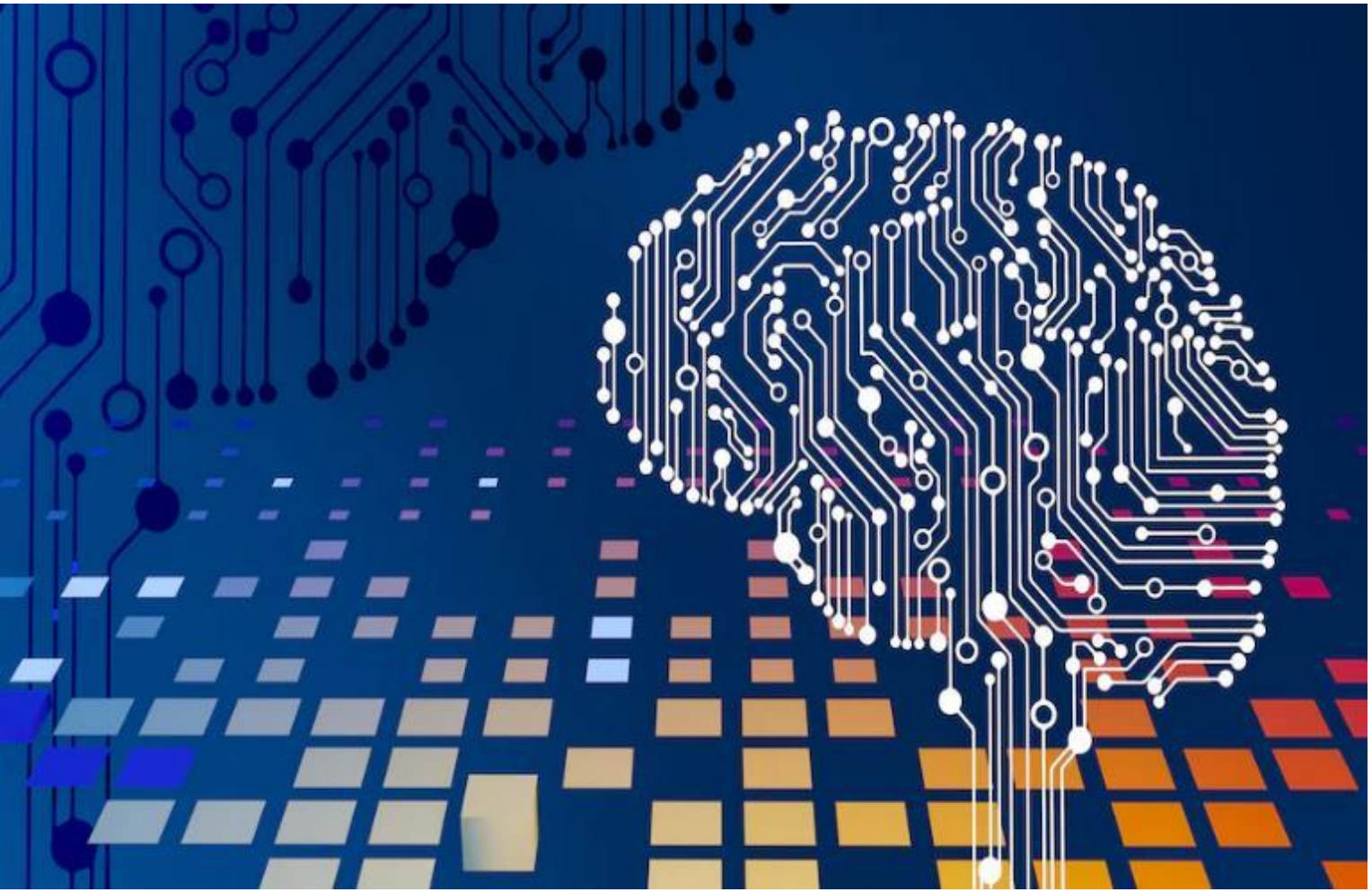


Ethical Considerations for Reviewing Research Studies Involving AI

Aline Talhouk

nature

Explore content | About the journal | Publish with us | Subscribe

nature > news > article

NEWS | 25 October 2023

AI 'breakthrough': neural net has human-like ability to generalize language

Humans are faced with an existential threat from Artificial General Intelligence

July 24, 2019

‘The Great Hack’: Cambridge Analytica is just the tip of the iceberg

It was the scandal which finally exposed the dark side of the big data economy underpinning the internet. The inside story of how one company, Cambridge Analytica, misused intimate personal Facebook data to micro-target and manipulate swing voters in the US election, is compellingly told in “The Great Hack”, a new documentary out today.

> “
> **One of the most urgent and uncomfortable questions raised in The Great Hack is: to what extent are we susceptible to such behavioural manipulation?**
>
> **Joe Westby**

