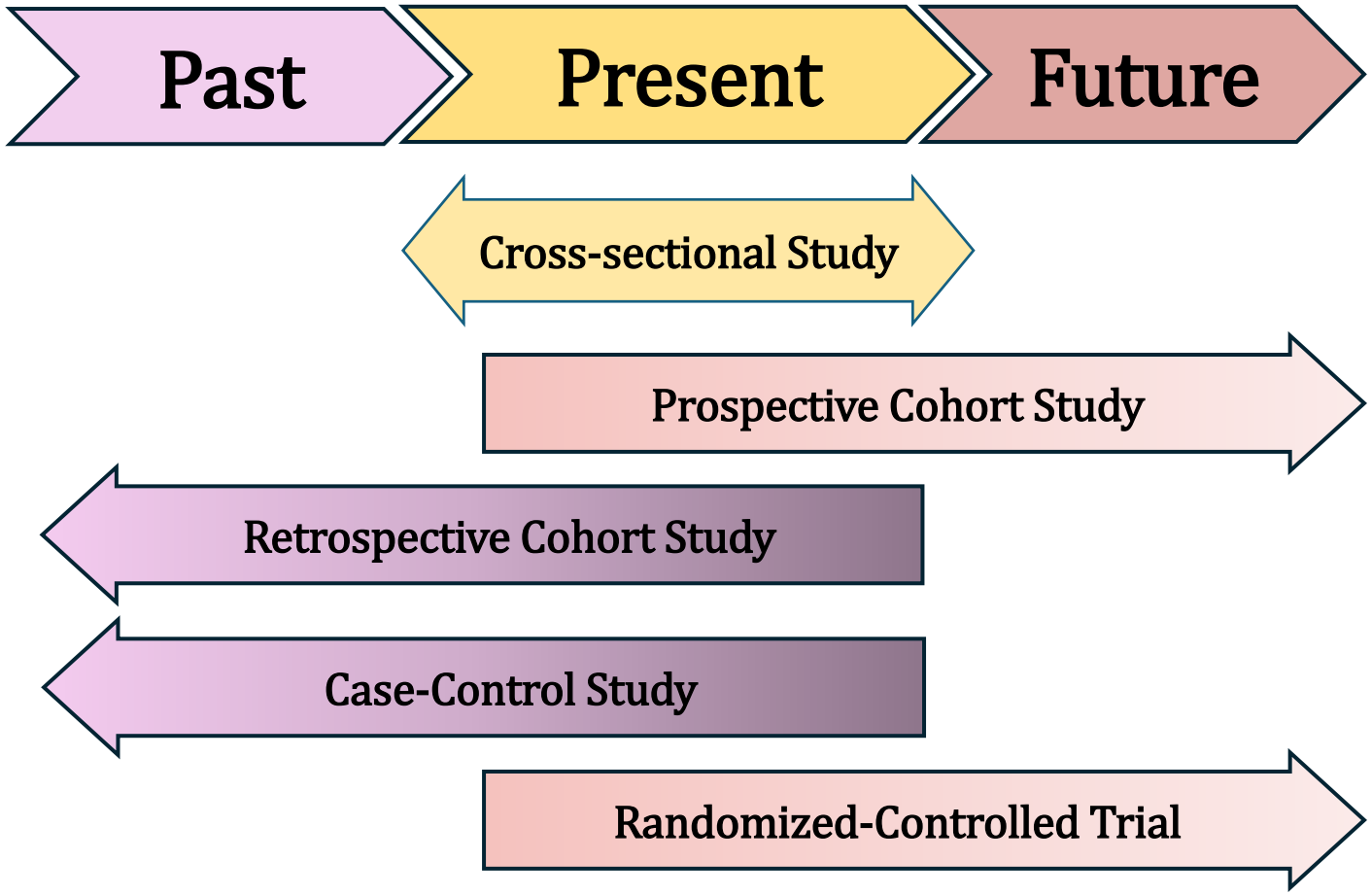
AI models can handle both structured and unstructured data

- unstructured data requires more preprocessing and feature extraction for training
- meaning of unstructured data can be highly context-dependent and ambiguous
- **i**ncorporating domain knowledge is required for interpreting unstructured data and building effective models

Data Design

Data Collection

- Does the data cover the necessary variables and time periods;
  - target population
  - outcomes
  - predictors
  - exposures (if any)
  - time frame



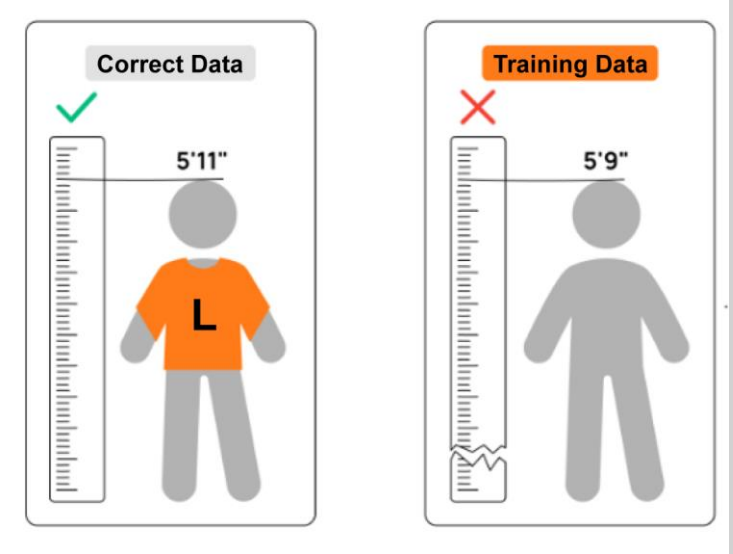
Use the easiest/most economical data design that can solve problem

Example: A cross-sectional data design suffices if one can solve the problem using the *prevalence* of the outcome in the exposed and unexposed groups.

Are there any potential biases introduced by the data design?

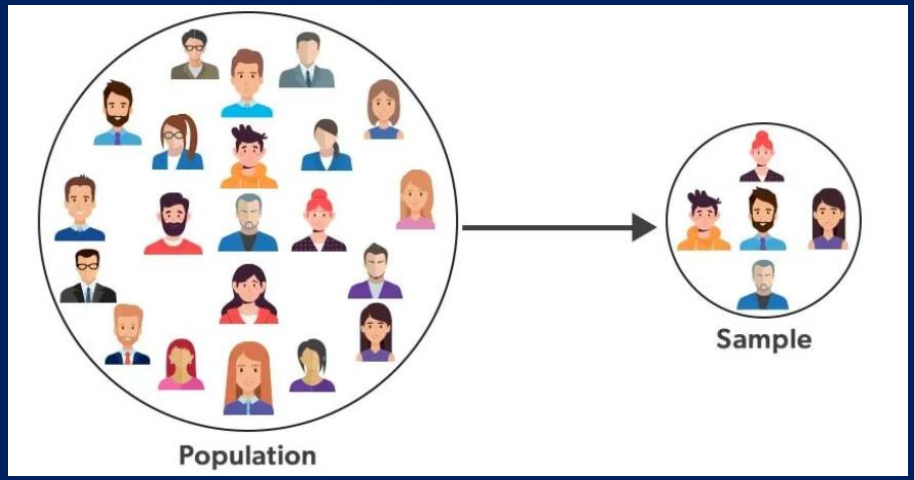
In the context of data design, **bias** are systematic deficiencies in the design that lead to errors in the produced models and corresponding problem domain solutions.

Data Collection



Measurement bias refers to the systematic error between the measured and the actual values of a variable.

Sampling bias occurs when some characteristic in the sample is over- or underrepresented relative to the population.



Confounding bias occurs when the association between an exposure and an outcome is distorted by the presence of a third variable, known as a confounder.

Slido Poll

Which of the following statements is NOT true about unstructured data used in AI models?

