



(Statistical) Machine Learning – Building and Evaluating Data-driven Prediction Models

IDeA – CTR Great Plains, March 12, 2020

Dr. Christian Haas

Department of Information Systems and Quantitative Analysis

Co-Director Data Science Lab



Dr. Christian Haas

Department of Information Systems and Quantitative Analysis

Educational Background

- MSc in Computer Science, Georgia Institute of Technology
- PhD in Information Systems and Computational Economics, Karlsruhe Institute of Technology, Germany
- Post-PhD: 3 years as Senior Data Scientist at IBM Germany
- Since 2017: Assistant Professor, UNO.

Research Interests



Applied Machine
Learning



Multi-objective Optimization
and Simulation



Algorithmic Fairness

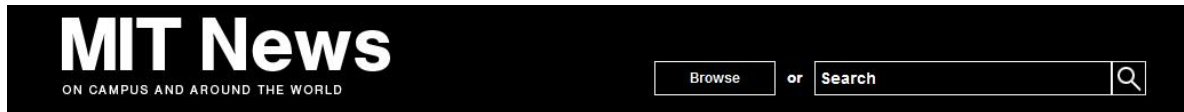


Agenda

Part 1: Data Science, Machine Learning, and Main Types of Models

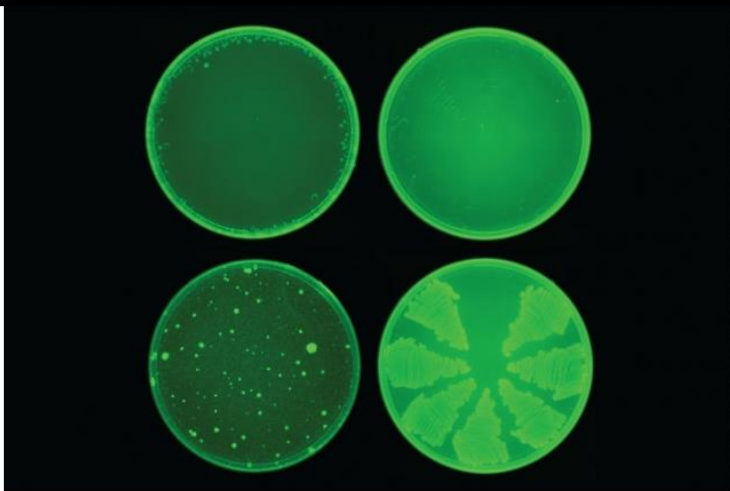
Part 2: Evaluation of ML models: how do we know if our model is working?

Part 3: Explainability, Fairness, and other considerations



MIT News
ON CAMPUS AND AROUND THE WORLD

Browse or Search



MIT researchers used a machine-learning algorithm to identify a drug called halicin that kills many strains of bacteria. Halicin (top row) prevented the development of antibiotic resistance in *E. coli*, while ciprofloxacin (bottom row) did not.

Image: courtesy of the Collins Lab at MIT

Artificial intelligence yields new antibiotic

A deep-learning model identifies a powerful new drug that can kill many species of antibiotic-resistant bacteria.