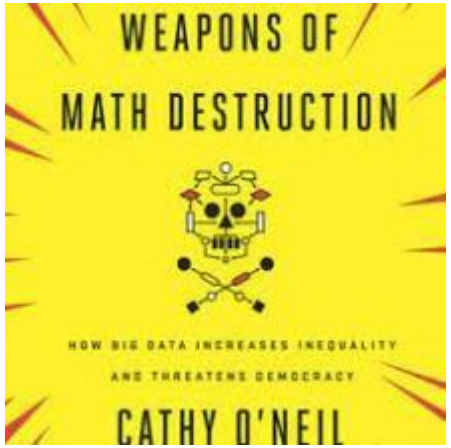
UBER / RIDE-SHARING / TRANSPO

Uber driver in first-ever deadly self-driving crash pleads guilty

/ Rafaela V
backup driv
autonomou
sentenced
probation f

By Andrew J. Hawkins, tra
covers EVs, public transpo
York Daily News and City &

Jul 31, 2023 at 12:48 PM I



Article

Amazon's sexist hiring algorithm could still be better than a human

Expecting algorithms to perform perfectly might be asking too much of ourselves

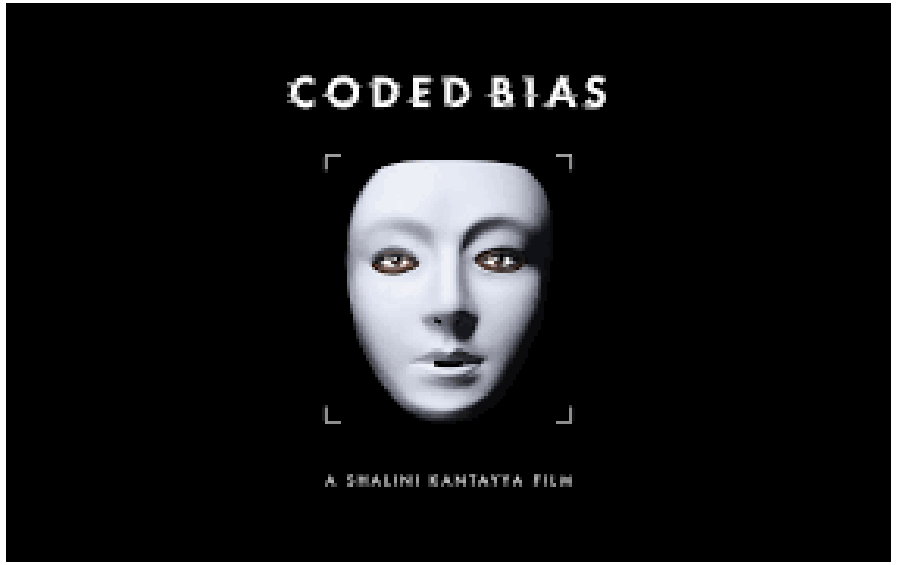
By Maude Lavanchy

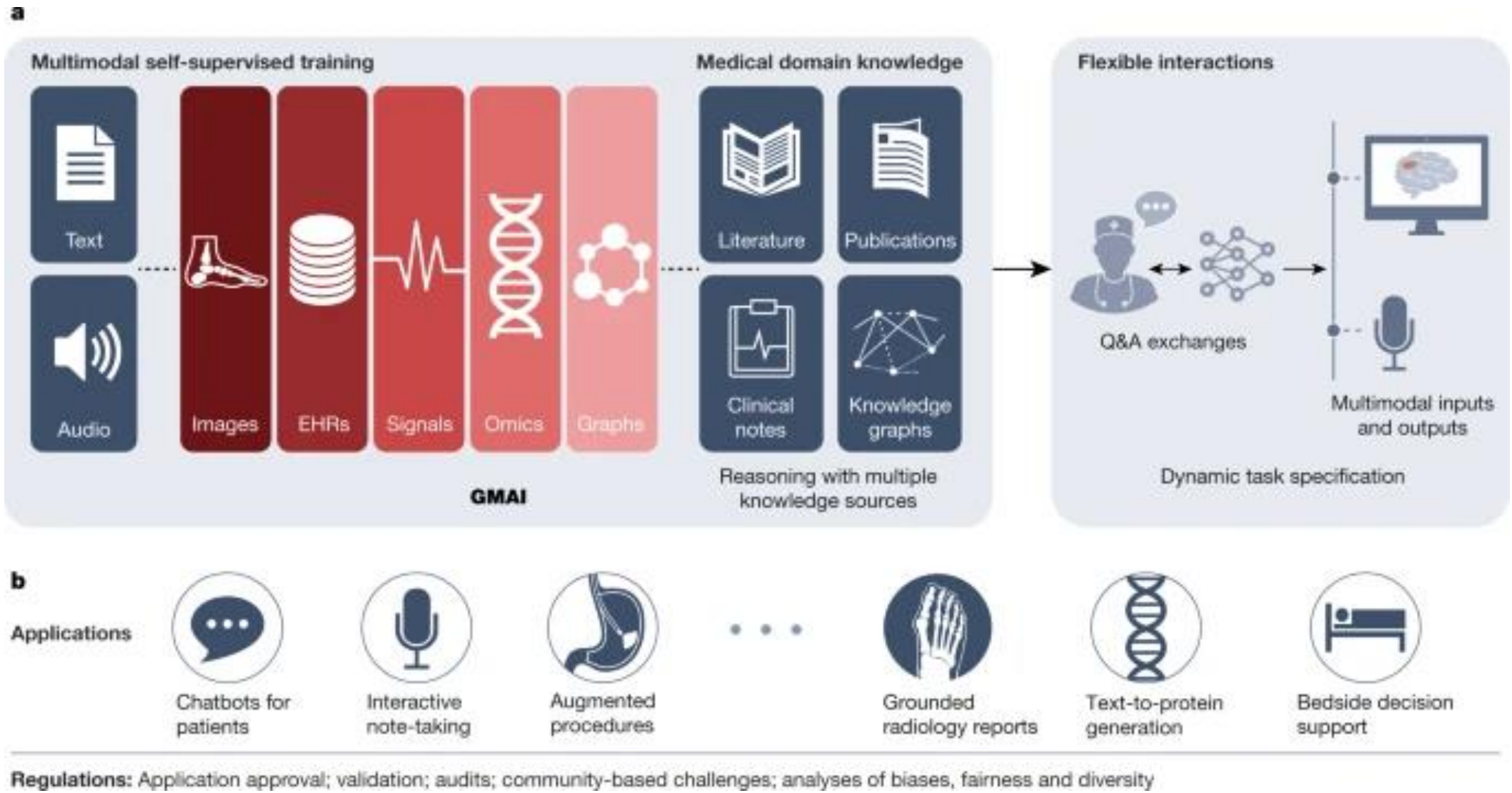
5 min

November 2018

Download full article

PRINTABLE PDF –Less than 1MB



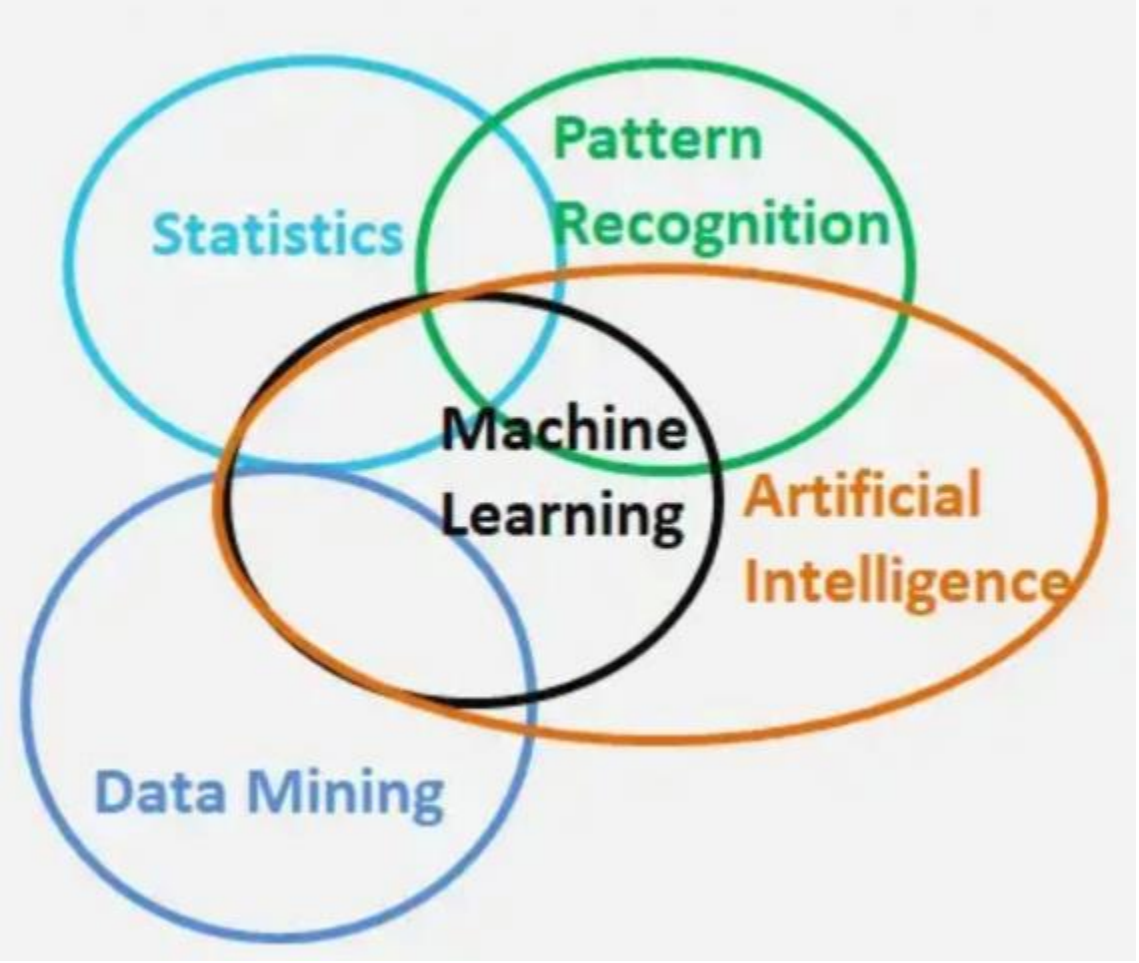


Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).

Statistics/Machine Learning/AI

Statistics:
- **Focus: Inference of parameters**
- Do models meet assumptions
- Models are often interpretable

Machine learning:
- **Focus: prediction**
- Makes few assumptions
- More flexible modeling
- Suitable for larger data
- Black box