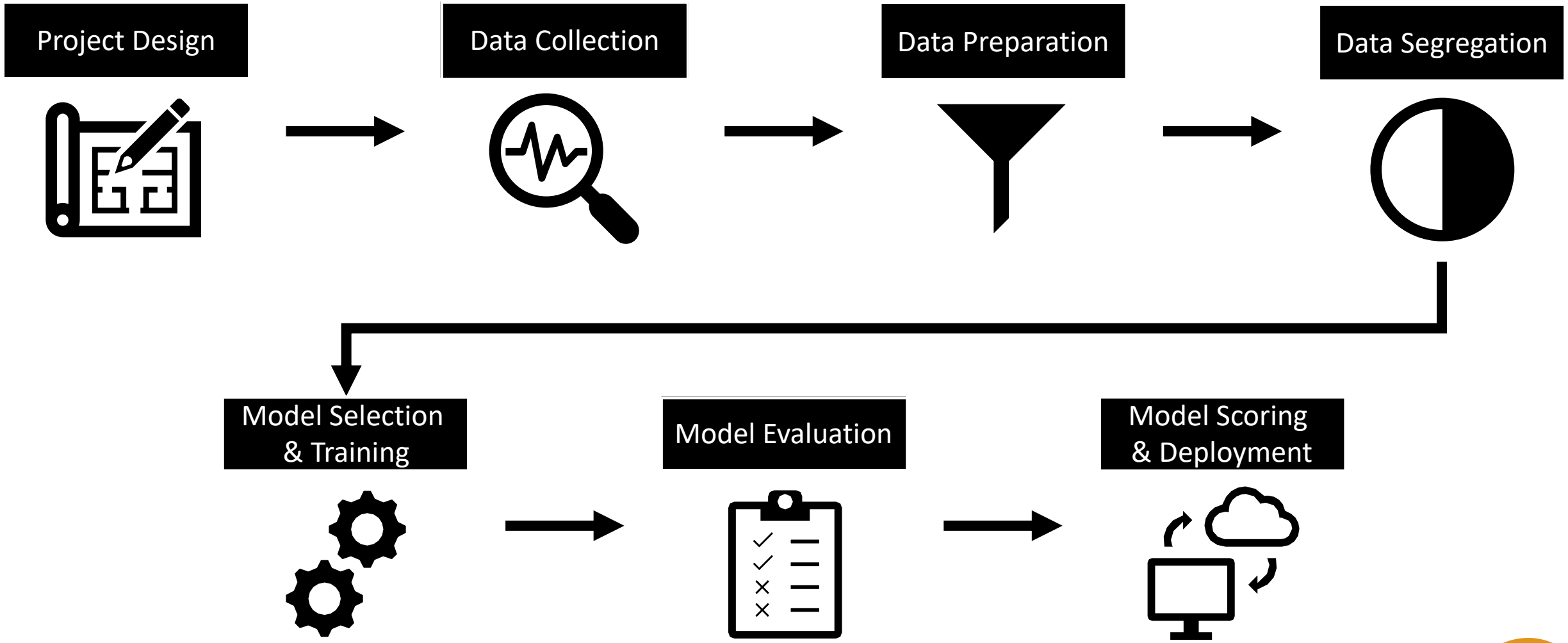
a) Unstructured data often requires extensive preprocessing and feature extraction for effective use in training AI models

b) Unstructured data is highly context-dependent and can be ambiguous, necessitating the incorporation of domain knowledge for proper interpretation

c) Unstructured data is straightforward and requires minimal preprocessing before it can be incorporated into AI models

d) Unstructured data includes forms such as text, images, and audio, which do not follow a predefined data model

AI development pipeline



Step 3. Data Preparation

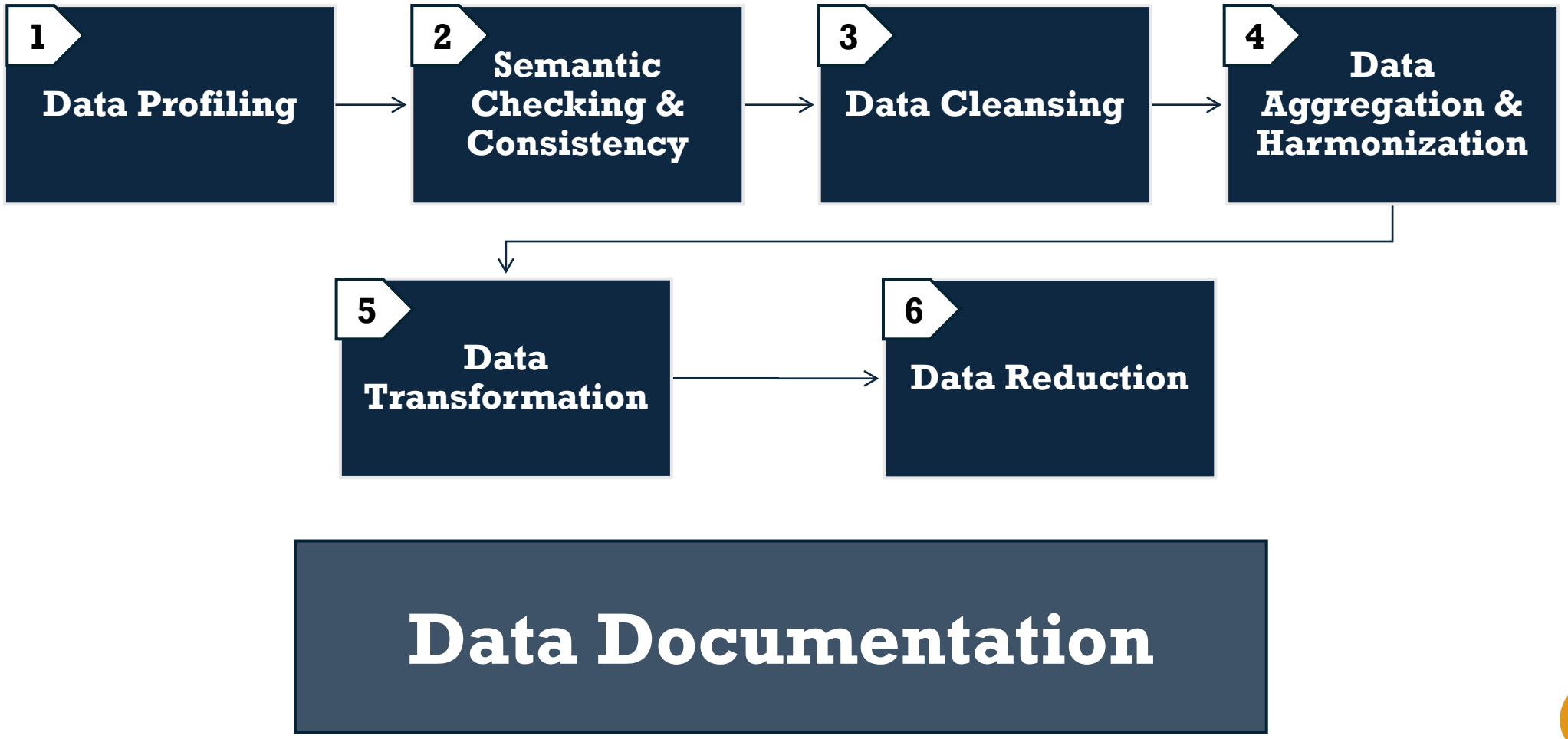
Definition The process of manipulating and organizing data prior to analysis

- AI data design relies heavily on the existence of digital data repositories that are created independently of the problem solving intent at hand
- Data often comes from various sources with different levels of accuracy and reliability
- Data can contain noise, outliers, and irrelevant information that can obscure meaningful patterns and mislead AI algorithms

Data Preparation

- To effectively aggregate or harmonize datasets for AI:
  - datasets must share compatible formats
  - uniform labeling across datasets is necessary to maintain clarity and utility
  - balanced datasets helps mitigate bias and enhances the robustness of AI models

