



# CHEP 801.3:

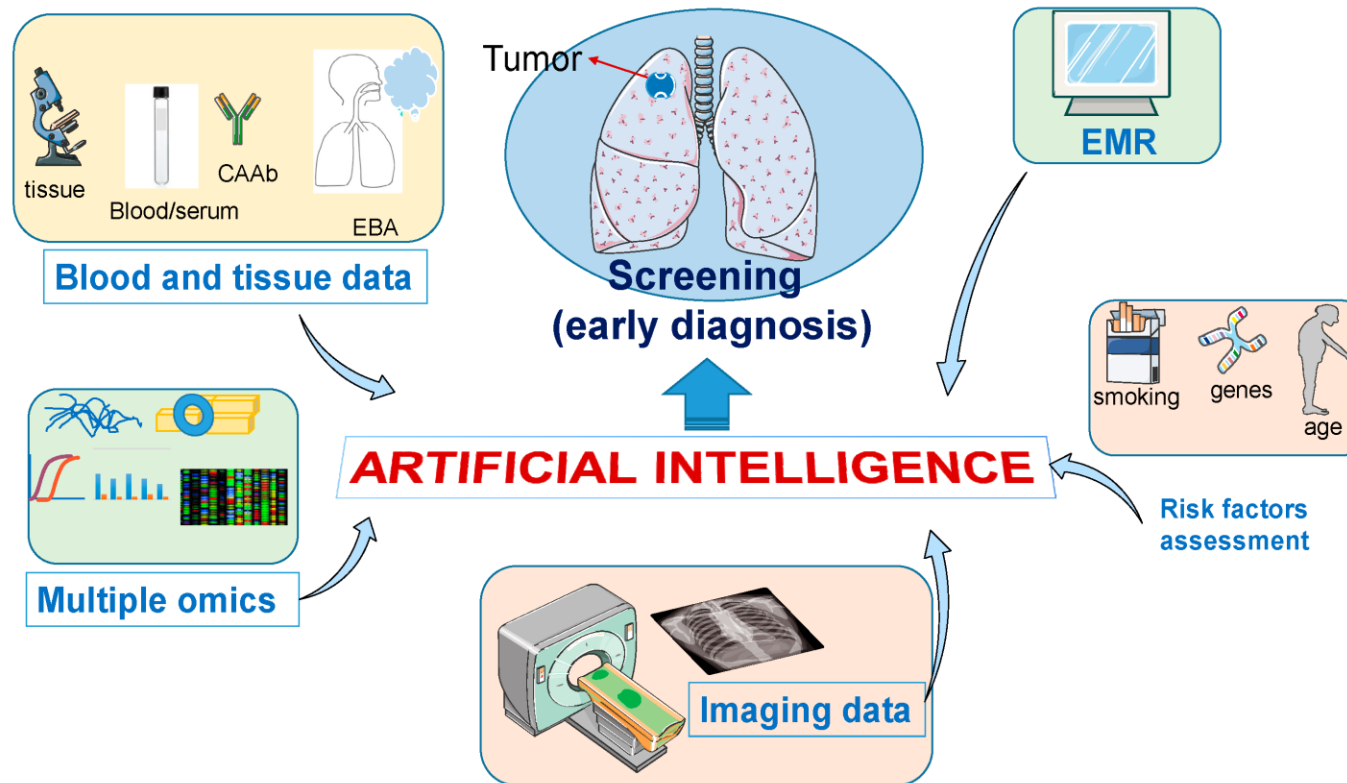
# Machine/statistical learning and predictive analysis

Department of Community Health and Epidemiology

# Classification of epidemiologic investigations

## Predictive analysis, Machine learning, AI, statistical learning

To predict or forecast health outcomes and exposures of patients or communities.



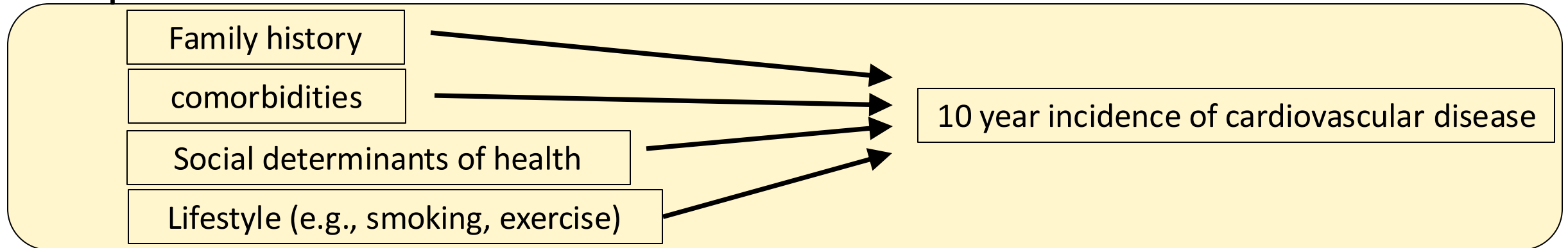
Espinoza et al (2020) <https://doi.org/10.3390/jcm9123860>

# Prediction V.S Explanation

- Etiologic (explanatory) analysis – estimating confounder-adjusted, unbiased association.
- Predictive analysis - Predicting values of the dependent variable from independent variable(s). Association between the dependent and independent variables can be confounded.