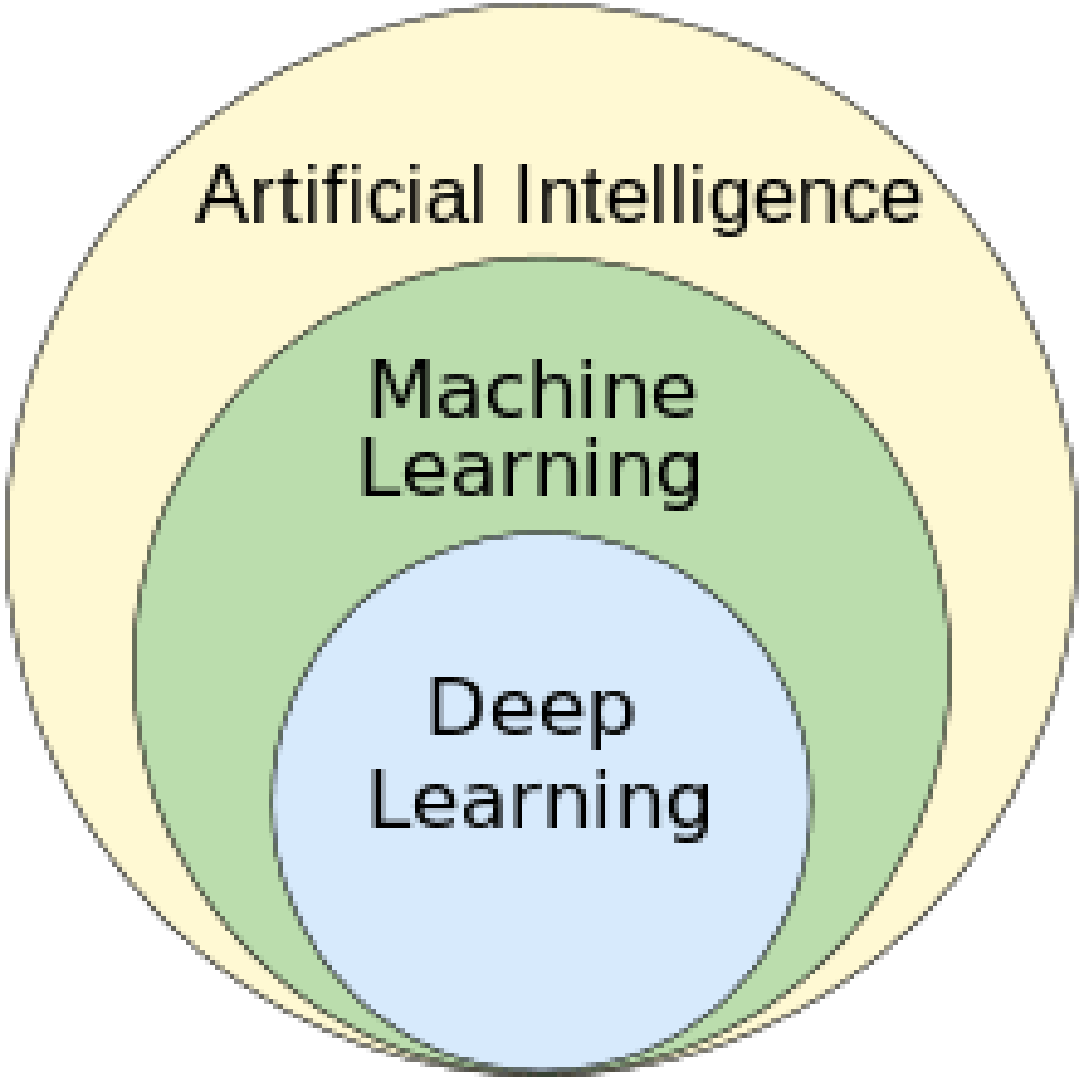


SAS institute

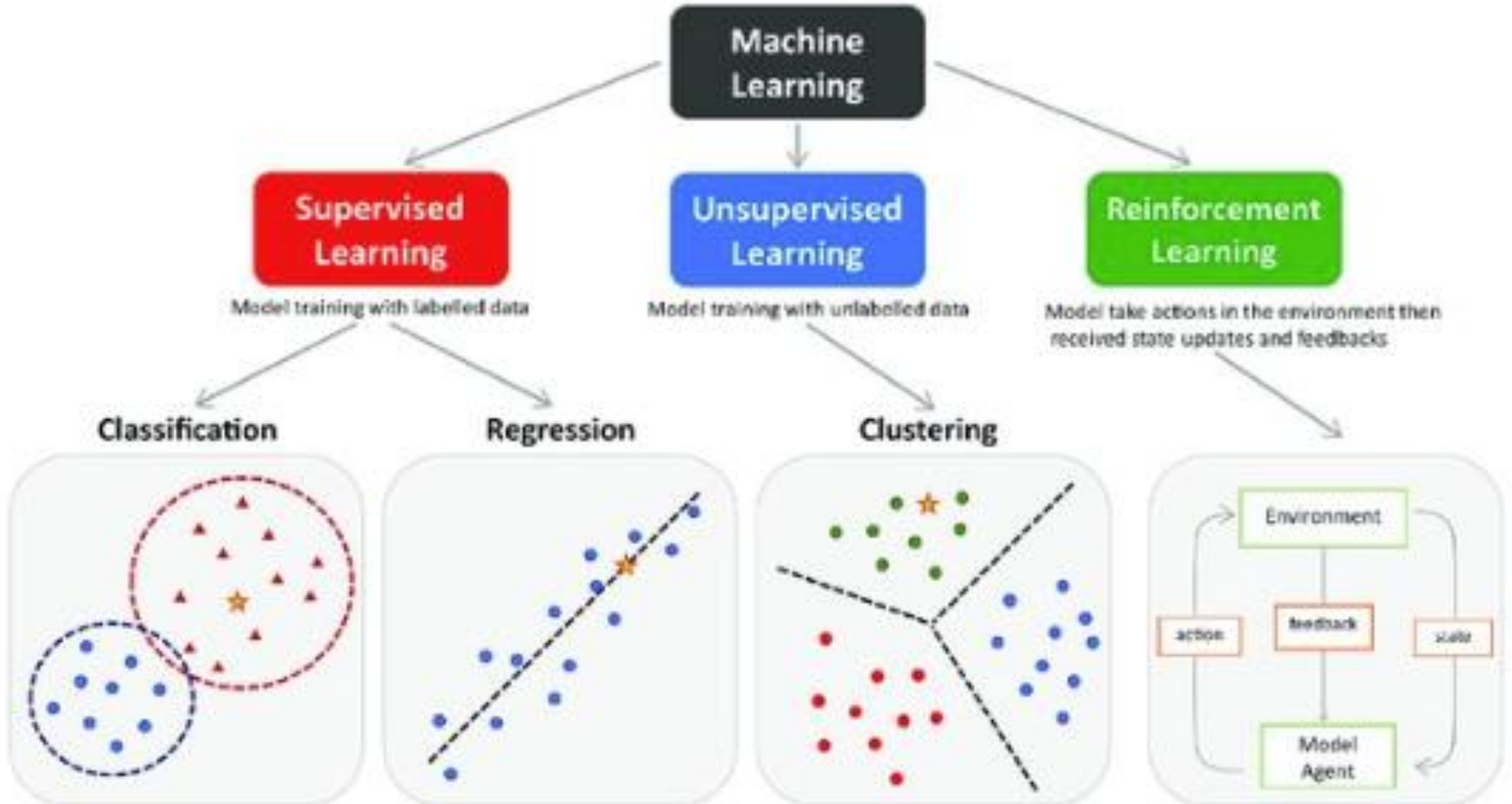
Ethical Considerations of ML vs Statistics

Added risks from statistical models

- Explainability and accountability due to the black-box nature of the algorithms
- Automated Decision-Making: lack of control
- Fairness: models can amplify and perpetuate societal biases



Types of ML



DOI: 10.3389/fphar.2021.720694

Supervised Learning

Objective: train a model to predict an outcome. In this case, the label or the target is known

Pros

- Clear evaluation metrics
- Well-established, efficient algorithms
- Continuous improvements
- Wide range of outcome prediction tasks (classification, regression)

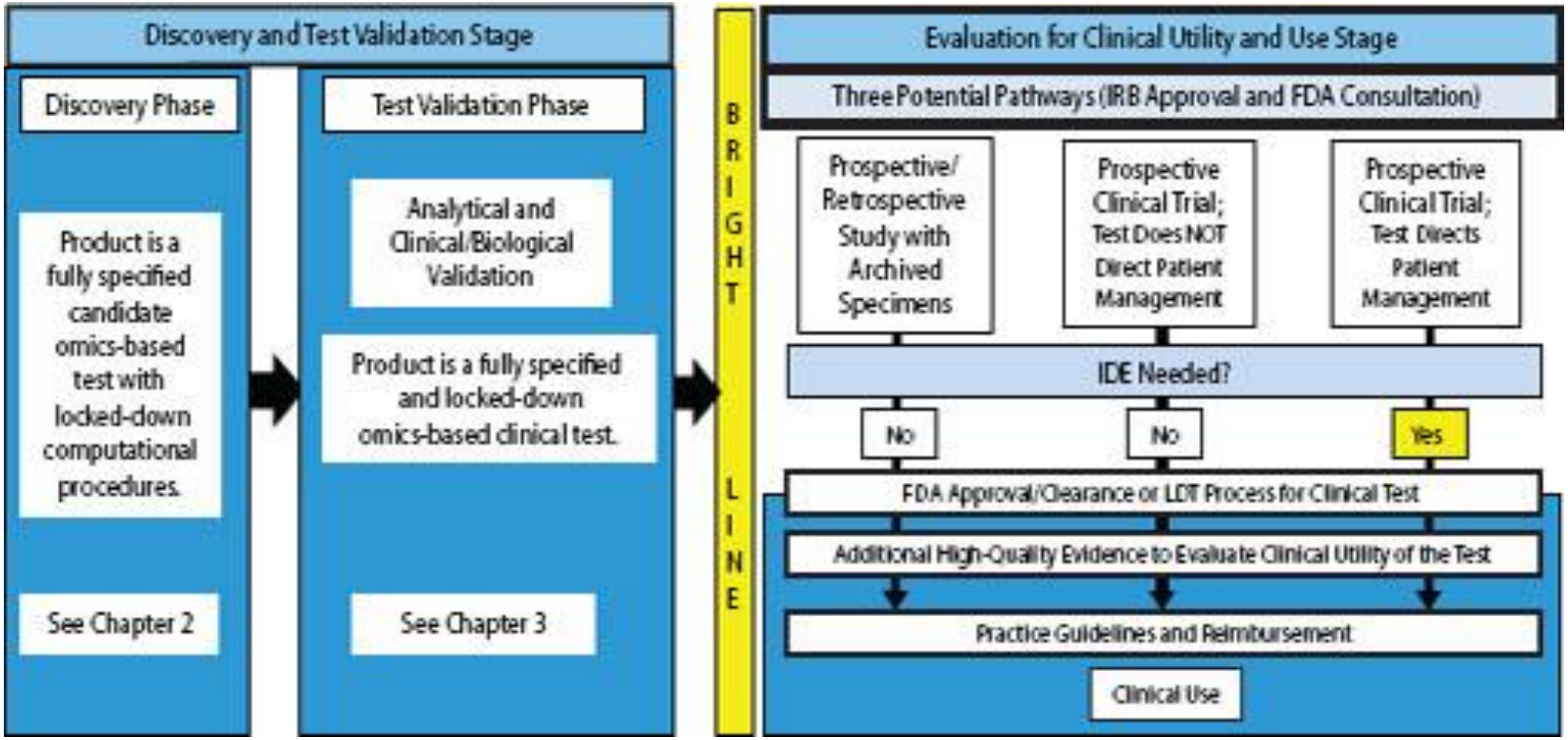
Cons

- Dependence on quality labeled data
- Not suitable for when there is no well-defined input/output relationship
- Risk of overfitting