Data Preparation Steps



Data Preparation

Data Profiling

Definition

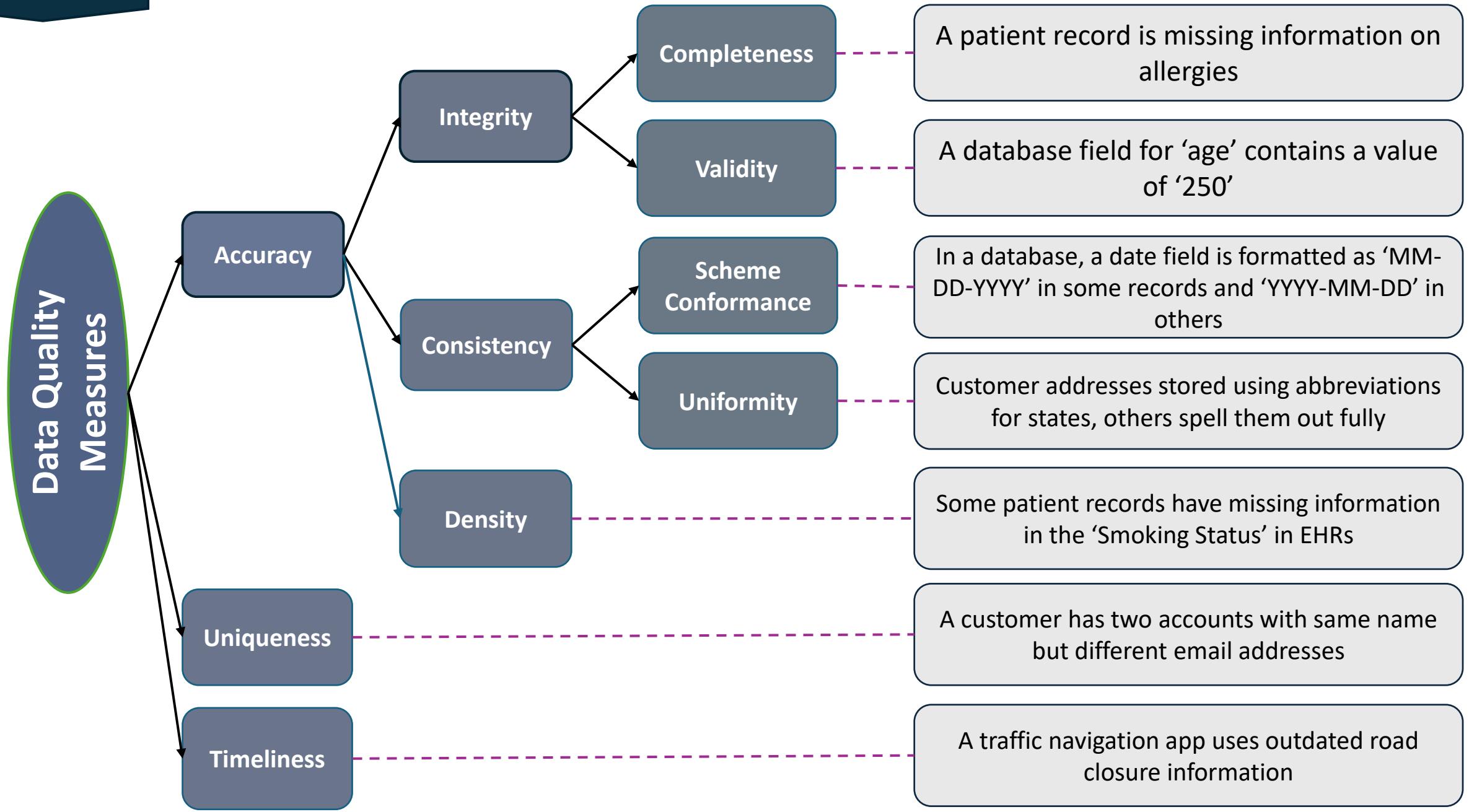
The process of carefully examining a dataset to understand its structure, content, and quality.

- Identify data types (e.g., text, numbers, dates), patterns, and relationships between different elements
- Uncover data quality issues like missing values, inconsistencies, errors, and outliers
- Explore the actual values within the data, identify common patterns, and detect anomalies

Data Preparation

- **Data quality** is the degree to which a dataset meets a user's requirements and is fit for the purpose it is intended.

Definition


Data Quality Measures
Accuracy
Integrity
Completeness
A patient record is missing information on allergies
Validity
A database field for ‘age’ contains a value of ‘250’
Consistency
Scheme Conformance
In a database, a date field is formatted as ‘MM-DD-YYYY’ in some records and ‘YYYY-MM-DD’ in others
Uniformity
Customer addresses stored using abbreviations for states, others spell them out fully
Density
Some patient records have missing information in the ‘Smoking Status’ in EHRs
Uniqueness
A customer has two accounts with same name but different email addresses
Timeliness
A traffic navigation app uses outdated road closure information

Data Preparation Steps

2 **Semantic Checking & Consistency**

Data Preparation

Data Documentation

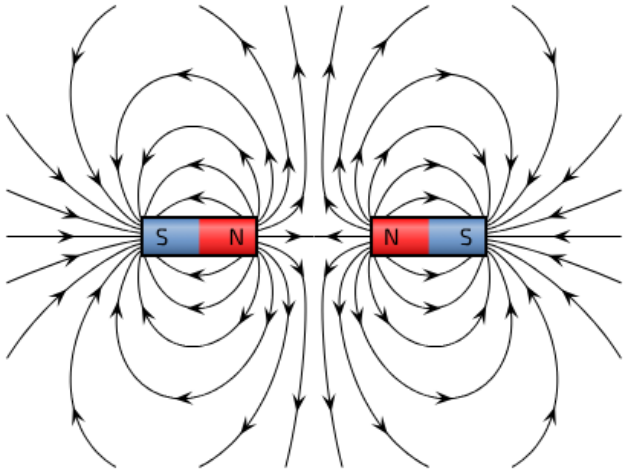
Semantic Checking and Consistency

Definition

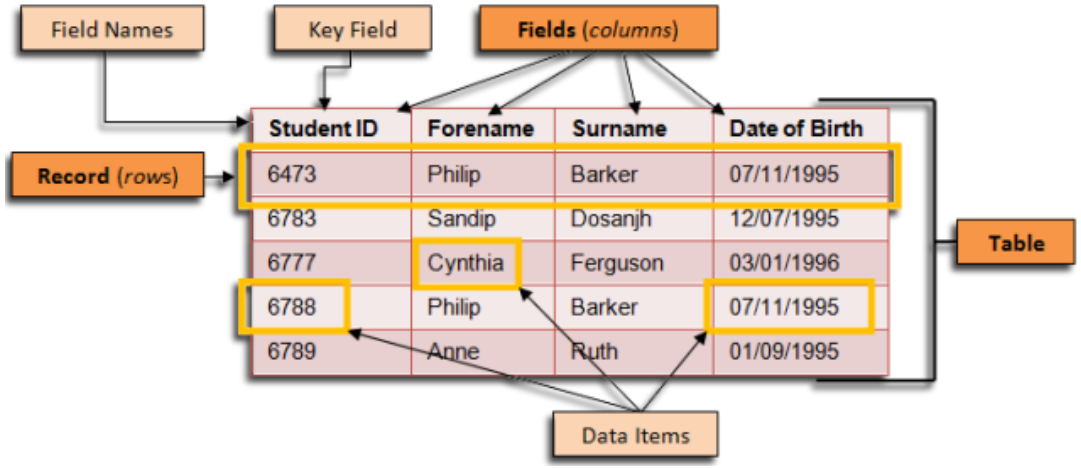
The process of identifying and resolving semantic inconsistencies by aligning data element meanings and interpretations with a common standard or ontology, ensuring uniform understanding across different sources and contexts.

An **ontology** is a representation of a domain of concepts, comprising concepts and their formal names, attributes of the concepts, and relationships among the concepts.