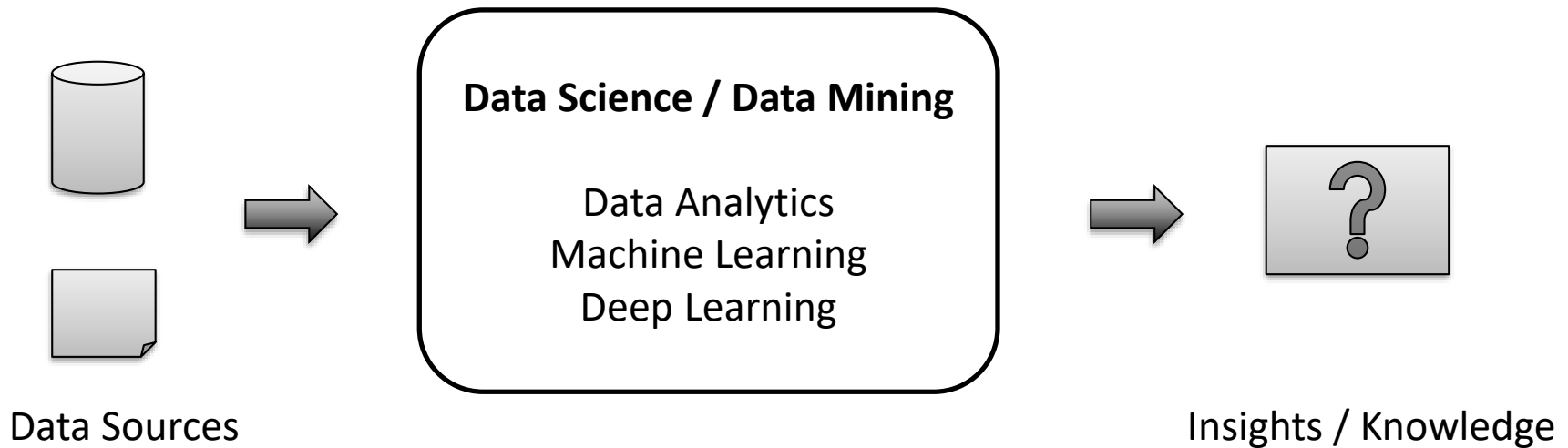
Anne Trafton | MIT News Office
February 20, 2020

Press Inquiries PRESS MENTIONS

“Using a **machine-learning algorithm**, MIT researchers have **identified a powerful new antibiotic compound**. In laboratory tests, the drug killed many of the world’s most problematic disease-causing bacteria, including some strains that are resistant to all known antibiotics. It also cleared infections in two different mouse models.”