Supervised Algorithms

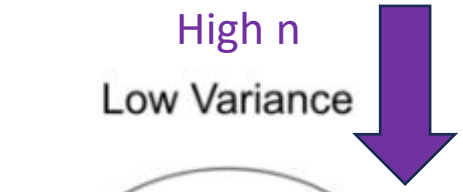
- Decision Trees
- Regression models with shrinkage parameters (Lasso, Ridge, Elastic Net)
- Support vector machines
- Ensemble learning (random forest, gradient boosting, bagging, stacking)
- Deep Neural Networks
- Ensemble methods: combining predictions from multiple algorithms

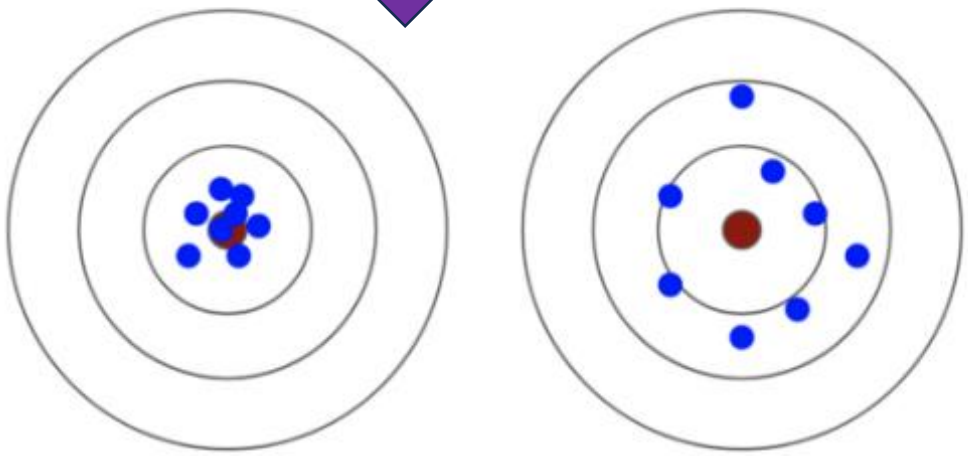
Institute of Medicine 2012



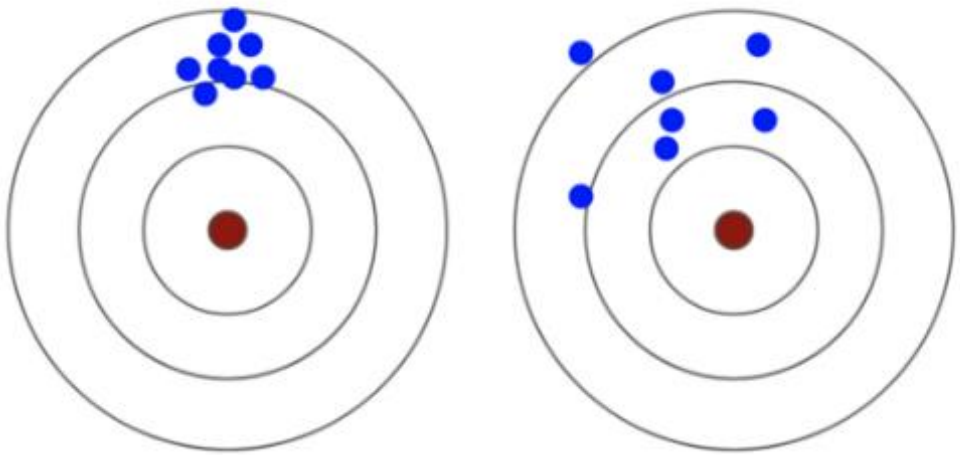
Bias - Variance Tradeoff

High n





Represent the data accurately but overfits