Research question: Association of 30-pack year intake of smoking and cardiovascular disease, after adjusting for other independent variables

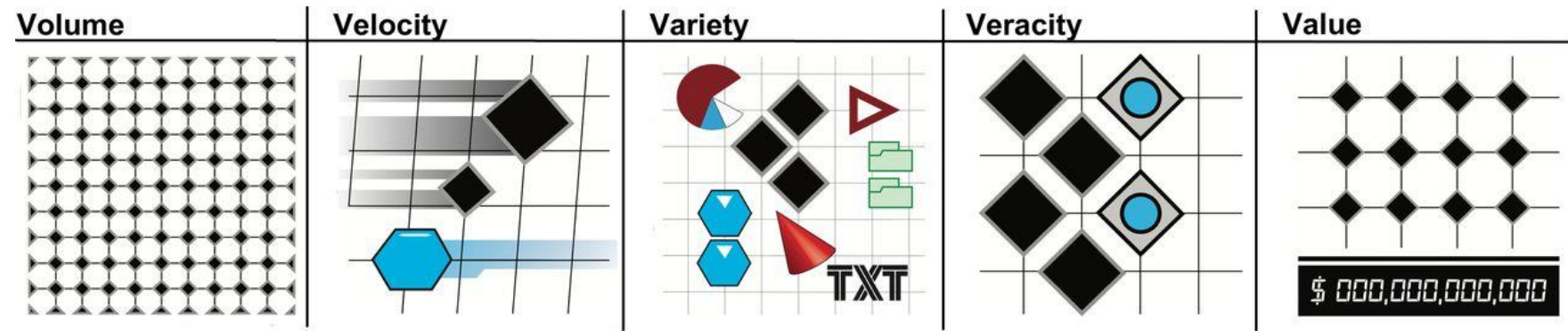
## Conceptual model



Research question: Which patients in your community gets cardiovascular diseases, given all the independent variables?

# Definitions

- Big Data
  - Complex and large amounts of information
  - Summarized by 5 V's
    - Volume
    - Velocity
    - Variety
    - Veracity
    - Value



# Machine Learning

- Big Data
  - You can analyse “big data” with traditional approaches we have learned in this class like linear and logistic regression
  - But often to get more out of the data you need methods more suited to the data
- Machine Learning
  - “A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E'$ )”

# Artificial Intelligence

- Artificial Intelligence
  - “Describe machines that perform human-like activities such as learning, perception, problem solving and playing games.”
  - AI’s often combine multiple different types of machine learning models to do different tasks and product an output.
    - Generate Christmas music from a picture
    - <http://www.cs.toronto.edu/songfrompi/>
      - Extract information from picture (one model)
      - Produce text (another model)
      - Produce music (another model)
      - Put it all together (maybe another model)

# Types of Machine Learning

- Supervised learning
  - *Supervised learning* is akin to the type of model-fitting that is standard in epidemiologic practice: The value of the outcome (i.e., the dependent variable), often called its “label” in machine learning, is known for each observation. Data with specified outcome values are called “labeled data.” Common supervised learning techniques include standard epidemiologic approaches such as linear and logistic regression, as well as many of the most popular machine learning algorithms (e.g., decision trees, support vector machines).
- Unsupervised learning
  - The algorithm attempts to identify natural relationships and groupings within the data without reference to any outcome or the “right answer”. Unsupervised learning approaches share similarities in goals and structure with statistical approaches that attempt to identify unspecified subgroups with similar characteristics (e.g., “latent” variables or classes). Clustering algorithms, which group observations on the basis of similar data characteristics (e.g., both oranges and beach balls are round), are common unsupervised learning implementations.

# Types of Machine Learning

- Semi-Supervised learning
  - Semisupervised learning fits models to both labeled and unlabeled data. Labeling data (outcomes) is often time-consuming and expensive, particularly for large data sets. Semisupervised learning supplements limited labeled data with an abundance of unlabeled data with the goal of improving model performance (studies show that unlabeled data can help build a better classifier, but appropriate model selection is critical)