Meaning of **'Field'**



Physics



Student ID	Forename	Surname	Date of Birth
6473	Philip	Barker	07/11/1995
6783	Sandip	Dosanjh	12/07/1995
6777	Cynthia	Ferguson	03/01/1996
6788	Philip	Barker	07/11/1995
6789	Anne	Ruth	01/09/1995

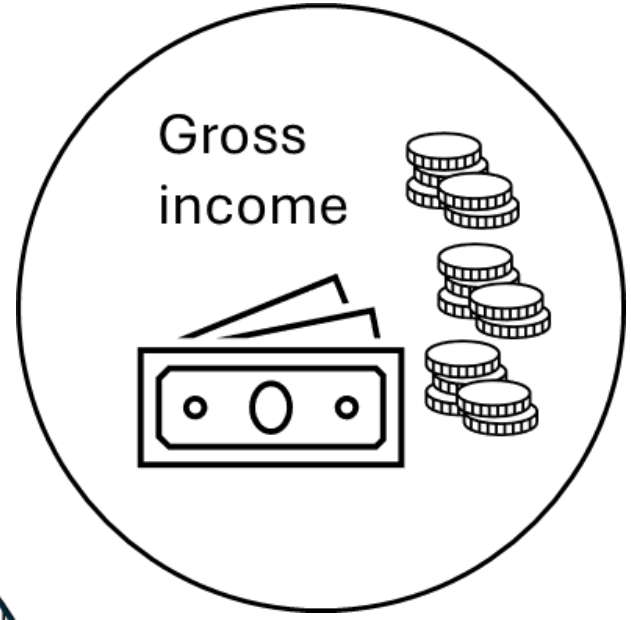
Computer Science

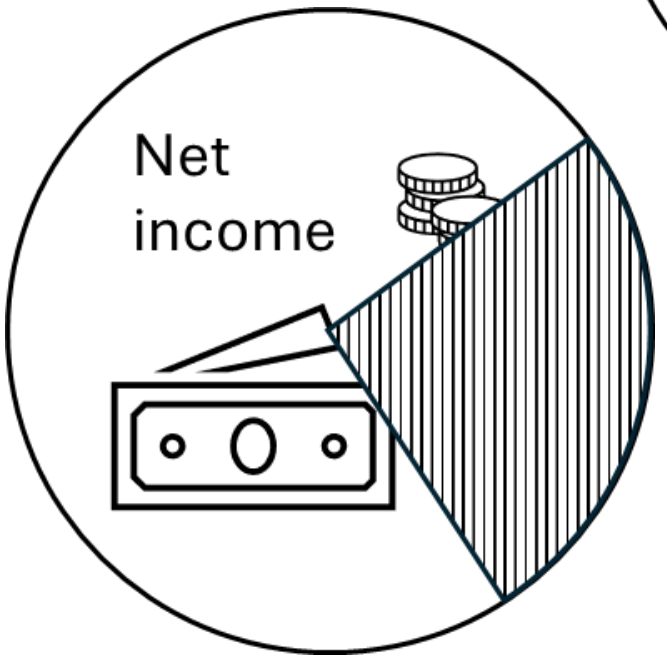
Agriculture

The same term may have different meanings in different contexts, leading to misinterpretations

Without a common, standardized vocabulary or ontology to define the meaning of data elements, different systems and individuals might use different terms and classifications, leading to semantic inconsistencies

'Income' variable in different datasets





Comparing "income" across these datasets without considering these different definitions would be misleading

Data Preparation Steps

3

Data Cleansing

Data Preparation

Data Documentation

Data Cleansing

Definition

The process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

❖ Also referred as data cleaning, data scrubbing, data reconciliation

Data cleansing steps:

- Clean and rectify errors in data such as incompleteness, incorrectness, inaccuracy, or irrelevancy by replacing, modifying, or deleting them
- Document error instances and error types
- Measure and verify to see whether the cleansing meets the user's specified tolerance limits in terms of cleanliness

Missing Data

- A data element is **missing**, if no data value is stored for a variable
- *Structural missingness*, where the data was not supposed to be collected in the first place.
  - Example: 'Number of previous pregnancies' variable missing for male patients
- *Informative missingness*, the reason for data being missing is related to the value that would have been observed
  - Example: Patients with the most severe depression might be less likely to complete questionnaires or attend appointments where this data is collected
  - Informative missingness can significantly impact the validity of results

EXAMPLE

Study goal: Test the effectiveness of a new medication for chronic pain.
Missingness: Patients experiencing severe side effects discontinued participation and did not provide reports on their pain levels.
Potential Bias: If the analysis is limited to data from patients who completed the study, a biased understanding of the medication's effectiveness might be generated, as data from those with potentially worse experiences would be excluded.

Handling Missing Data

Before choosing a handling strategy, carefully consider the reasons for missing data

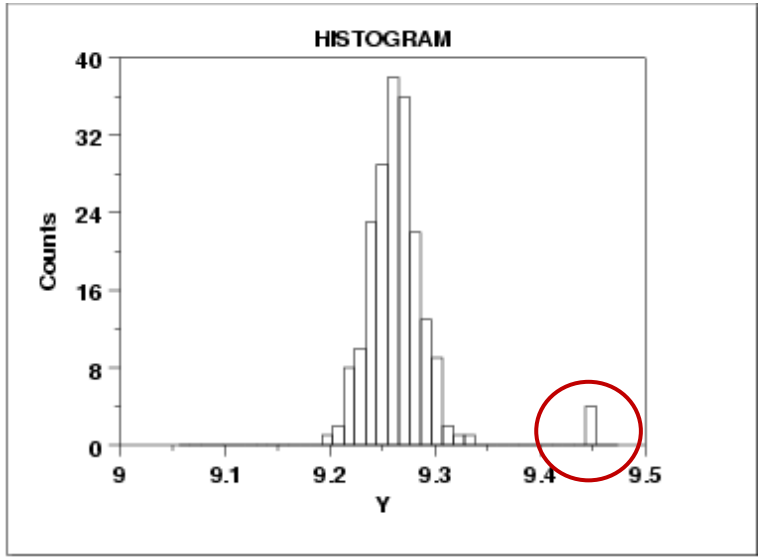
Missing Data Mechanism	Description	Example	Handling Strategies
Missing completely at random (MCAR)	The reason that the data are missing are not related to the missing data itself and every observation has an equal chance of being missing	A lab test did not produce a result or an automated blood pressure cuff does not record results periodically	Listwise deletion, pairwise deletion, mean/median/mode imputation
Missing at random (MAR)	There is a relationship between one or more observed pieces of data that is associated with the missing data	A patient has no HgA1c results (because they are not diabetic)	Multiple imputation, Maximum Likelihood Estimation
Missing not at random (MNAR)	The probability that data is missing depends on the data itself	A patient does not report a quality of life score (because they are depressed)	Model missingness, analyze different pattern of missingness separately, conduct sensitivity analysis

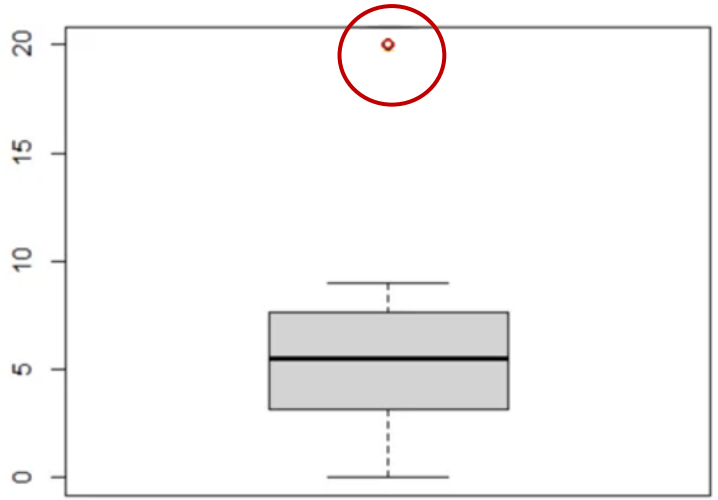
The amount and type of missingness in the dataset, and how missingness were handled should be documented

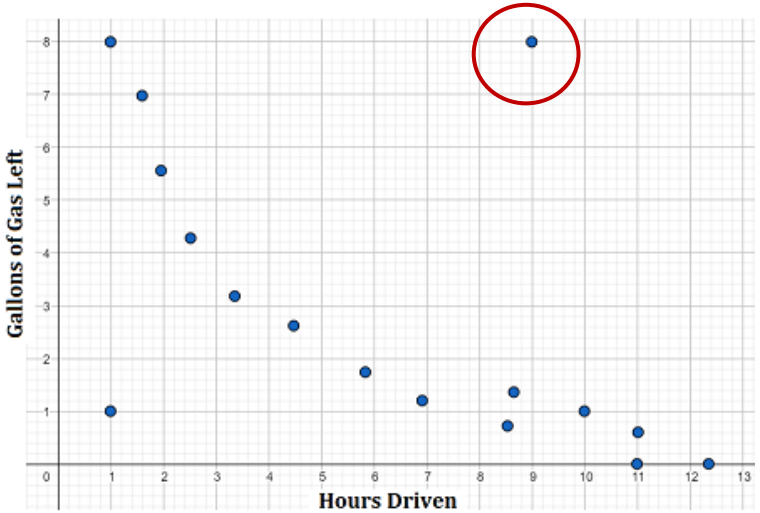
Handling Outliers

Definition Data do not conform to the overall data distribution.