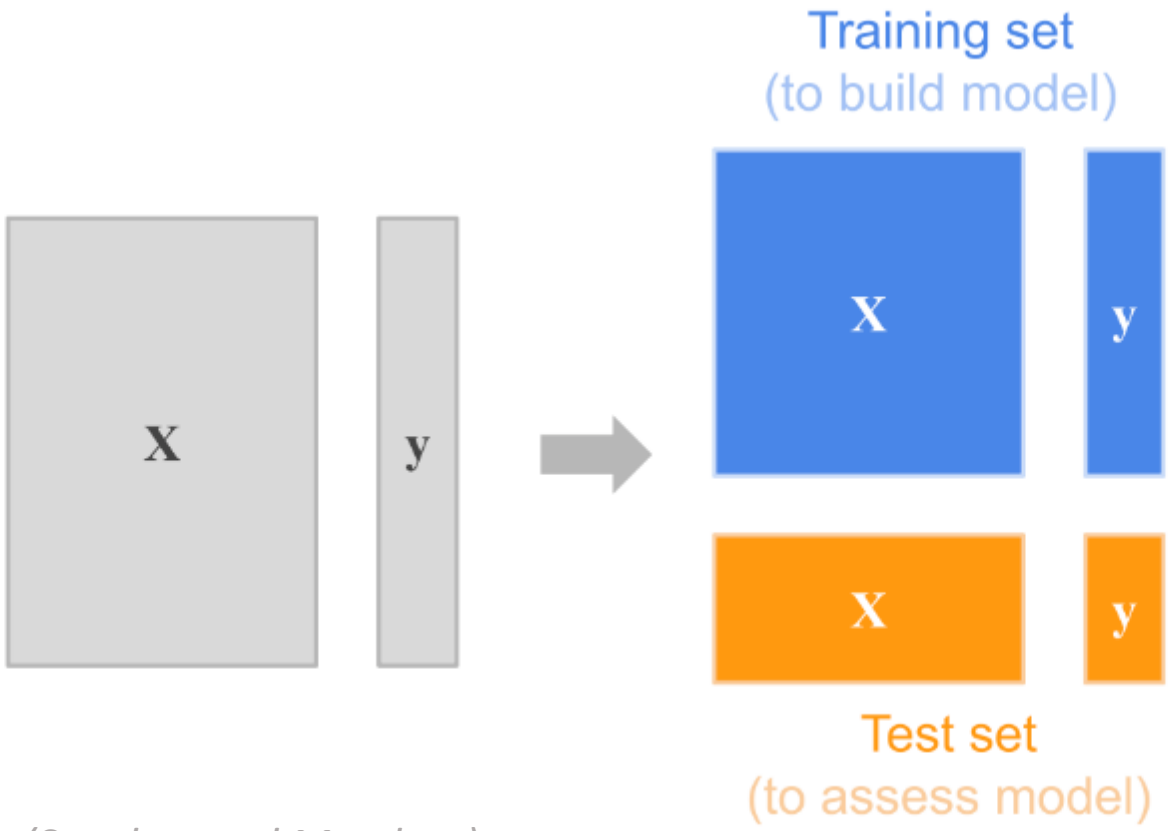


Very sure about a biased answer

Underfits the data, does not generalize

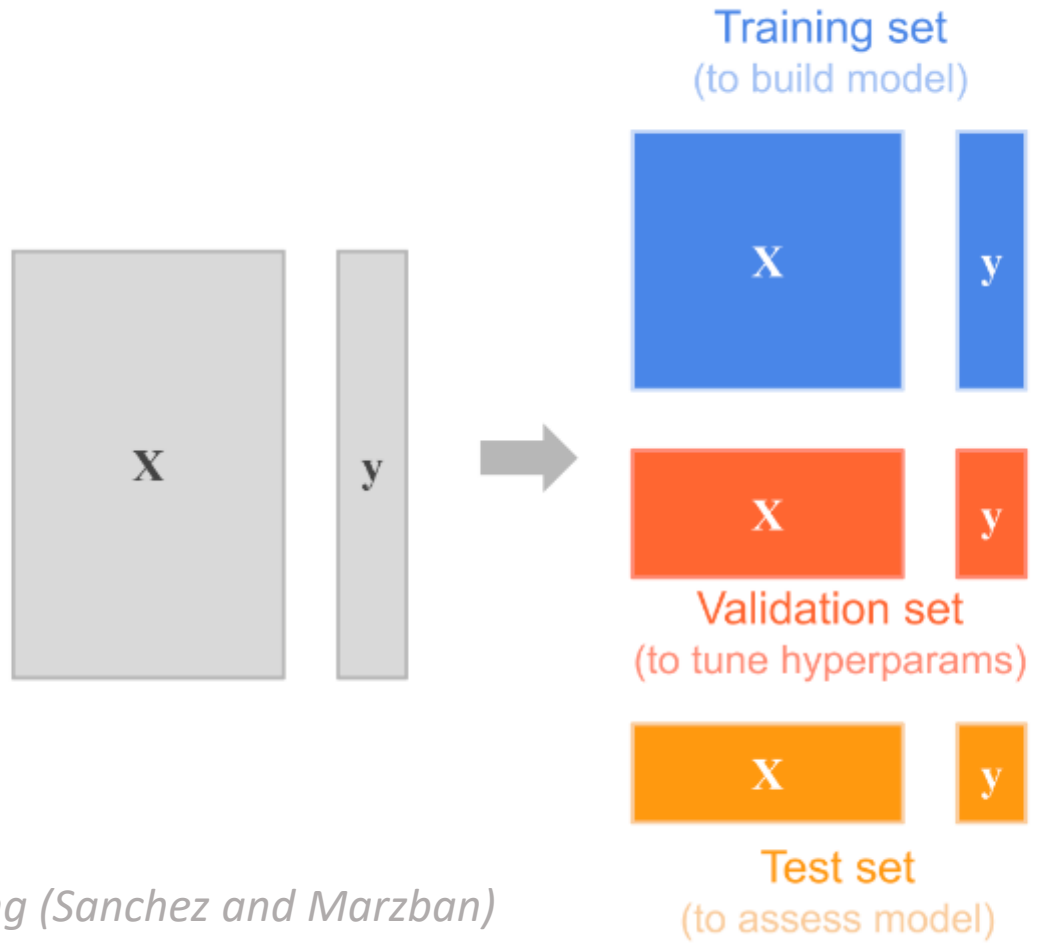
Image by Angela Shi

Hold out set



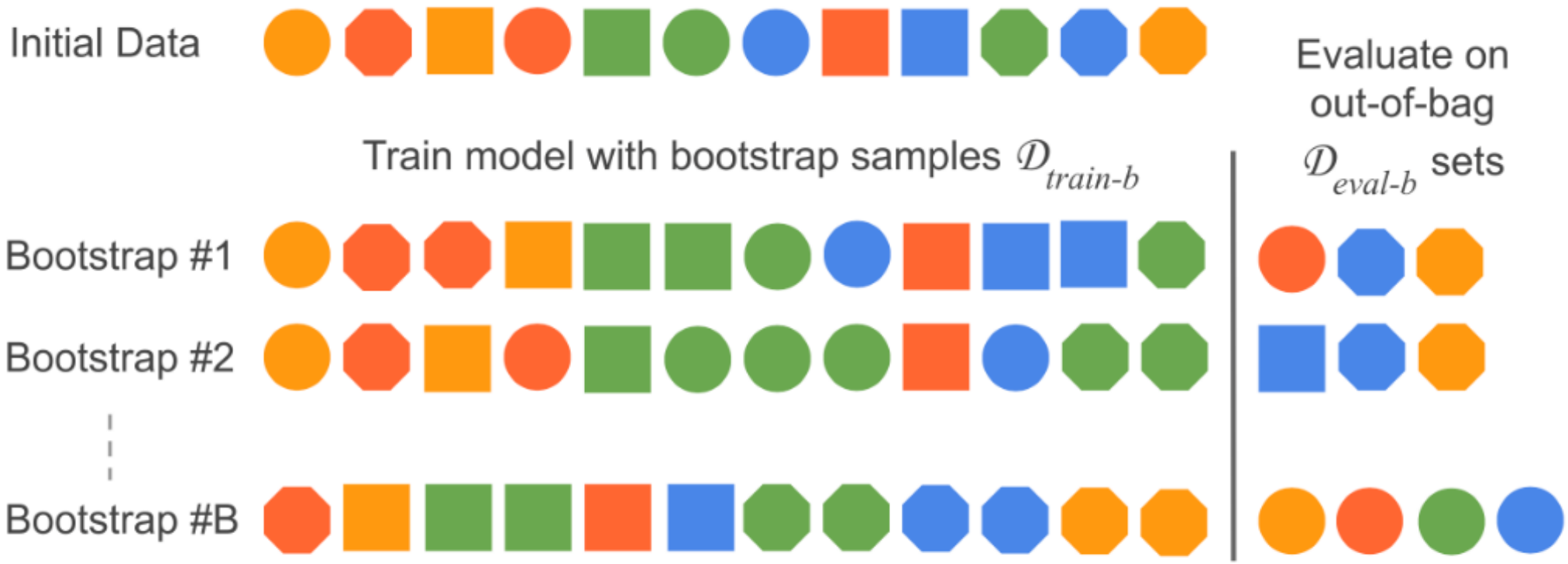
From all models are wrong (Sanchez and Marzban)

Three-way holdout



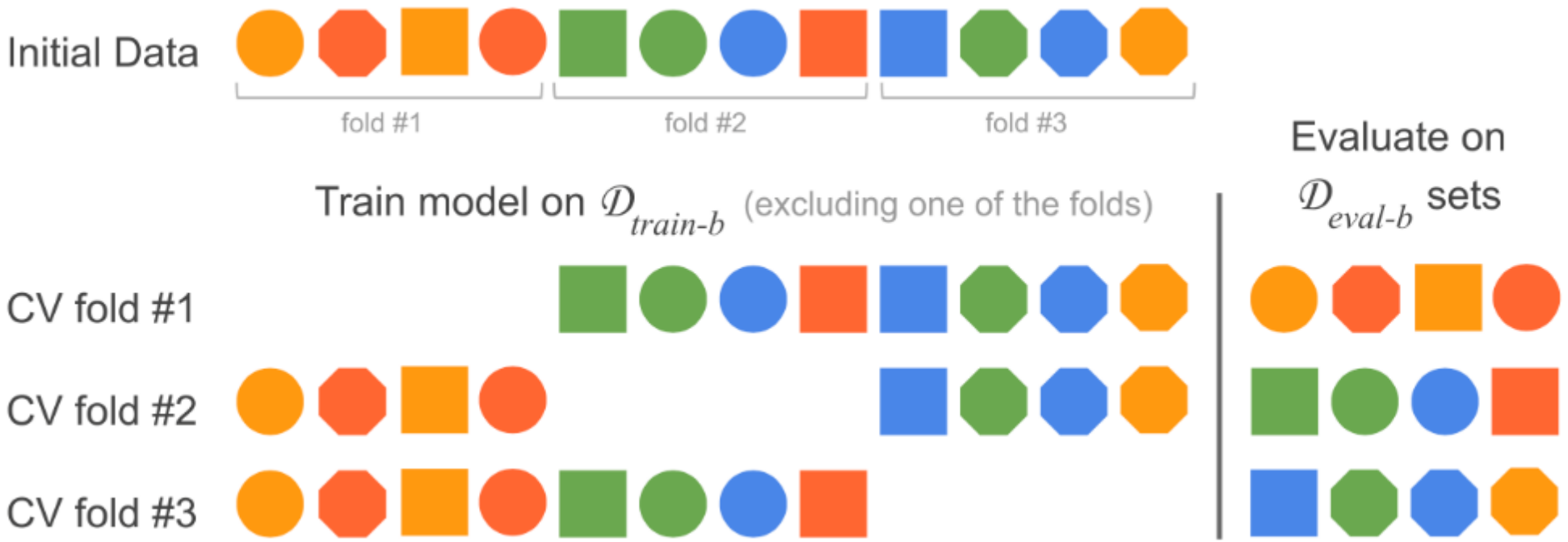
From all models are wrong (Sanchez and Marzban)

The Bootstrap



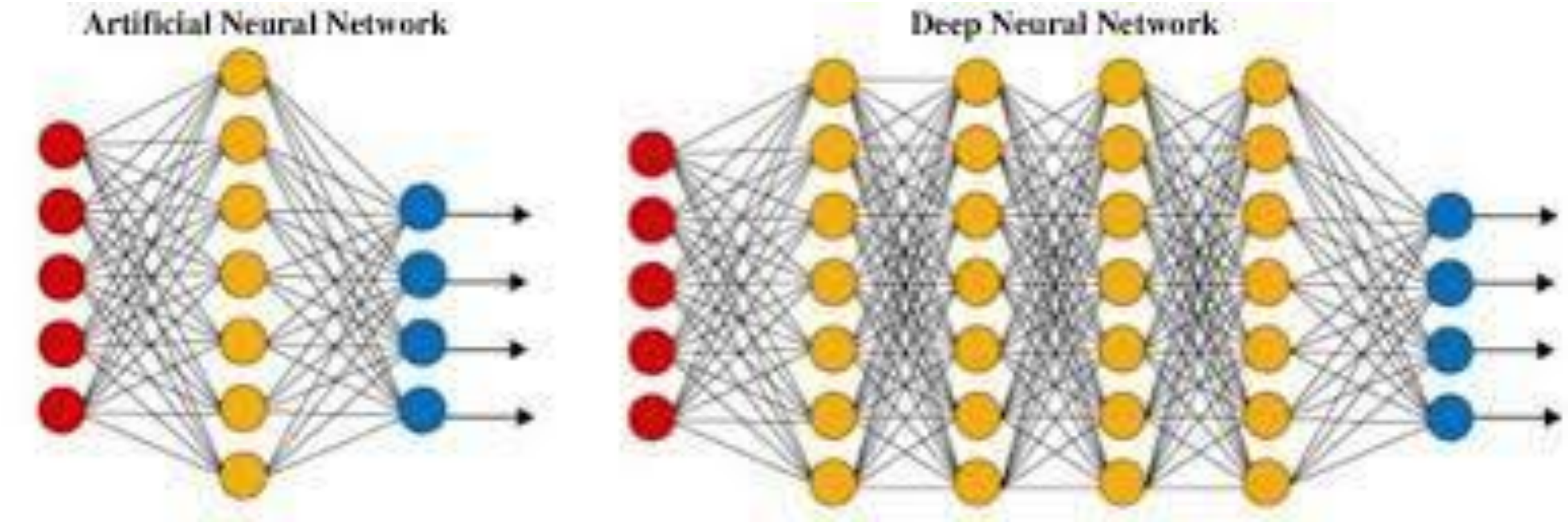
From all models are wrong (Sanchez and Marzban)