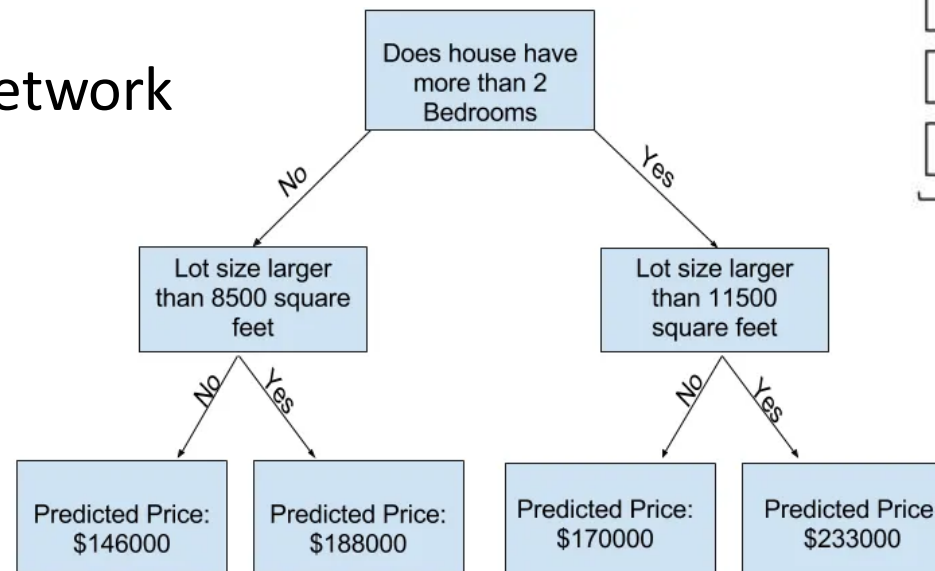
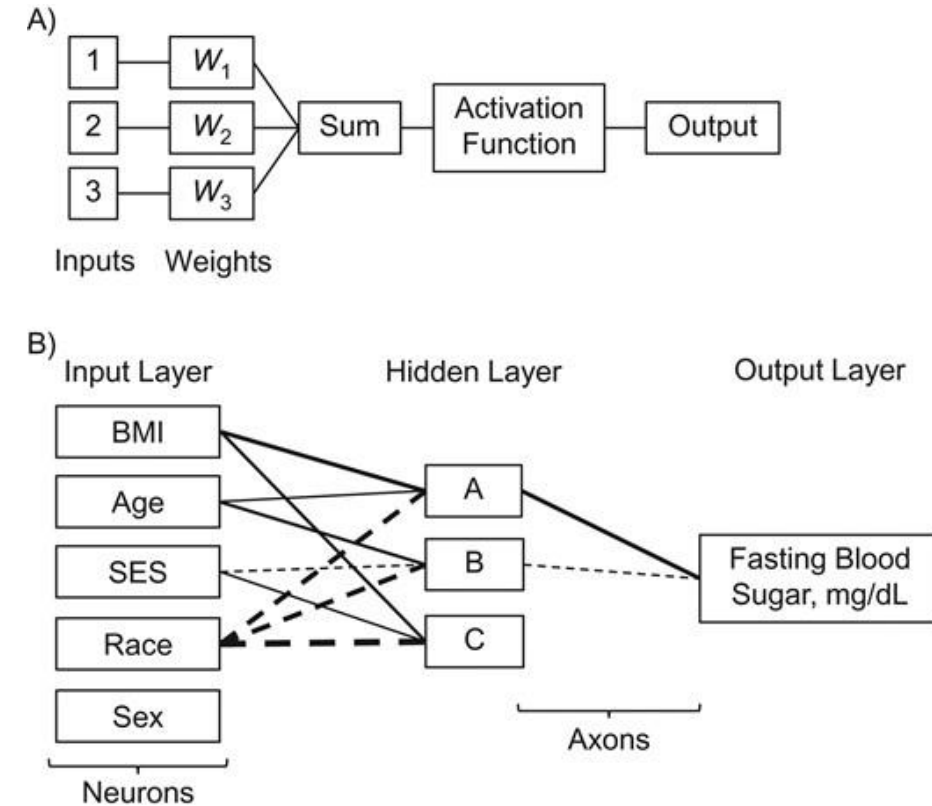


# Types of Machine Learning

- Classification
  - Identifying the category where an observation belongs, given known category labels. Logistic regression is an example of a classifier from statistics.
- Regression
  - Predicting a continuous outcome. Linear regression is an example of a classifier from statistics.

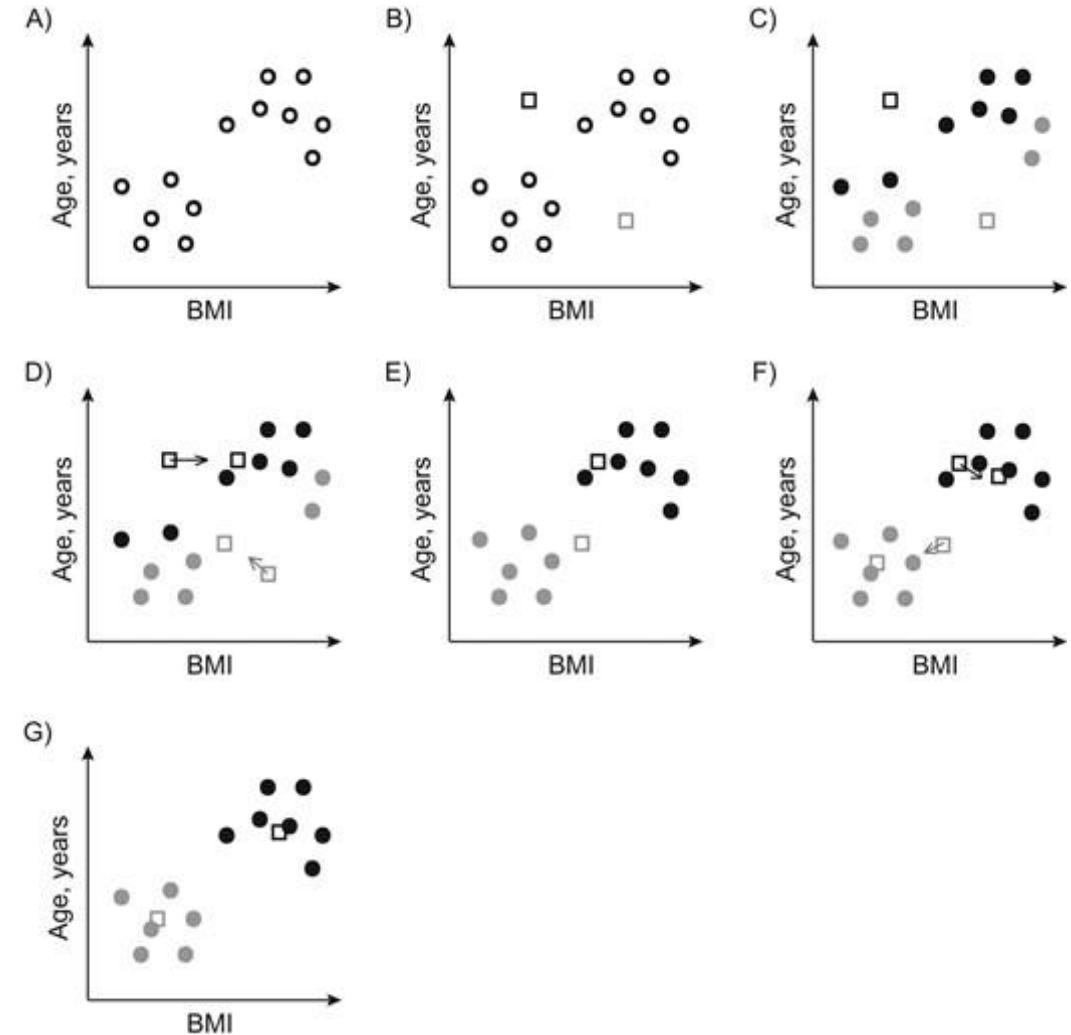
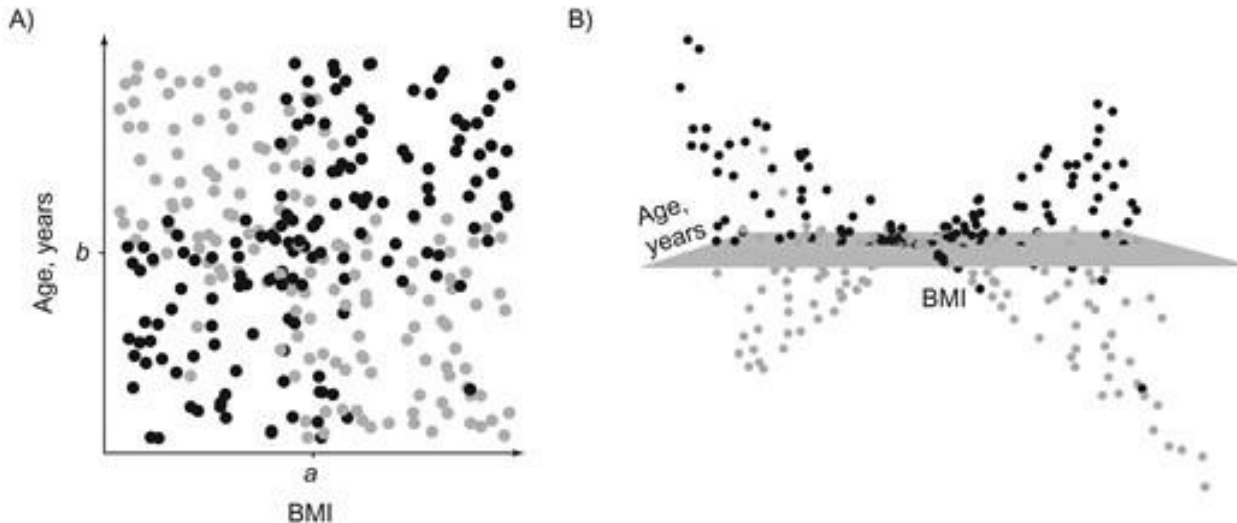
# Supervised Learning - Regression