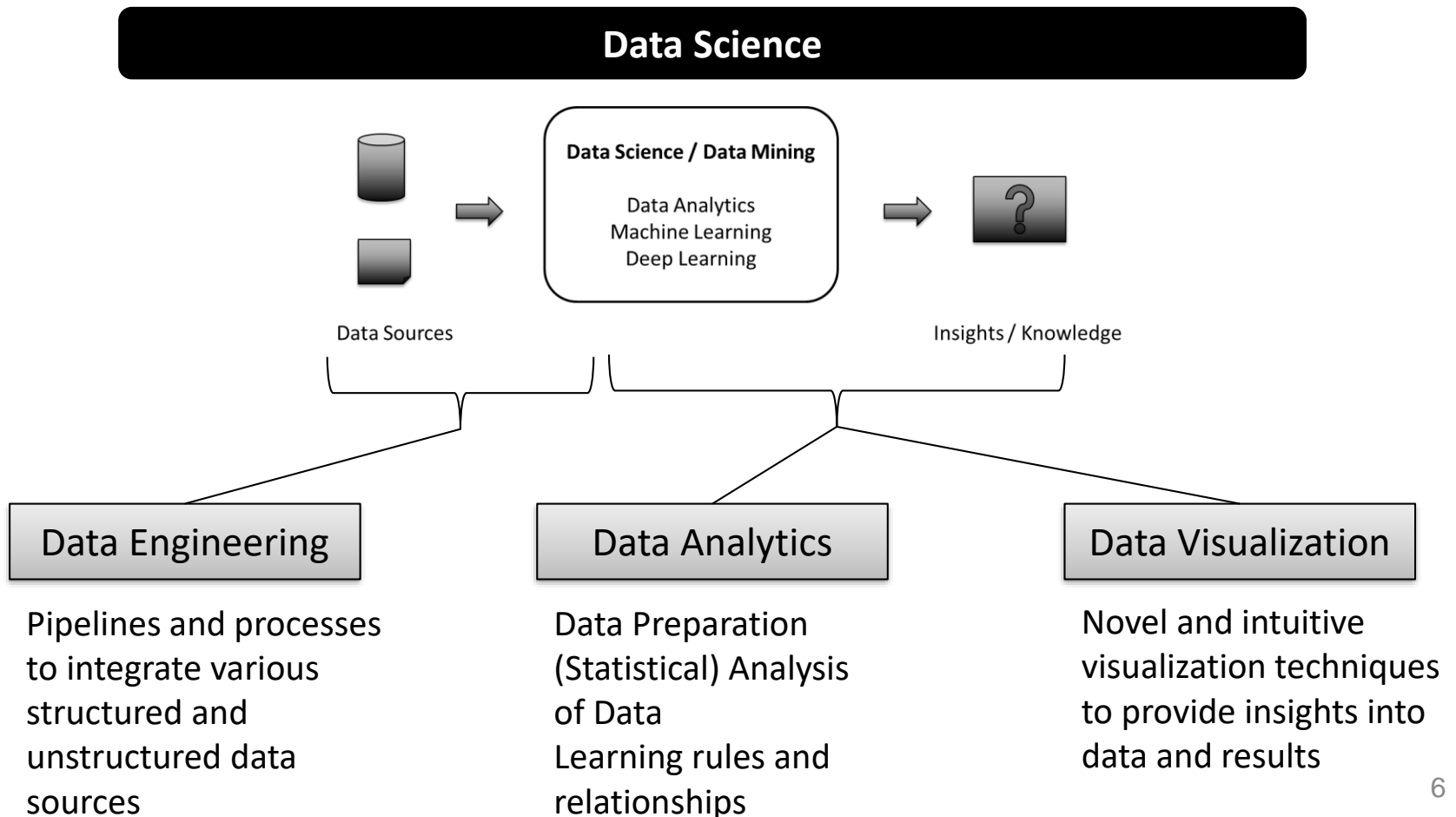
Data Science – An Overview



At a high level, data science is a set of fundamental principles, processes, techniques and technologies that guide the **extraction of knowledge from data.**