K-Fold Cross Validation



From all models are wrong (Sanchez and Marzban)

Deep Learning a Black Box



Black-box: inability to fully understand an AI's decision-making process and the inability to predict the AI's decisions or outputs

Unsupervised Learning

Objective: Identify patterns in the data without specific instruction of what to do with it

Pros

- Can identify complex patterns that may not be obvious to experts
- Less costly and more flexible because does not need data labels
- Facilitates data exploration and hypothesis generation

Cons

- Difficult to evaluate
- Subjective Interpretation of Clusters
- Vulnerable to noise in the data

Unsupervised Algorithms

- Hierarchical clustering
- K-means clustering, PAM
- Principal component analysis (PCA), Factor analysis
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- t-SNE (t-Distributed Stochastic Neighbor Embedding)
- Gaussian mixture models
- Ensemble methods

Semi-supervised Learning

Objective: combines elements of both supervised and unsupervised learning. Data labels are only available on a small subset of the data which also contains un-labelled data.

Pros

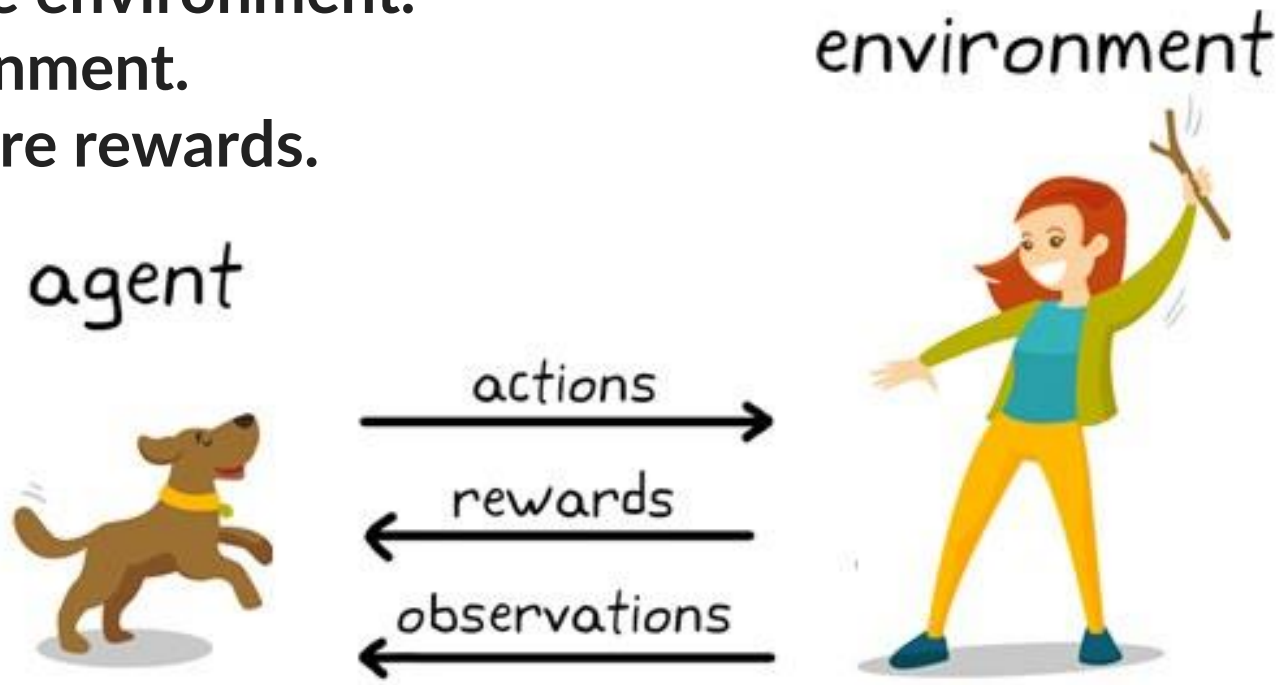
- Can make use of a large amount of unlabelled data
- More cost effective because only small subset requires label
- Weak supervision improves generalization

Cons

- Difficult to balance the distribution of labeled and unlabeled data
- More complex to implement than other algorithms
- Vulnerable to the quality of unlabeled data

Reinforcement Learning

1.Start in a state.
2.Take an action.
3.Receive a reward or penalty from the environment.
4.Observe the new state of the environment.
5.Update your policy to maximize future rewards.



Reinforcement Learning Example – KDNuggets

Reinforcement Learning (Expert Systems)

Objective: Agent learns to make decisions by interacting with its environment. The agent takes actions receives feedback in the form of reward and punishment. The goal is to develop a policy to maximize the cumulative reward over time.

Pros

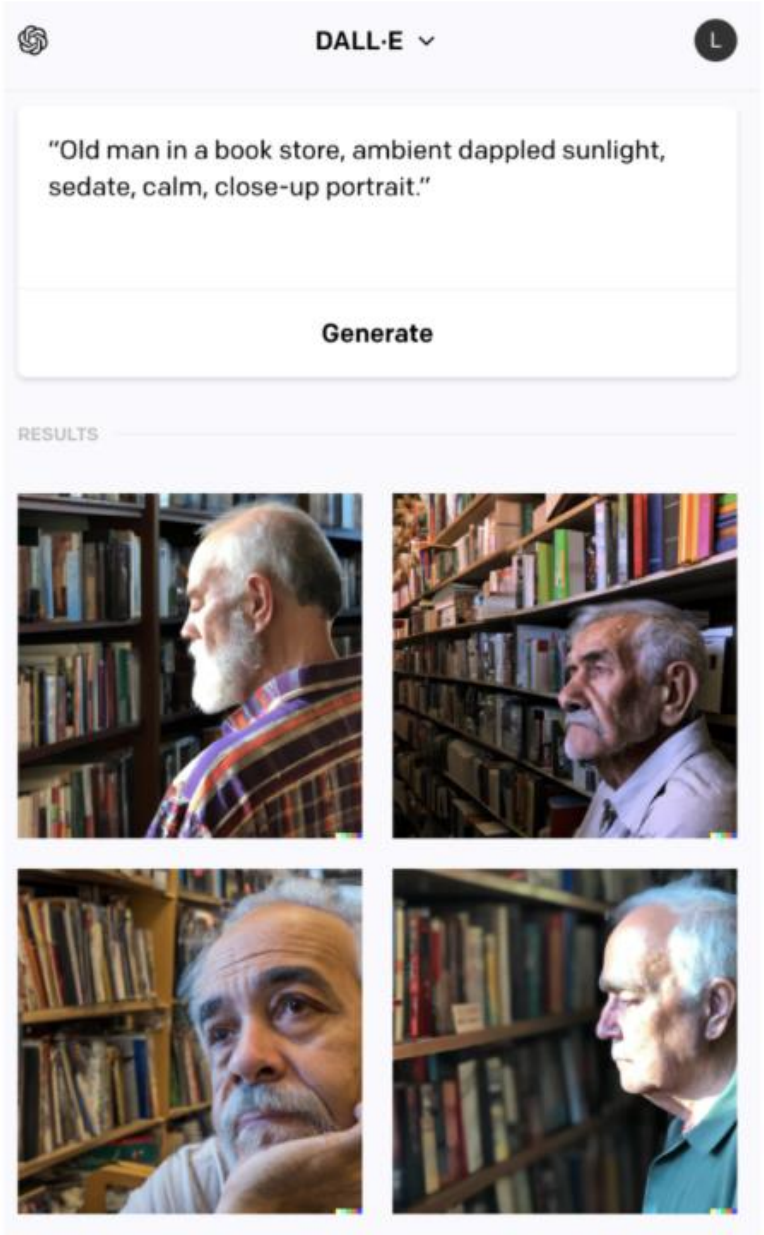
- Work well a complex and dynamic environment
- Grounded in a decision theoretic framework
- Learn by trial and error through interactions and improves over time

Cons

- May require large number of interactions
- Instability and limited understanding of internal representation
- May optimize a pre-set reward at the expense of other consequences

Generative Models

- Generative AI are ML models able to dynamically generate synthetic output after being trained on real data
- Generative AI learns the data generation mechanism and replicates it
- The creation of synthetic output, like images, texts, sonnets or code, is what distinguishes generative AI from other ML models