- Linear Regression
  - Wait what?
- Decision Tree
  - Combination of a bunch of logistic regressions (basically)
- Random Forest
  - Ensemble (combination of) a bunch of decision trees
- Artificial Neural Network



# Supervised Learning - Classification

- Logistic Regression
  - Wait what?
- Support Vector Machine
- Naïve Bayes
- Random Forest
- Artificial Neural Network



# Unsupervised Learning

- Principal Component Analysis
- K Nearest Neighbour
- t-distributed Stochastic Neighbor Embedding (t-SNE)
- Natural Language Processing (NLP)

# Models

- Within supervised and unsupervised there are LOTS (hundreds) of different model types. These might have slightly different names or variations for specific tasks or to address different types of data.
- There are many flavours (variations) of Random Forest and Artificial Neural Network type models.
  - It can be hard to keep up.
- Pet Peeve!!!
  - Saying “using machine learning” is like saying “using statistics” it’s nearly completely uninformative in terms of describing what you actual did for your analysis.

# Quick Glossary

Machine Learning Term(s)	Epidemiology Term(s)	Definition and Notes
Attribute, feature, predictor, or field	Independent variable	Machine learning uses various terms to reference what epidemiologists would consider an “independent variable,” including <i>attribute</i> , <i>feature</i> , <i>predictor</i> , and <i>field</i> .
Domain	Range of possible variable values	The domain is the set of possible values of an attribute. It can be continuous or categorical/binary.
Input and output	Independent (exposure) and dependent (outcome) variables	In machine learning, “input” refers to all of the predictors or independent variables that enter the model, and “output” generally refers to the predicted value (whether a number, classification, etc.) of the dependent variable or outcome.

# Quick Glossary

Machine Learning Term(s)	Epidemiology Term(s)	Definition and Notes
Classifier, estimator	Model	“Classifiers” or “estimators” are used generally in the machine learning literature to refer to algorithms that perform a prediction or classification of interest. Their less common, though more technical, usage specifically refers to fully parameterized models that are used to predict or classify.
Learner	Model-fitting algorithm	A learner inputs a training set and outputs a classifier. Usually, but not always, <i>learner</i> refers to the fitting algorithm, while <i>classifier</i> refers to the fitted model.
Dimensionality	No. of covariates	No. of independent variables under consideration in a model.