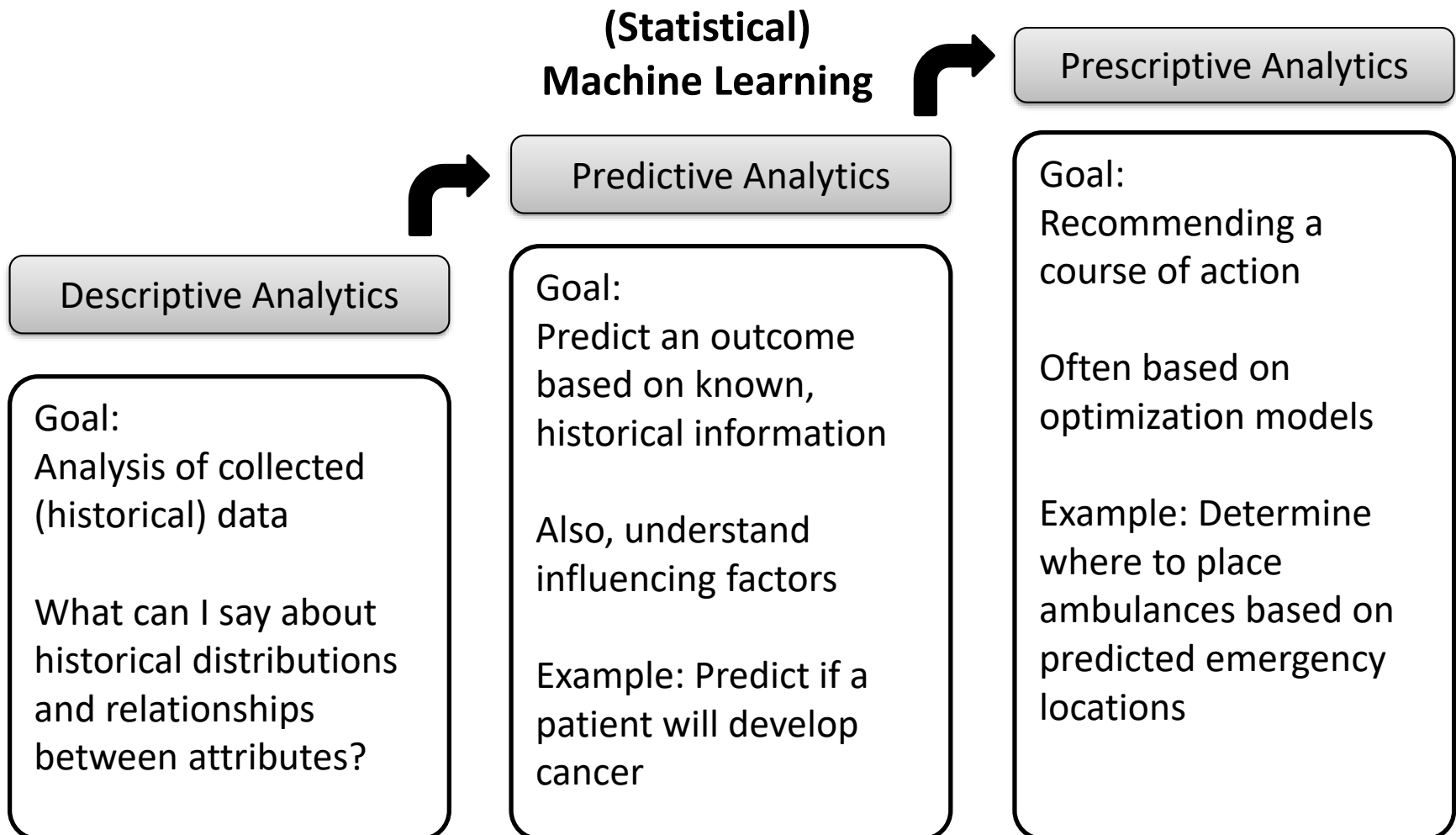


Data Science – (Some) Subdomains





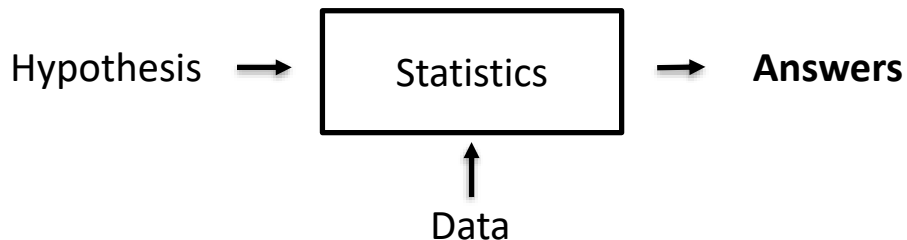
Data Analytics Areas – A Differentiation





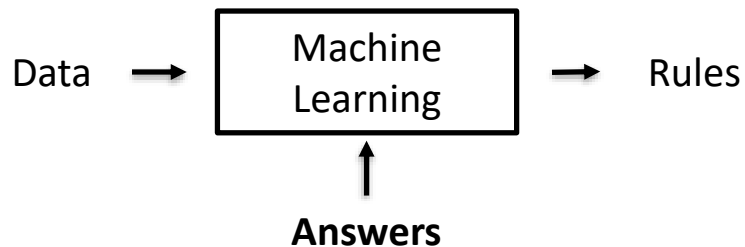
Machine Learning vs Traditional Statistics

Statistics



- Start with formulating hypotheses that you want to test
- Run experiments to collect data
- Goal: Confirm or reject hypotheses

Machine Learning



- Start with observed data
- Generalize these rules for new data
- Goal: Find rules that describe the relationship between the data and the answers/outcomes



Statistical Machine Learning – The Main Idea

Data

$$Y = \begin{pmatrix} Yes \\ No \\ \dots \end{pmatrix}$$

Outcome:
Disease, Churn, Condition, ...

$$X = \begin{pmatrix} 50 & No & 65 \\ 70 & No & \dots 87 \\ \dots & \dots & \dots \end{pmatrix}$$

Predictor variables:
Age, Conditions, ...