Outliers can be detected using data visualization techniques







Data Preparation

Outliers can be trimmed, transformed, or treat as missing data.

It should be documented by summarizing the outlier detection and handling procedures.

Slido Poll

Which of the following best describes the element of 'consistency' in the context of data quality?

a) Ensuring that data values are unique across datasets

b) Maintaining that data conforms to specific schema requirements and is uniformly presented across different systems

c) Guaranteeing data records are up-to-date and reflect the current situation or the most recent data

d) Verifying that all required data is present and that all data meets defined validity constraints

Data Preparation Steps

4 **Data Aggregation & Harmonization**

Data Preparation

Data Documentation

Data Harmonization and Aggregation

Definition

Data Aggregation: The process of gathering information from databases in order to get readily combined datasets for processing.

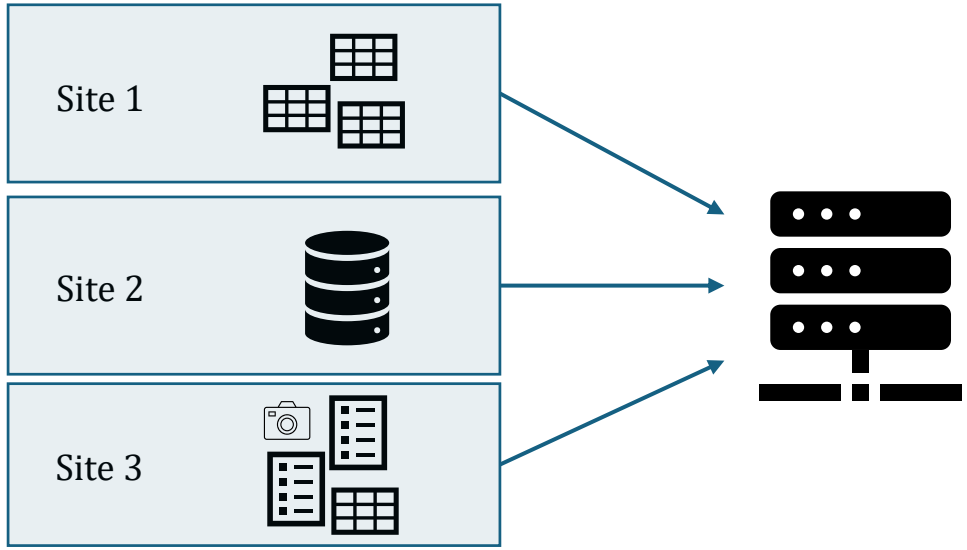
Data Harmonization: The process of bringing data together from different sources and transforming it into a common format that allows for meaningful comparison and analysis.

Definition

Data Preparation

Harmonization works by mapping data elements to standardized ontologies, ensuring

- bringing same variables and values to common computer readable codes
- accurate search for data elements
- interoperability across groups and sites



What is Interoperability?

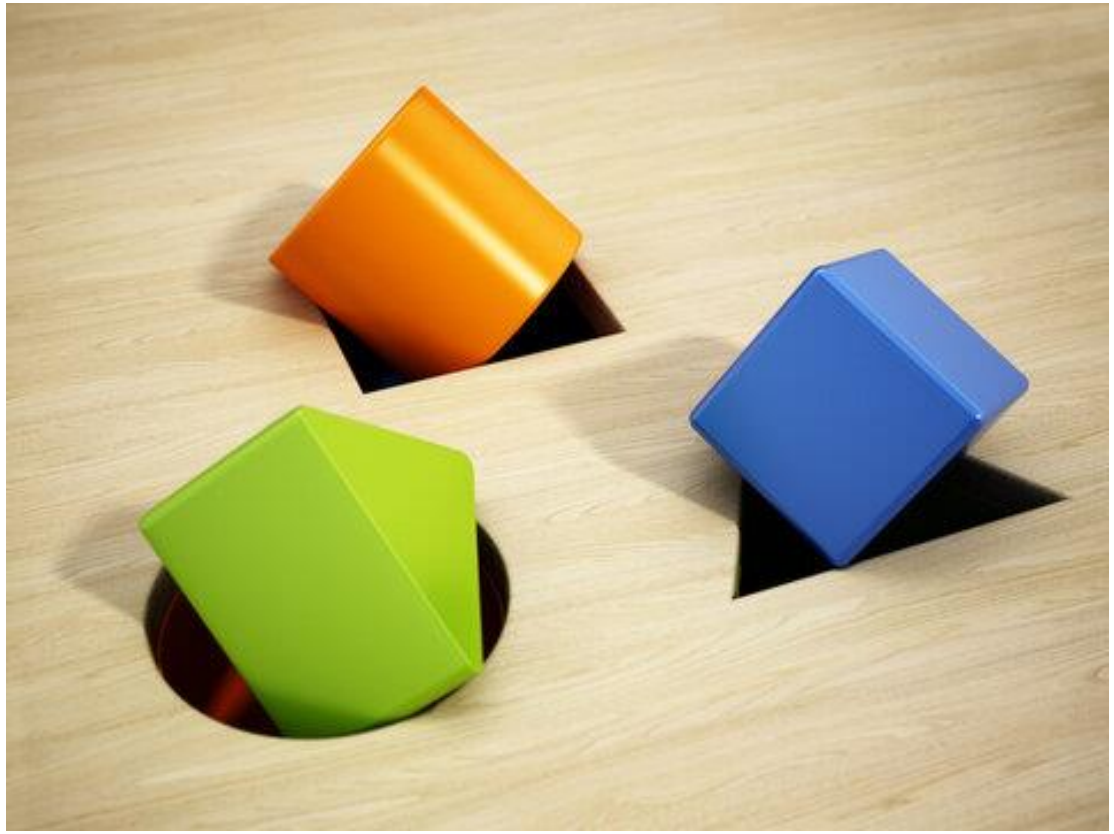
Definition

The ability of applications and systems to securely and automatically exchange data irrespective of geographical, political, or organizational boundaries.

- Interoperability facilitates data sharing and collaboration among researchers from different institutions and disciplines
- By combining data from multiple sources, researchers can increase the sample size of their studies

Data Harmonization Challenges

- Variability in data implementation and definitions across different sources including
  - different coding systems
  - varying levels of granularity
  - conflicting definitions
- Lack of widely adopted standards, incompatibilities between systems
- Inconsistent, incomplete, or inaccurate data