# Quick Glossary

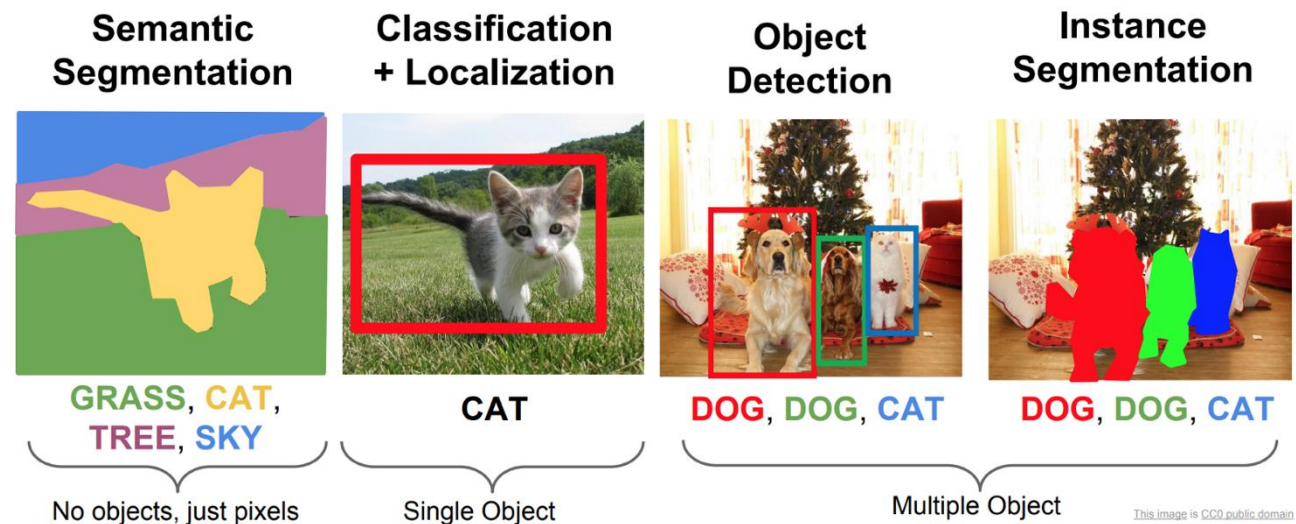
Machine Learning Term(s)	Epidemiology Term(s)	Definition and Notes
Label	Value of dependent variables, outcomes	A variable's label is its value for each observation (e.g., 0 or 1). Although labels can technically describe any variable, common shorthand is that "labeled data" refers to data in which the dependent variable assumes a value for all observations.
Imbalanced data	Data set in which some cases or risk categories occur much less frequently than the others	In imbalanced machine learning data sets, the outcome or another risk category of interest occurs much less frequently, either because of the intrinsic nature of the problem (e.g., a rare disease in a database of medical records) or because of the sampling strategy (e.g., prevalence of cases in the study population is much lower than that in the target/source population).
Loss function	Error measure	In machine learning, a loss function is generally considered a penalty for misclassification when assessing a model's predictive performance.



# General Applications

- NLP
  - Chat GPT (<https://chat-gpt.org/chat>)
  - Bing (<https://www.bing.com/new>)
- Computer Vision (Video or images)
  - Yolo (<https://docs.ultralytics.com/>)
  - Google (<https://cloud.google.com/vision/>)
- Human Activity Recognition (my stuff)

Examples for public health??



Write a reference letter for graduate student for an application to MIT computer science.

Dear Admissions Committee,

I am writing to enthusiastically recommend [Student Name] for the MIT computer science program. I had the pleasure of working with [Student Name] during their tenure in my research laboratory as a graduate student, and I was continually impressed by their skills, work ethic, and dedication to their studies.

As a graduate student in [discipline], [Student Name] possessed a particularly strong aptitude for computer science concepts and quickly developed a deep understanding of the field. Their technical abilities were demonstrated through their contributions to several successful research projects, where they were responsible for designing, implementing, and analyzing complex algorithms. [Student Name] consistently tackled difficult problems with diligence and creativity, often developing innovative solutions to challenging tasks.

Moreover, [Student Name] was an excellent communicator, both in writing and in presentations. They frequently interacted with research supervisors and peers, providing well-articulated, concise explanations of their work that effectively conveyed their ideas to a range of technical experts, including those without a computer science background. I am confident that [Student Name] will be an even stronger asset in graduate coursework that requires communication of technical findings to a broader audience.

Overall, I strongly recommend [Student Name] for admission to the MIT computer science program. They will undoubtedly excel in this program, and I look forward to seeing their future research contributions.

# Epidemiologic Applications

- Matching (Key challenge for epidemiologists)
  - Machine learning algorithms often deal implicitly with interactions and nonlinearities, whereas such high-order terms must be explicitly specified (and are commonly ignored) in logistic regression.
  - Machine learning algorithms also perform well in estimating propensity scores in the presence of high-dimensional data and can reduce underlying model misspecification.
- Effect heterogeneity
  - Machine learning methods have been used understand heterogeneity in treatment effects across subpopulations.
  - “Casual trees” that create decision trees where groupings are based on treatment effect and provide principled estimates of treatment effects within these strata.

# Epidemiologic Applications

- Causal structure learning
  - A group of exploratory techniques that identify an optimal directed acyclic graph consistent with conditional independence relationships in the data and provided background knowledge. Approaches to causal structure learning include Bayesian network approaches and linear, nongaussian, acyclic models (LiNGAMs).

# Critiques of Big Data/ML/AI

- **Automating research changes the definition of knowledge**
  - Minimisation of the importance of other areas of knowledge creation and ignores limitations of big data.
- **Claims of objectivity are misleading**
  - Big data can perpetuate the myth that quantitative data analysis is inherently objective.
- **Bigger data are not always better data**
- **Not all data are equivalent**
- **Just because it is accessible does not make it ethical**
  - Researchers must adhere to general ethical principles including respect for persons, concern for welfare and justice.