Goal of (Statistical) Machine Learning

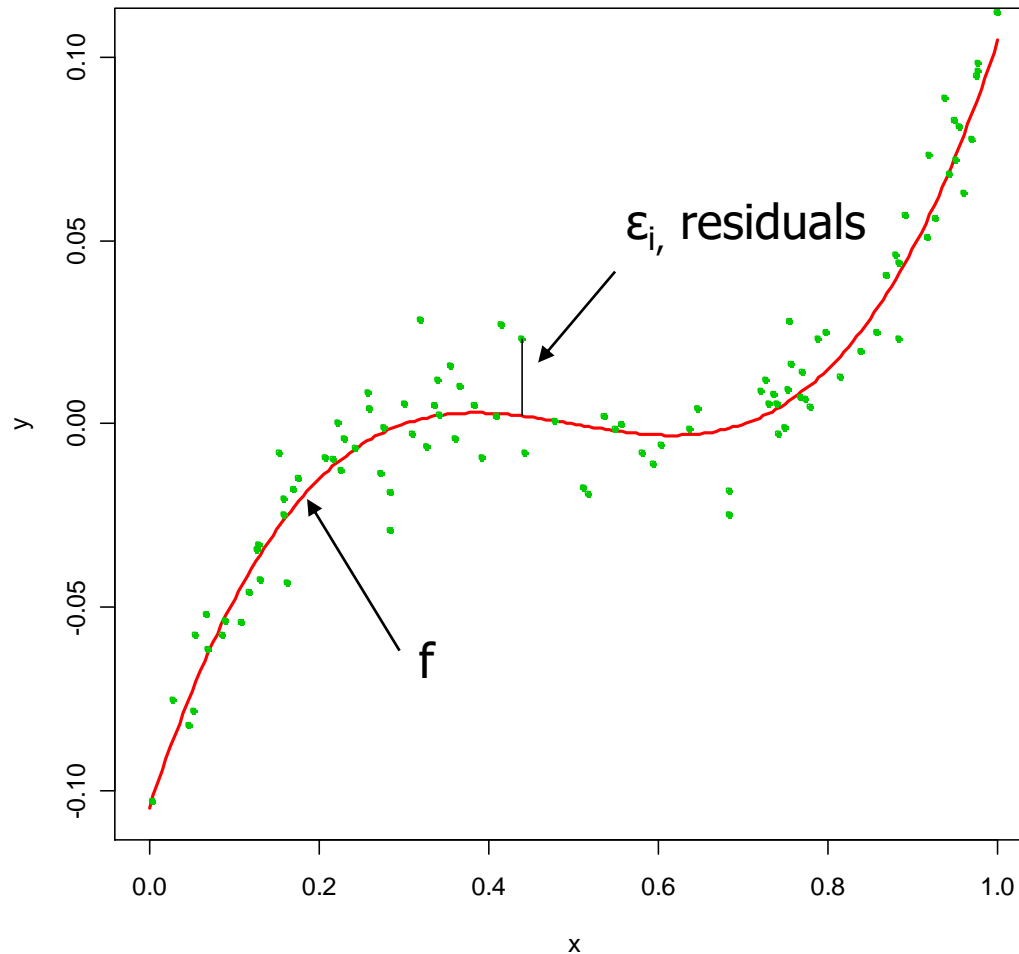
$$Y_i = f(X_i) + \epsilon_i$$

Unknown function
that we want to
learn

(Random)
measurement or
other errors

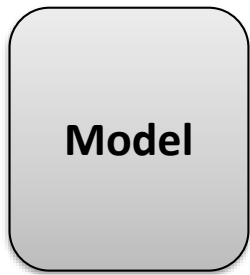


A Simple Example – Fitting a Function f





Types of Statistical Machine Learning Models



$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \end{pmatrix} = f \left(\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \end{pmatrix} \right) + \epsilon_i$$

Supervised Learning

- We know / are given Y as observed outcomes
- Goal: finding function f that most closely allows us to predict Y
- Subtypes:
 - Regression
 - Classification

Unsupervised Learning

- We don't know / are not given Y
- Goal: gaining a better understanding of the relationships of the X variables
- Subtypes:
 - Clustering
 - Factor Analysis / PCA
 - Association Rules



Supervised Machine Learning: Two Main Tasks

1

Inference

We have a particular variable (called the target variable) and we want to learn how the target variable depends on the other variables / predictors.

- Goal: Understand which variables influence the target variable
- Example: which variables significantly influence the probability of getting a disease?
- Focus on models that provide insights into variable 'importance'

2

Prediction

We have a particular variable (the target variable) and we want predict it as closely as possible for our / new data.