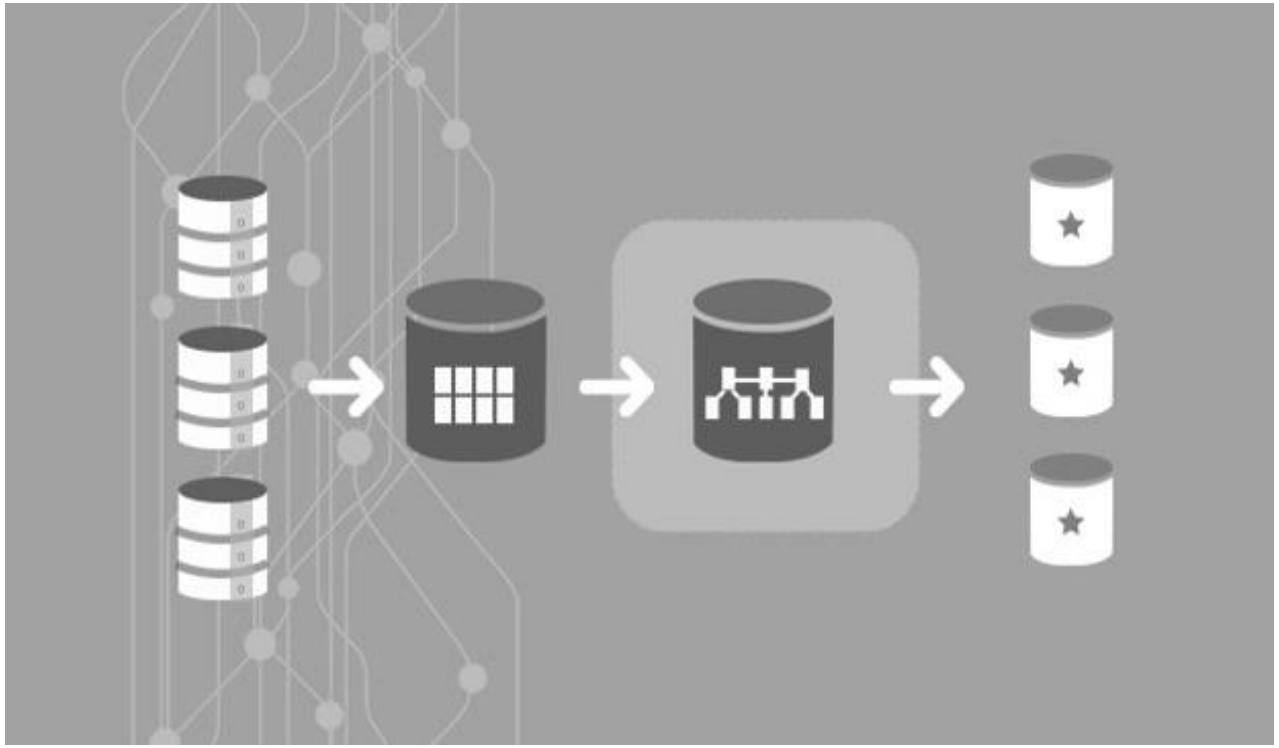
Data Harmonization Steps

1. Identify a data model (common data elements)

2. Map Source Data to CDEs

3. Transform Data to a Common Format

4. Validate Data and Create a Harmonized Dataset



Data harmonization creates a unified view of data by addressing

- inconsistencies in data structure
- format
- semantics

1 Identify a Data Model (Common Data Elements)

- Determine the key data elements relevant to your analysis or research question (age, sex, income)
- Select a data model that are widely recognized and used across different settings and systems (ScHARe CDEs https://cde.nlm.nih.gov/)

Definition

Common data elements (CDEs) are standardized, precisely defined questions paired with a set of specific allowable responses, used systematically across different sites, studies, or clinical trials to ensure consistent data collection.

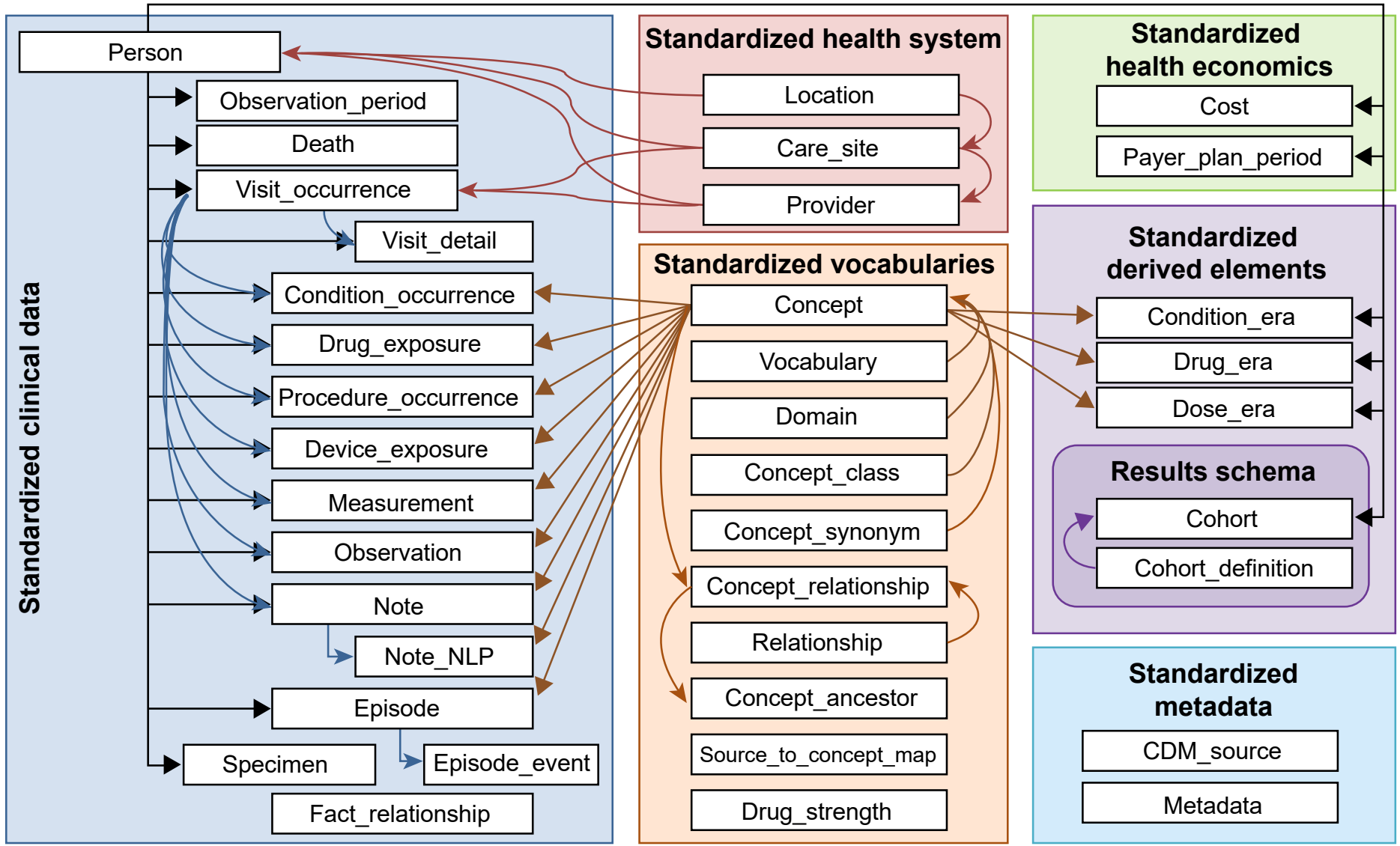
- CDEs ensure consistency across different studies, making it easier to combine and analyze data from multiple sources.
- By using CDEs researchers can create larger and more diverse datasets that are critical for AI model development

Data Preparation

Definition A data model is an abstract model that organizes elements of data and standardizes how they relate to one another and to the properties of real-world entities.

Example: Data Models

OMOP Common Data Model



Example: Common Data Elements

14. English proficiency

We are interested in your own opinion of how well you speak English. Would you say you speak English:

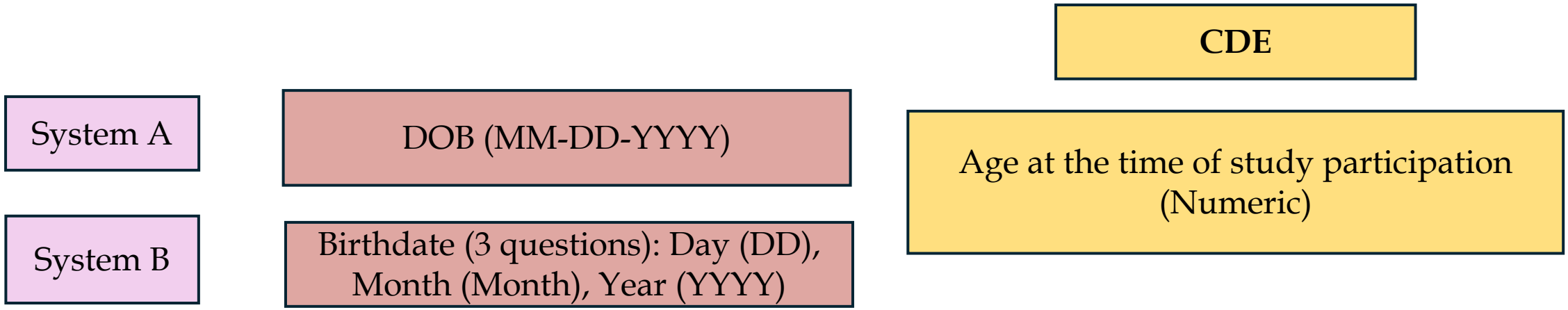
- ☐ Very well
- ☐ Well
- ☐ Not well
- ☐ Not at all
- ☐ Refused
- ☐ Don't Know