# Critiques of Big Data/ML/AI

- **Limited access to big data creates new digital divides**
  - Big data is often owned by the entity that collects it, giving them control over who can access their data and at what cost. High cost of data access can create disparities in who can conduct research.
    - Recent GPT-4 Paper: <https://openai.com/research/gpt-4>
- **Big data hubris**
  - With sufficient volume and velocity can compensate for or eliminate the need for high veracity data, and high-quality study designs

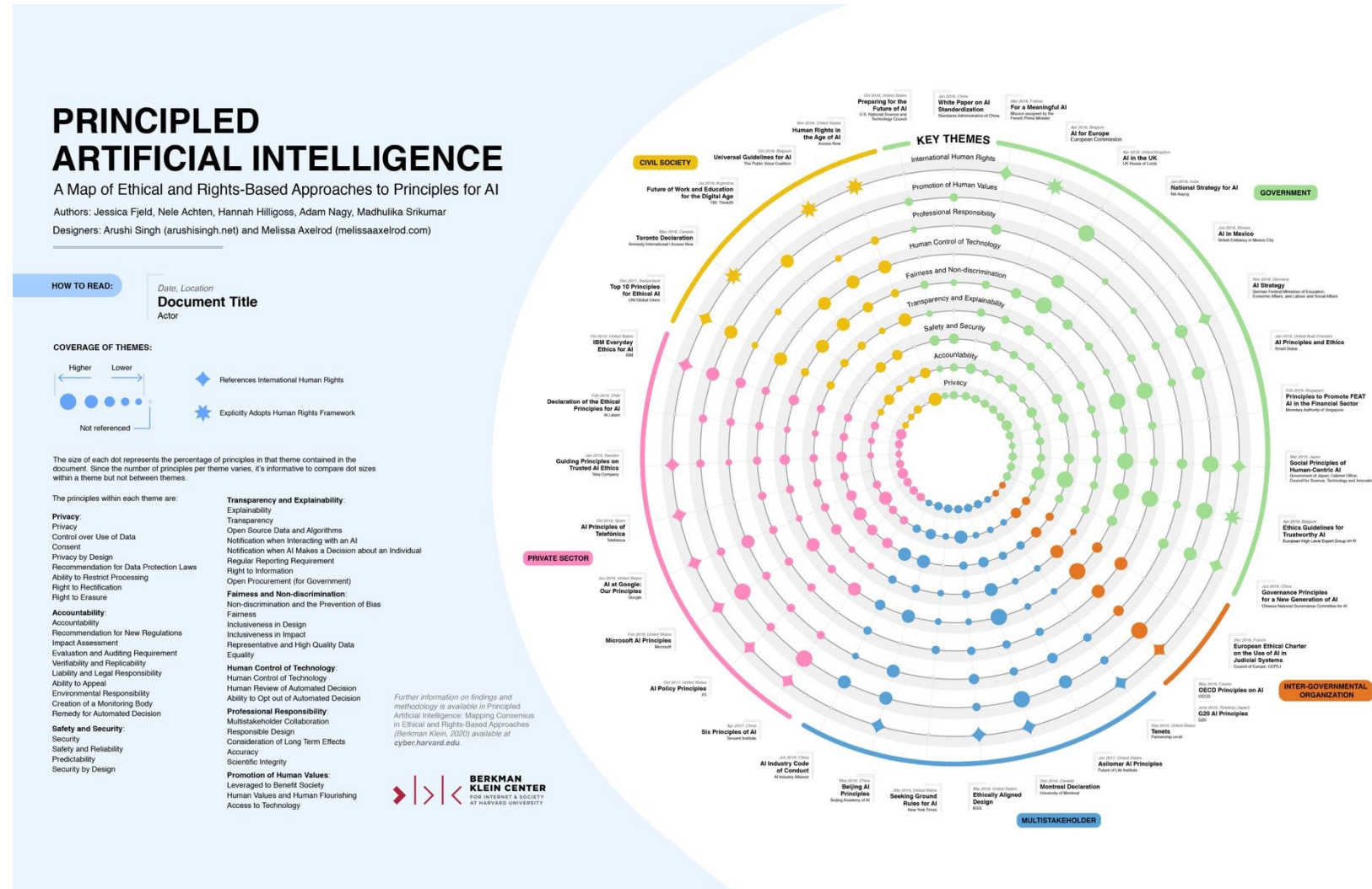
# Equity and Bias

- Many important concerns with different types of ML and AI applications
  - [Racist soap dispenser](#)
  - [Homophobic and racist chat apps](#)
  - "Predictive" policing
  - Many others...
- Leaders
  - Timnit Gebru (<https://time.com/6132399/timnit-gebru-ai-google/>)
  - Emily Bender (<http://faculty.washington.edu/ebender/>)

# Principled AI

Fjeld, Jessica and Achten, Nele and Hilligoss, Hannah and Nagy, Adam and Srikumar, Madhulika, Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center Research.

<http://dx.doi.org/10.2139/ssrn.3518482>





# What do I actually do?

- Key data concepts
  - **Labelled data**
    - Data where the value of the variable to be predicted is known.
  - **Training data**
    - Data used to train a machine learning model. Machine learning methods use features (ie, variables) to predict an outcome. However, unlike more traditional methods, the primary objective of the training process is predicting as much of the variance in the data as possible.
  - **Test set data**
    - Labelled data that are not included in the training process, but that are used to validate the model developed using the training data.