- Goal: Minimize the prediction error
- Example: minimize the false positive and false negative rate in disease prediction
- Models with high predictive power are often more complex than models used for inference



Regression Models

Model

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \end{pmatrix} = f \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \end{pmatrix} + \epsilon_i$$

Y is quantitative

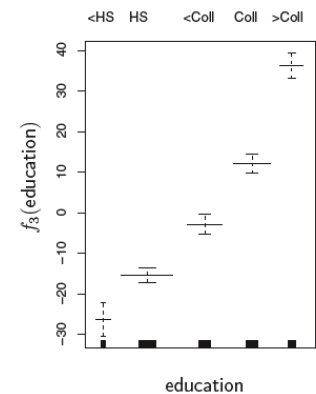
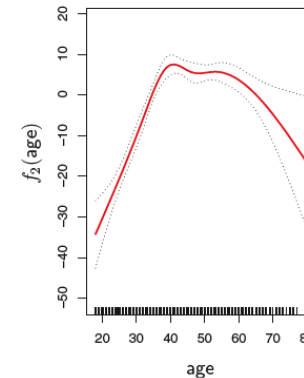
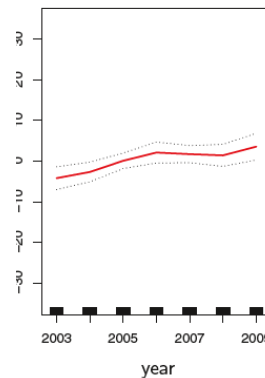
f can be linear or nonlinear

Example

Goal: Predict numerical Y

Sample Models:

(Non-)linear regression,
Splines, Regression Trees,
Generalized Additive
Models, etc.





Use Case Example: Predicting Infections

Scenario

- Monitoring of current health threat
- Both short-term and long-term factors affect the future rate of infections
- Emergency planning requires a good estimate of future infections for resource planning

Goal

- Build a regression model that takes into account all relevant (and known) factors
- Estimate the quality of the model on previous forecasts

Methods and Challenges

- Model Types: Multiple linear regression, Advanced (non-linear) regression, splines, etc.
- Estimating all relevant potentially unknown factors can lead to high uncertainty of the prediction



Classification Models

Model

$$Y = \begin{pmatrix} Yes \\ No \\ \dots \end{pmatrix} = f \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \end{pmatrix} + \epsilon_i$$

Y is categorical,
often binary

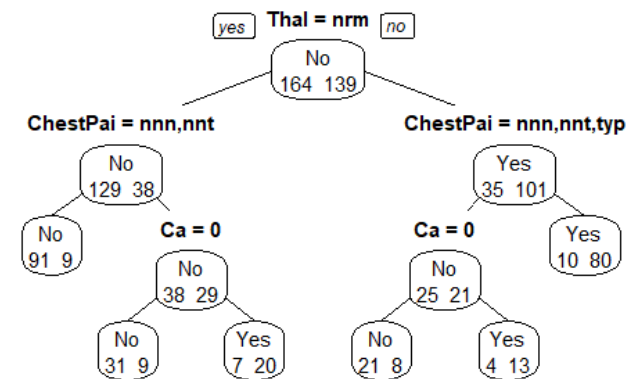
f usually nonlinear

Example

Goal: Predict binary/ categorical outcome

Algorithms usually give an additional probability estimate

Examples: Logistic Regression, Decision Trees, Boosted / Bagged Trees, Support Vector Machines, etc.





Use Case Example: Disease Incidence Prediction

Scenario

- Understanding of influencing factors that make it more likely to develop a disease
- Better understanding can help with individual predictions, risk scores, and better treatment options
- Being able to identify high-risk patients allows for a better management of resources

Goal

- Predict the likelihood that a patient is going to develop a specific disease
- Follow-up with specific treatment options for low risk vs high risk patients

Methods and Challenges

- Model Types: decision trees (random forests), logistic regression, SVMs, etc.
- Other relevant health factors (e.g., comorbidities) need to be taken into account



Unsupervised Learning

Model

$$\begin{pmatrix} x_{11} & x_{12} & x_{1p} \\ x_{21} & x_{22} & \dots x_{2p} \\ \dots & \dots & \dots \end{pmatrix}$$

Clustering

Factor Analysis,
Principal Component Analysis

Example

**Factor Analysis,
Principal Component
Analysis:**

Understand relationship
between columns.