Created using DALL-E

Large Language Models

Language models are trained on text data and they try to predict the next word

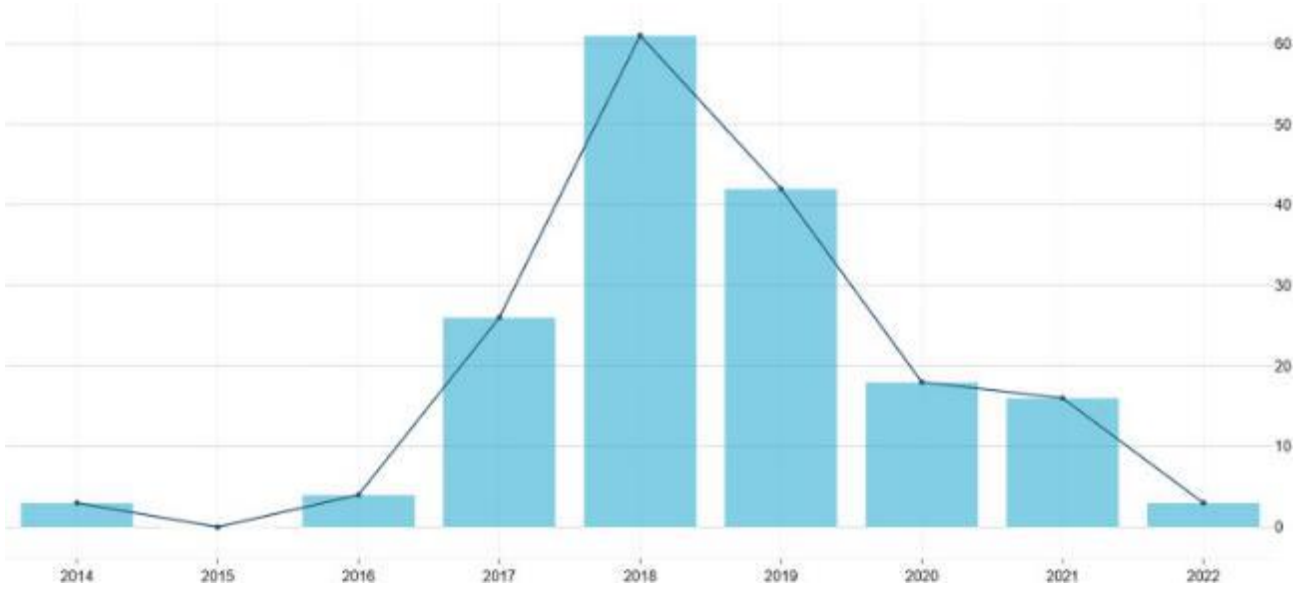
Generative Pre-Trained Transformer (GPT) are a class of Large Language Models based on transformer architecture

They are fed a huge amount of textual data from the internet. The algorithm learns to predict what words will follow given all the word combinations it has seen in training

ML Algorithms are as good as the data they are trained on

- Data hungry they work best with large n **and** large p
- Data may not capture the biological signal, and may be too noisy or incomplete or of poor quality
- How the data is handled and processed could drown signal
- Data may not be representative of the population
- Data may shift over time and space and the algorithm could become stale and fail to generalize

AI Ethics Boom

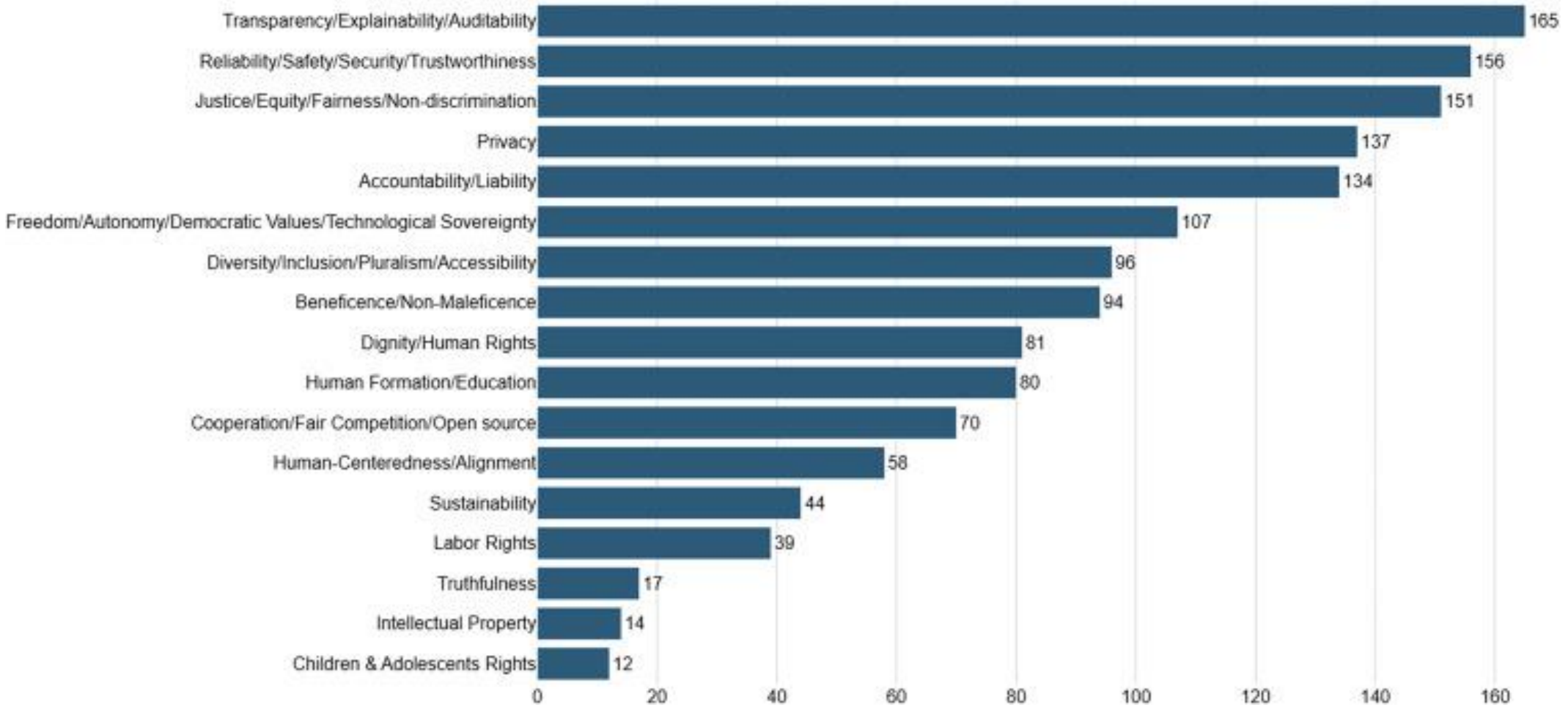


Since 2014, 5 time increase in pubs related to AI ethics:

2014 fairness, reliability, dignity
2016 accountability, beneficence, and privacy
2018 transparency and explainable AI

Patterns Volume 4, Issue 10 (October 2023)

19 Ethics Principles Identified



Patterns Volume 4, Issue 10 (October 2023)

Transparency/explainability/auditability the use and development of AI technologies should be transparent for all interested stakeholders and understandable to nonexperts and, when necessary, subject to be audited

Reliability/safety/security/trustworthiness AI technologies should be reliable, safe and robust, promoting user trust and adoption

Justice/equity/fairness/non-discrimination regardless of the different sensitive attributes that may characterize an individual, algorithmic treatment should happen "fairly”