English Proficiency	Value List	Very Well |	A subjective response of strong agreement or ability. |	C159856 |	NCI Thesaurus |
		Well |	A subjective response of good agreement or ability. |	C182125 |	NCI Thesaurus |
		Not Well |	A subjective response of poor agreement or ability. |	C182124 |	NCI Thesaurus |
		Not at All |	A subjective answer of no agreement or ability. |	C91213 |	NCI Thesaurus |
		Refused |	Used to indicate when a respondent makes a decision to not answer a question.|	C51024 |	NCI Thesaurus |
		Don't Know	The answer is not known by the person answering.	C67142	NCI Thesaurus

ScHARe CDEs (https://cde.nlm.nih.gov/)

2 Map Source Data to CDEs

- Examine the data from each source system
- Identify how their data elements correspond to the chosen CDEs
- Map the source data fields to the CDEs, taking into account any differences in terminology, coding systems, or data formats



3 Transform Data to a Common Format

- Convert the source data into a standardized format that aligns with the CDEs by
  - translating codes from one system to another
  - standardizing units of measurement
  - using consistent data labels
  - reformatting dates and times
  - resolving inconsistencies in data definitions

		Required Transformation
System A	DOB (MM-DD-YYYY)	Extract the month, day, and year components from this format, determine the date of study participation, calculate the time difference, calculate age in years
System B	Birthdate (3 questions): Day (DD), Month (Month), Year (YYYY)	Determine the date of study participation, calculate the time difference, calculate age in years

4 Validate Data and Create a Harmonized Dataset

- Check the transformed data for accuracy, completeness, and consistency
- Identify and address any data quality issues, such as missing values, outliers, or inconsistencies
- Combine the transformed data from all sources into a single, harmonized dataset

Slido Poll

What is the primary purpose of using Common Data Elements (CDEs) in research?

a) To allow each research site to use its unique data collection methods
b) To promote variability in data collection and analysis methods
c) To ensure consistent and standardized data collection across different studies or sites
d) To completely automate the data analysis process across multiple studies

Data Preparation Steps

Data Preparation

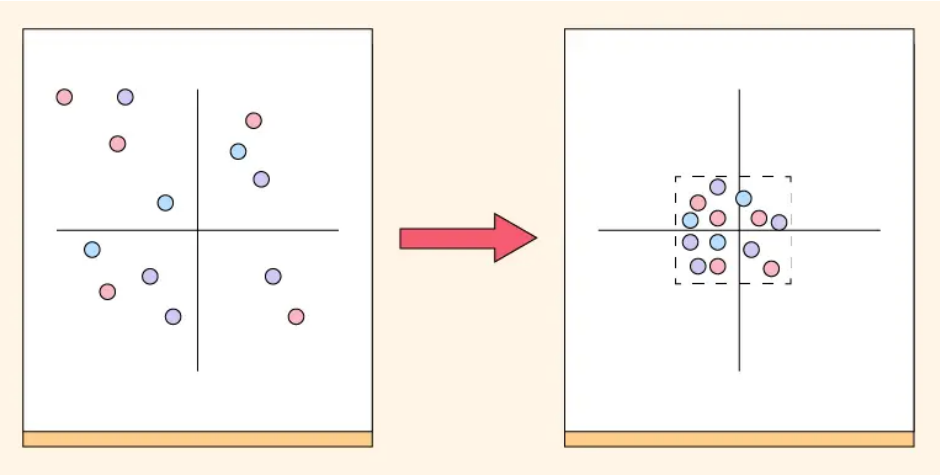
5 Data Transformation

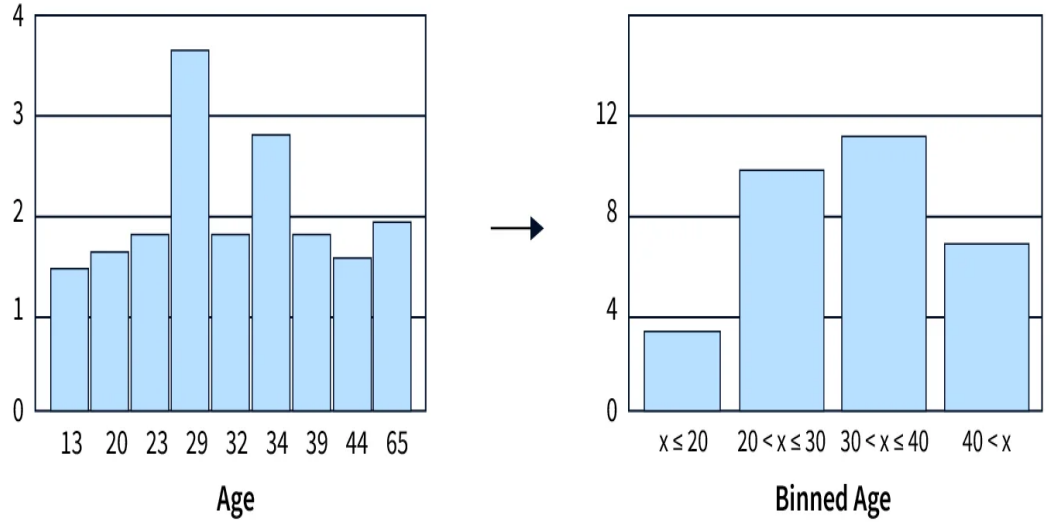
Data Documentation

Data Transformation

Definition

The process of transforming existing variables, or developing additional variables, or features, from the source data that can enable the use of a particular learning algorithms, improve its performance, or interpretability.





Scaling refers to adjusting the range of a feature to a specific scale.

Binning groups continuous data into discrete intervals or bins.

Gender	Gender
Male	1
Female	0
Male	1
Female	0

Indicator variables refers to converting categorical variables into numerical representations by creating binary (0 or 1) variables

Gender	Female	Male
Male	0	1
Female	1	0
Male	0	1
Female	1	0

One-hot encoding creates a separate binary variable for each category of a categorical feature