# What do I actually do?

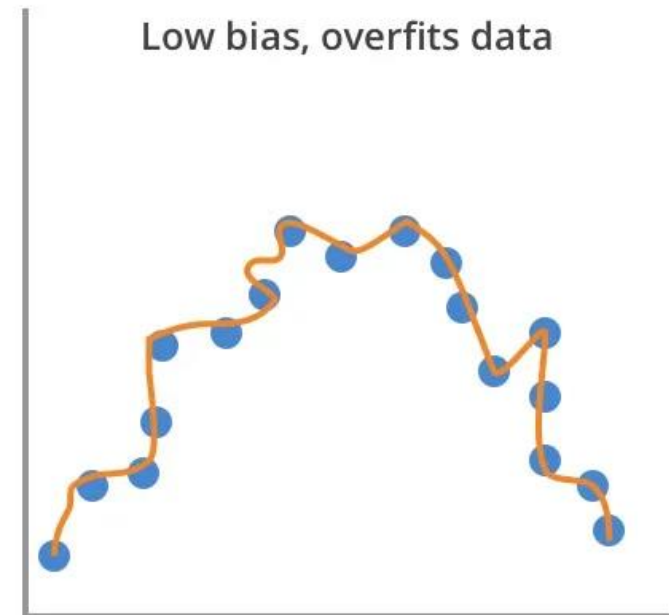
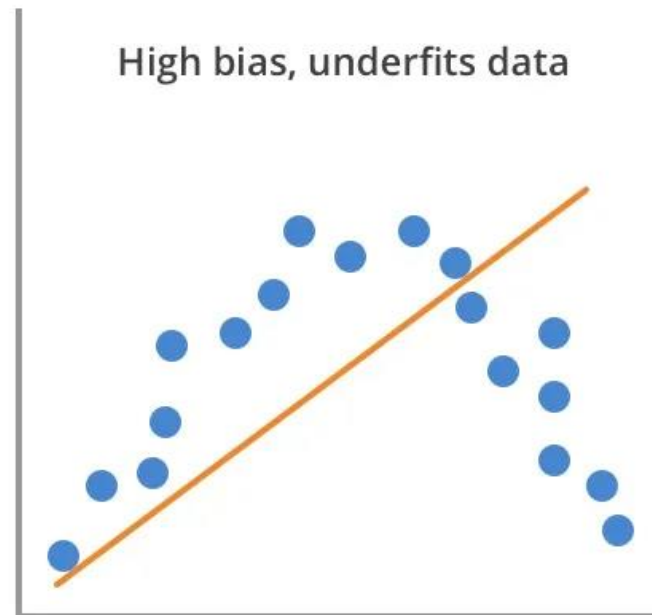
- Key data concepts
  - **Bias Variance Trade off**

$$\textit{Model Error} = \textit{Irreducible error} + \textit{Bias} + \textit{Variance}$$

- *Irreducible error*
  - Error caused by noise in the data, random variations that don't represent a real pattern in the data, or the influence of variables that are not yet captured as features.

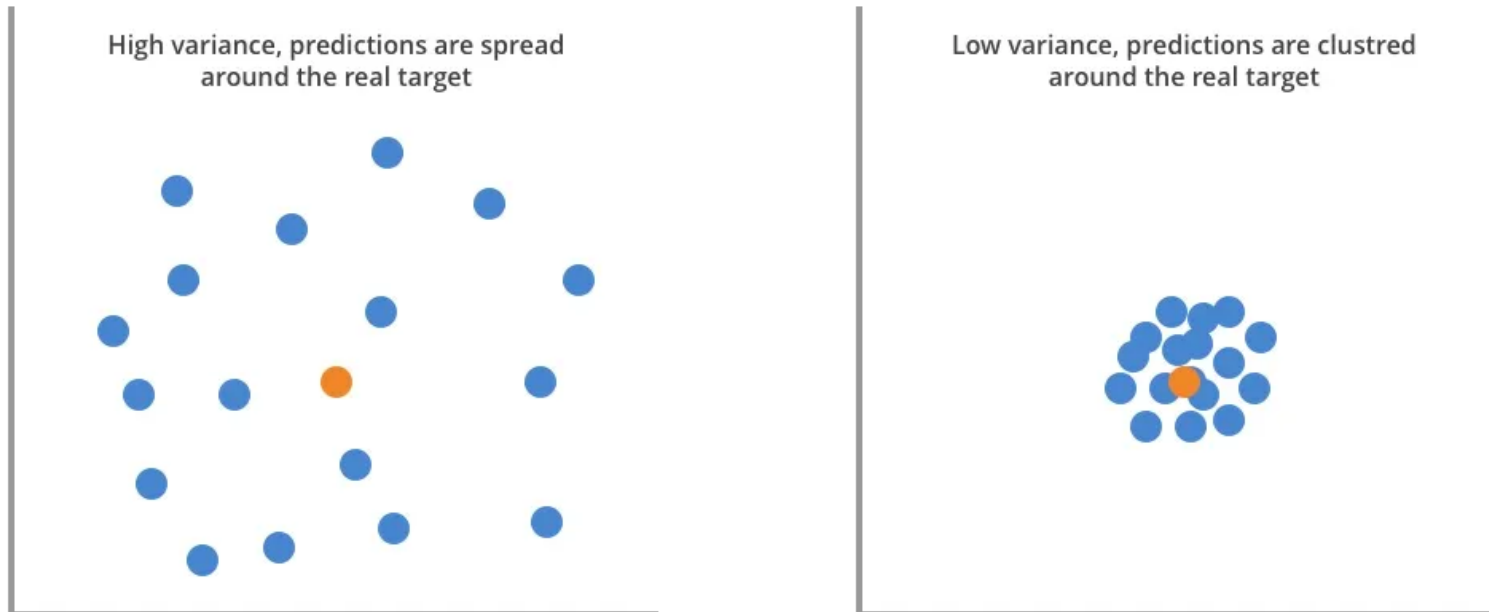
# What do I actually do?