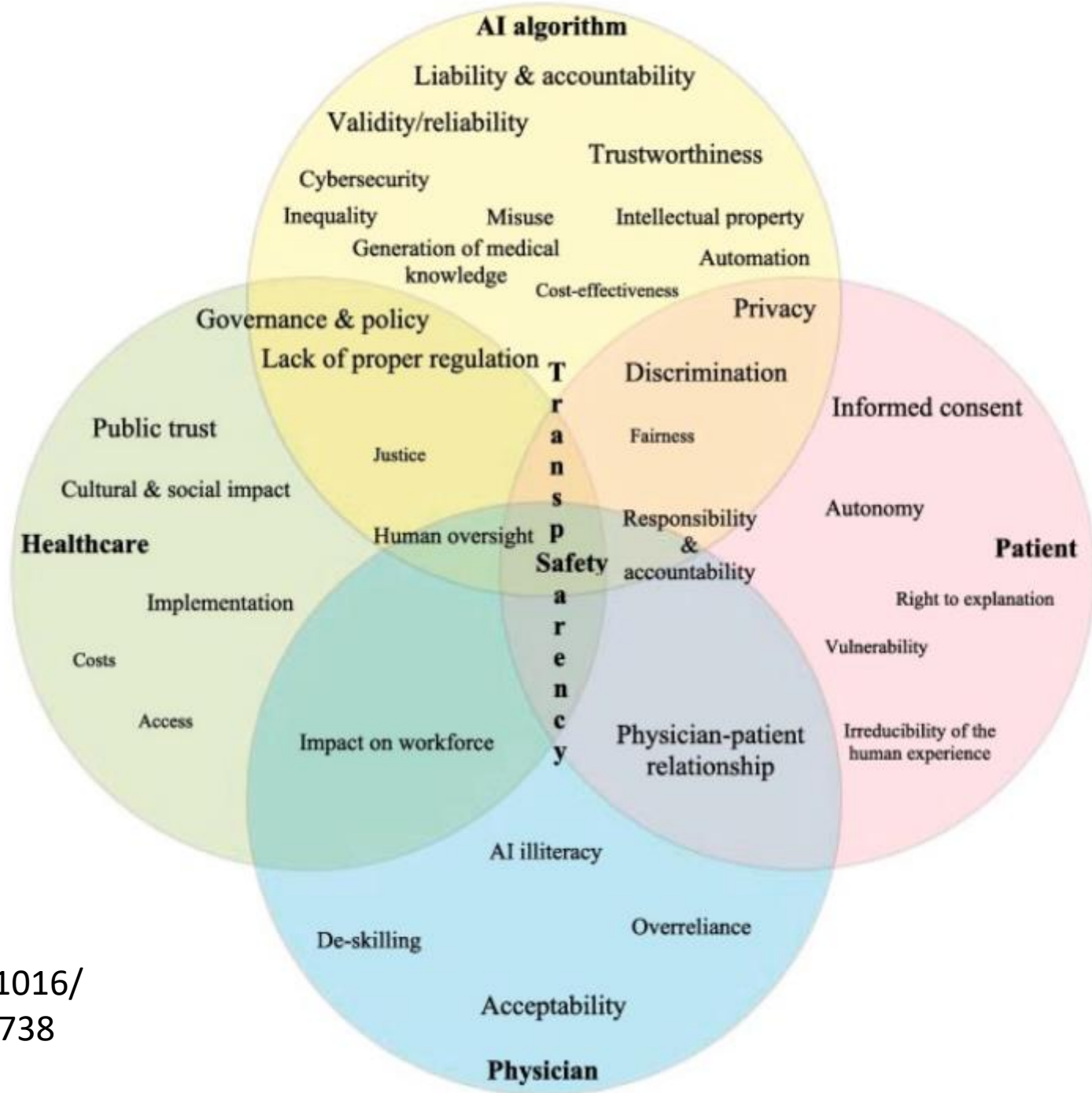
Privacy the right to "expose oneself voluntarily, and to the extent desired, to the world”. Also related to data protection such as data minimization, anonymity, informed consent, and others

Accountability/liability developers and deployers of AI technologies should be compliant with regulatory bodies and accountable for their actions and the impacts caused by their technologies

Issues in Health Care

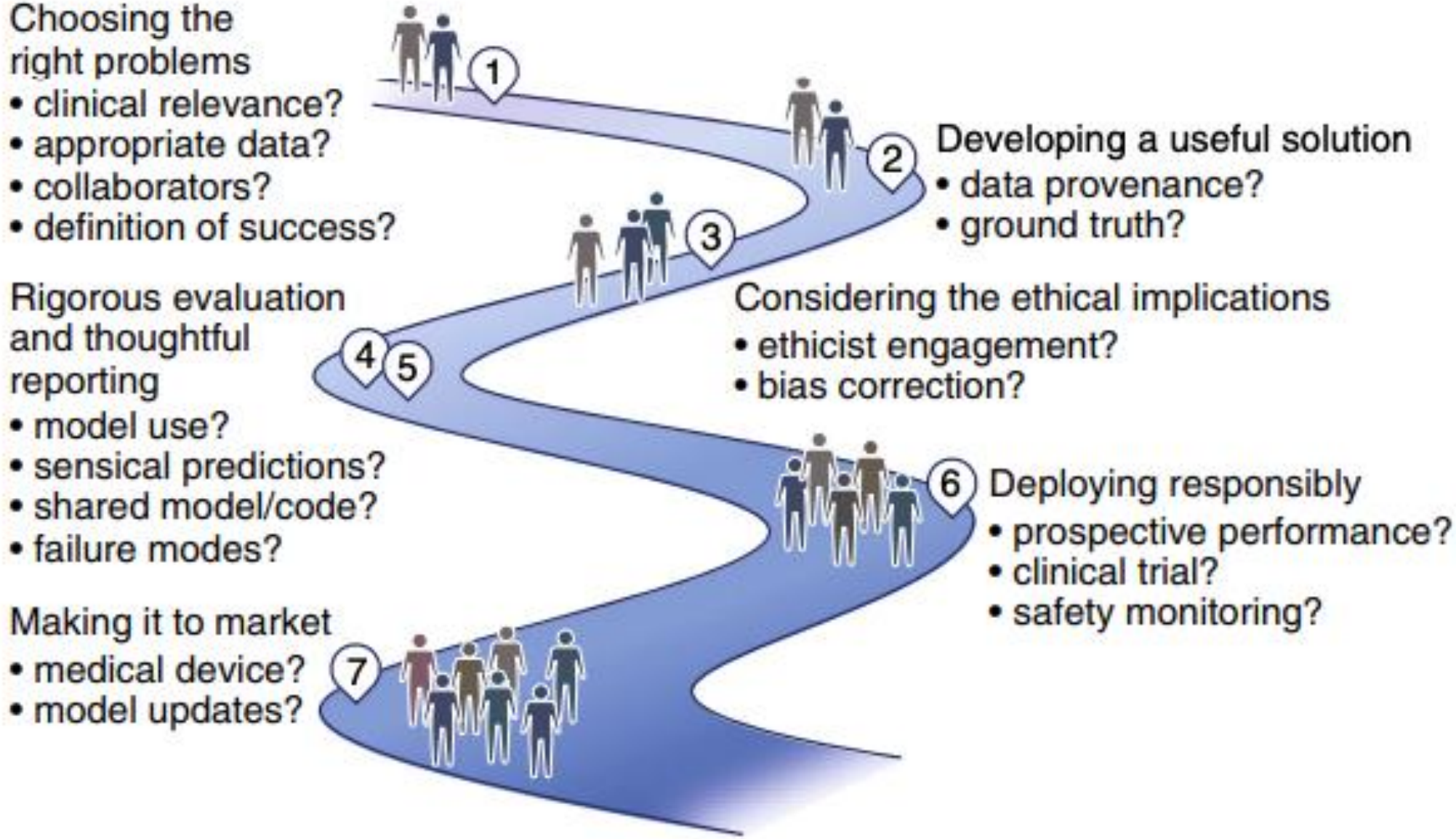
- Patient Safety
- Transparency/explainability/blackbox/opacity
- Bias in decision-making
- Accountability/Responsibility/Liability
- Undermine patient trust

https://doi.org/10.1016/
j.ijmedinf.2022.104738



https://doi.org/10.1016/j.ijmedinf.2022.104738

Consideration during ethical review



Nat Med. 2019 Sep;25(9):1337-1340.

Choosing the right problem

- The problem is defined clearly and the use of AI is well motivated
- Clear articulation of how and where the AI will be used
- Collaboration with domain experts to gain perspective on clinical relevance
- Considerations were given to who will be impacted, who makes decisions, who uses the system, and who benefits
- Researchers have consulted with diverse perspectives and are aware of the direct and indirect impacts on users/patients, and ensured fair benefits

Developing a useful solution: Engagement

- Engagement with operational, administrative leaders, patients, government, etc.
- Engagement with the research ethics board or ethicist

Developing a useful solution: Data

- Specify the data used for training, detailing its collection methods, timing, and purpose
- Evaluate if whether data used for training and evaluation is generated under a context similar to its intended use in practice (e.g. data distribution)
- Distinguish between observational data and data from different designs, considering biases and consistency (e.g. indication biases in observational data)
- Do available data exacerbate inequities?
- Are there privacy concerns and what is the risk of re-identification? Informed consent?

Developing a useful solution: Algorithm/Architecture

- Dealing with missing values is it ethical to impute? HIV or smoking status if the patient didn’t want to share the data?
- Specify the ML model and its architecture
- Clearly define the model's objective function, outlining what it is being trained to accomplish.
- Justify the selection of performance metrics and explicitly define success criteria for the model.
- Prioritize reproducibility; document procedures to ensure the model's consistent outcomes.

Rigorous Evaluation

- Implement train/test split to prevent overfitting during model training
- Avoid double-dipping by separating data used for training and evaluation
- Clearly define and report the model's likely success and failure scenarios
- Validate model performance internally and externally to ensure reliability
- Beyond a single AUC measure, consider sensitivity, specificity, PPV, NPV, and net benefit in evaluation
- Beyond discrimination, model calibration and net benefit evaluation

Deploy Responsibly

- Ensure models are plausible and clinically interpretable for practical application
- Be cautious of privacy leaks during the model training and deployment
- Conduct post-hoc evaluations to detect and rectify biases in the model
- Regularly calibrate, monitor drift, and retrain models for ongoing effectiveness
- De-risk deployment through study design

Thank you!