Data Preparation Steps

Data Preparation

6

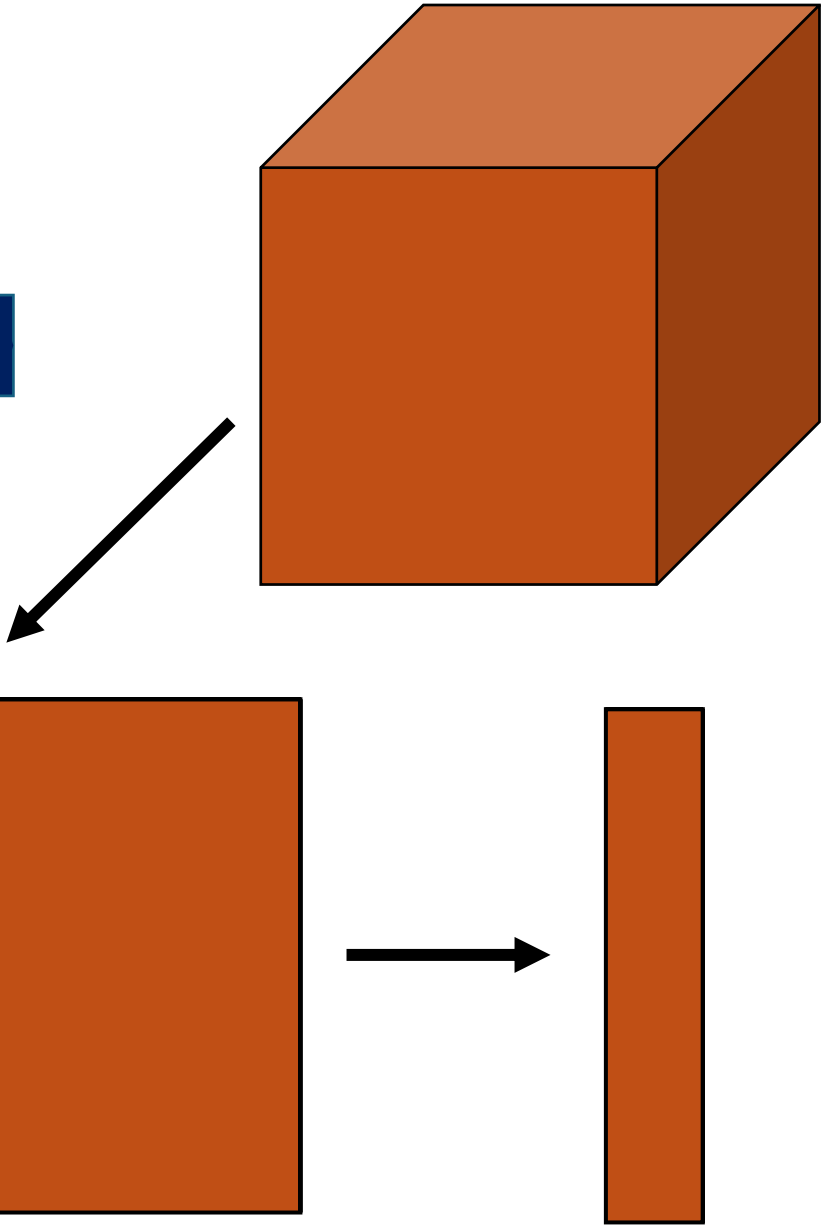
Data Reduction

Data Documentation

Data Reduction

The process of simplifying data without losing critical information, making it easier for AI algorithms to process and learn from it.

Definition

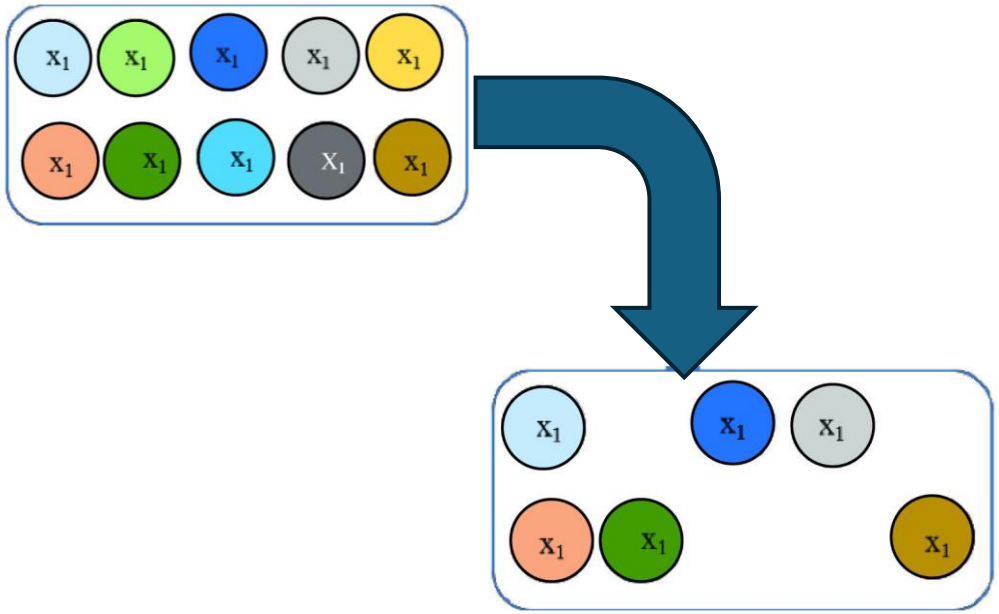


Eliminating irrelevant or redundant data can

- improve the accuracy and efficiency of AI algorithms
- make it easier to identify patterns and relationships

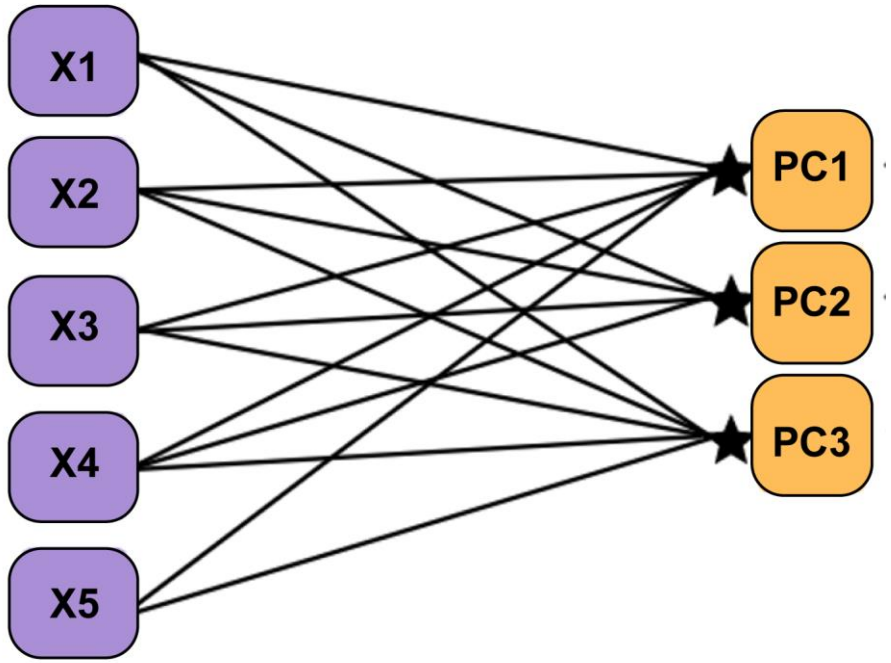
Data Reduction Methods

- Feature Reduction: Selecting the most relevant features (variables) for the AI model.
- Sample Selection: Selecting a smaller, representative subset of the data for analysis.



Data Reduction Methods

- Feature Extraction: Creating new features by combining existing ones (e.g., Principal Component Analysis)



Data Preparation Steps

Data Preparation

Data Documentation

Data Documentation

Data documentation provides a clear picture of how data was collected, processed, and transformed, and

- promotes **transparency in AI**
- enables others to **reproduce** the data preparation steps
- facilitates **collaboration** among different teams and stakeholders
- helps address **ethical concerns** related to data privacy, fairness, and accountability

Transparency ensures **trustworthiness** and **ethical** usage by enabling consumers and communities to **understand how AI** models **function** and make decisions.

Data dictionary

Definition

A **data dictionary** is an inventory of all data elements and their "meta-data" including

- the type of a data element (e.g. date, string, integer, real-value numeric)
- expectations of the data (e.g. feasible values, range),
- the terminology it is mapped to (if any)
- a brief narrative of what the data is supposed to represent

Data Preparation

Variable	Variable name	Mesaurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential order
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender
Date of survey	SURVEY	mm/dd/yyyy	01/01/2015 – 01/01/2016	When the participant completed the survey
Self-reported consumer spending	SPEND	Numeric	0-100,000,000	Self-reported average yearly expenditure
Market sentiment	SENTIMENT	Numeric	1 = negative 2 = neutral 3 = positive	Sentiment towards US domestic economy
Actual GDP growth	GDP	Numeric	-5.0-5.0	Average US yearly GDP growth

Data Documentation Best Practices

- Record where the data originated (e.g., specific databases, APIs, public repositories) including dates of acquisition and any relevant access agreements
- Calculate and document key summary statistics for each variable
- Document any data quality issues identified and the methods used to address them.
- Describe all data cleaning steps performed, such as handling missing values, correcting errors, and removing duplicates.
- Provide clear and concise definitions for all variables and their units of measurement.
- Document any rules or constraints applied to ensure data consistency across different sources or variables.
- Describe any validation checks performed to ensure the data conforms to expected formats and ranges.
- Maintain a clear record of how data is combined and transformed throughout the aggregation and harmonization process.
- If data reduction techniques are used (e.g., feature selection, sampling), document the specific methods applied.
- Describe the criteria used to select or exclude data (e.g., feature importance, relevance to the AI task).
- Track different versions of the cleaned dataset and document the changes made in each version.