- Bias
  - Bias is about the ability to capture the true patterns in the dataset.
  - A model with high bias will underfit the data. It will take a simplistic approach to model the true patterns in the data.
  - A model with low bias is more complex than it should be. It will overfit to the data it utilized to learn and fit poorly to the testing data.

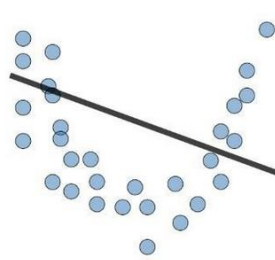
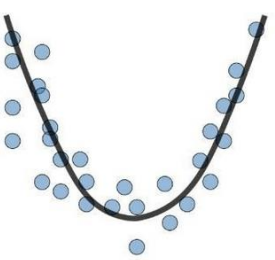
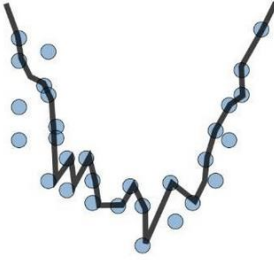
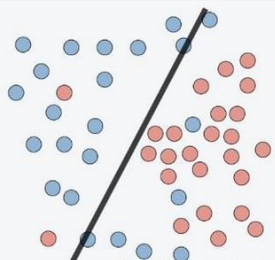
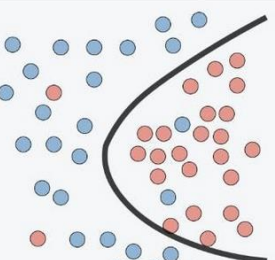
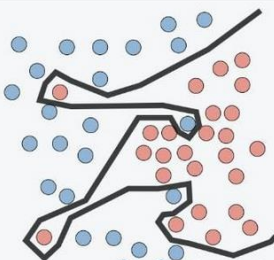
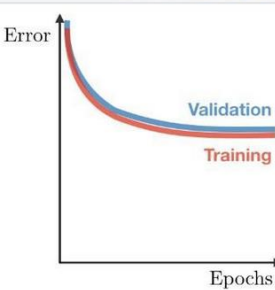
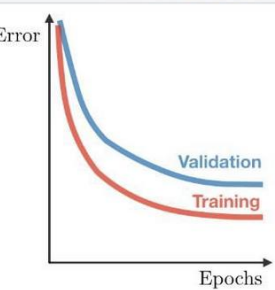
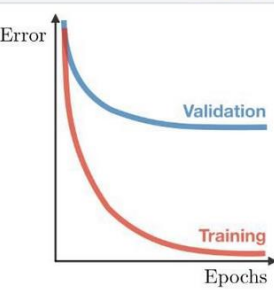


# What do I actually do?

- Variance
  - Variance captures the range of predictions for each data record.
  - A measure of how far off each prediction is from the average of all predictions for that testing set.
  - Reduce variance is to build a model with more training data. The model will have more examples to learn from and improve its ability generalize its predictions.



- Bias-Variance Trade off
  - Need to try and balance bias with variance
  - There is no right answer
  - Much of the testing and training is done qualitatively

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"> <li>• High training error</li> <li>• Training error close to test error</li> <li>• High bias</li> </ul>	<ul style="list-style-type: none"> <li>• Training error slightly lower than test error</li> </ul>	<ul style="list-style-type: none"> <li>• Very low training error</li> <li>• Training error much lower than test error</li> <li>• High variance</li> </ul>
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	<ul style="list-style-type: none"> <li>• Complexify model</li> <li>• Add more features</li> <li>• Train longer</li> </ul>		<ul style="list-style-type: none"> <li>• Perform regularization</li> <li>• Get more data</li> </ul>

# What do I actually do?

## 1. Have a research question!

- Not helpful to just run a bunch of analyses for not much purpose

## 2. Select features (variables of interest)

- We are interested in a set of features that predict the outcome as a whole (prediction). We are less interested in etiology of associations between variables.

## 3. Check balance

- Unlike our RR/OR in Epi, ML models are highly sensitive to imbalance

## 4. Split the data

- Basic ML tutorials will say do a 80/20 or 70/30 split. Use 70% of the data for training a 30% for testing.
- Other ways to split are k-fold cross validation or leave-one-out cross validation.
  - This decision is super important and study design specific.

# What do I actually do?

## 4. Train some models

- ML typically you will test a number (4-5) models and see which is better. Very different from Epi and Stats approaches.

## 5. Test models

- Select evaluation metrics that are meaningful for your problem
  - We will go over this
- Make decisions about the best performing model keeping in mind bias-variance trade off.
  - A model with 99% accuracy is very suspect.

## 6. Decide which models you will present and how

- Confusion matrix?
- Which evaluation metrics?
- Feature importance?

# Some key readings

- Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. *Am J Epidemiol*. 2019 Dec 31;188(12):2222-2239. doi: 10.1093/aje/kwz189. PMID: 31509183.
- Serghiou S, Rough K. Deep Learning for Epidemiologists: An Introduction to Neural Networks. *Am J Epidemiol*. 2023 Nov 3;192(11):1904-1916. doi: 10.1093/aje/kwad107. PMID: 37139570.
- Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. *Annu Rev Public Health*. 2020 Apr 2;41:21-36. doi: 10.1146/annurev-publhealth-040119-094437. Epub 2019 Oct 2. PMID: 31577910.
- Fuller D, Buote R, Stanley K. A glossary for big data in population and public health: discussion and commentary on terminology and research methods. *J Epidemiol Community Health*. 2017 Nov;71(11):1113-1117. doi: 10.1136/jech-2017-209608. Epub 2017 Sep 16. PMID: 28918390.