Example: Comorbidities

Clustering:

Identify observations with similar
characteristics

Example: Patients with similar health
conditions / history

Methods: k-means, hierarchical



Use Case Example: Identifying Patient Groups with Similar Health Patterns

Scenario

- Based on clinical measurements and general health, patients can be grouped into different categories: healthy, hypertension, diabetes, etc.
- Different patient groups have different needs and treatment options: certain combinations of conditions / measurements can occur more frequently in certain groups

Goal

- Identify and model different patient groups based on their health patterns and conditions
- The identified groups can then be managed individually with custom treatments

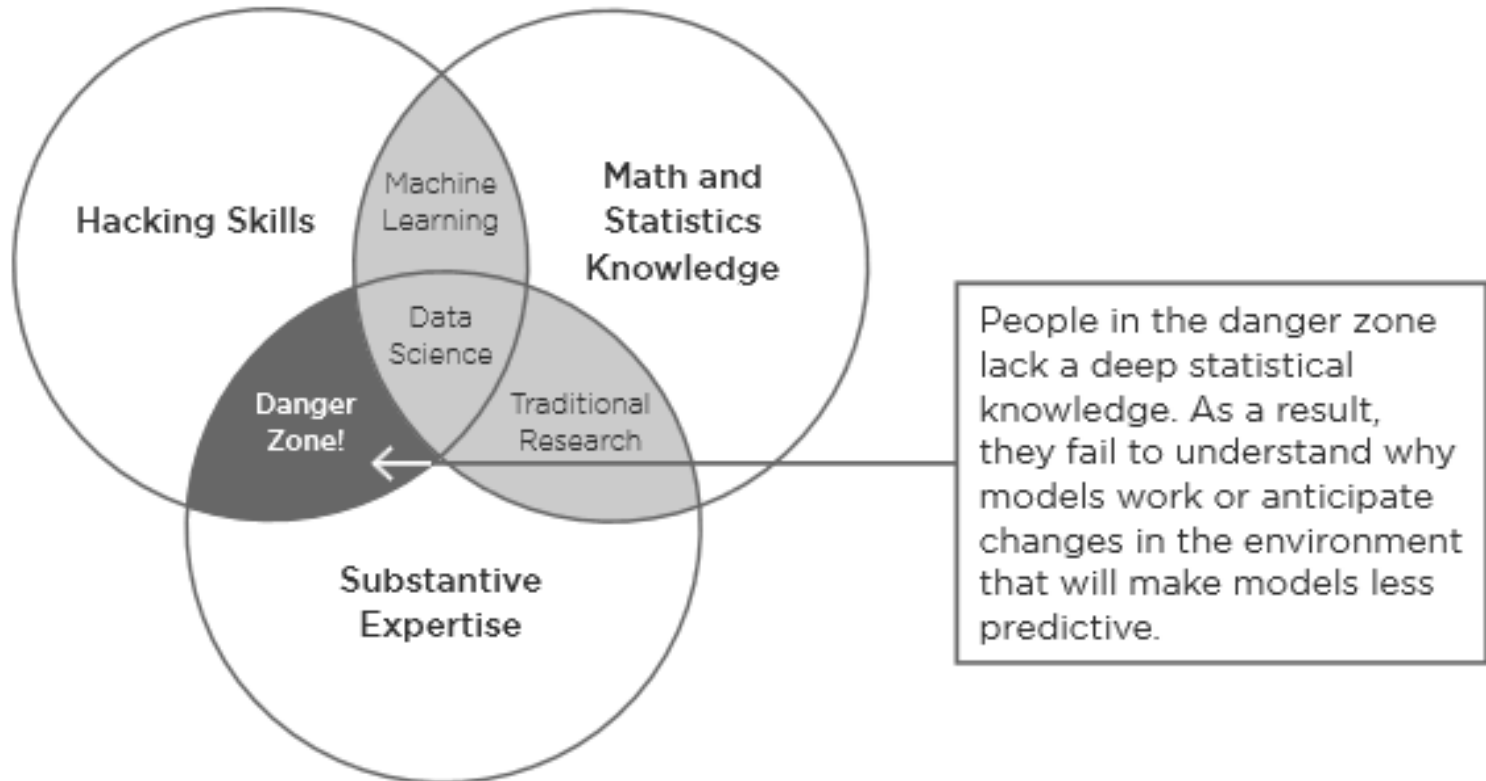
Methods and Challenges

- Model Types: k-means or hierarchical clustering.
- Patients can switch between groups over time, hence models need to be re-evaluated after a while



Data Science / Machine Learning – What Skills are needed?

Data Science: A Popular Definition



Source: Drew Conway.⁴



Agenda

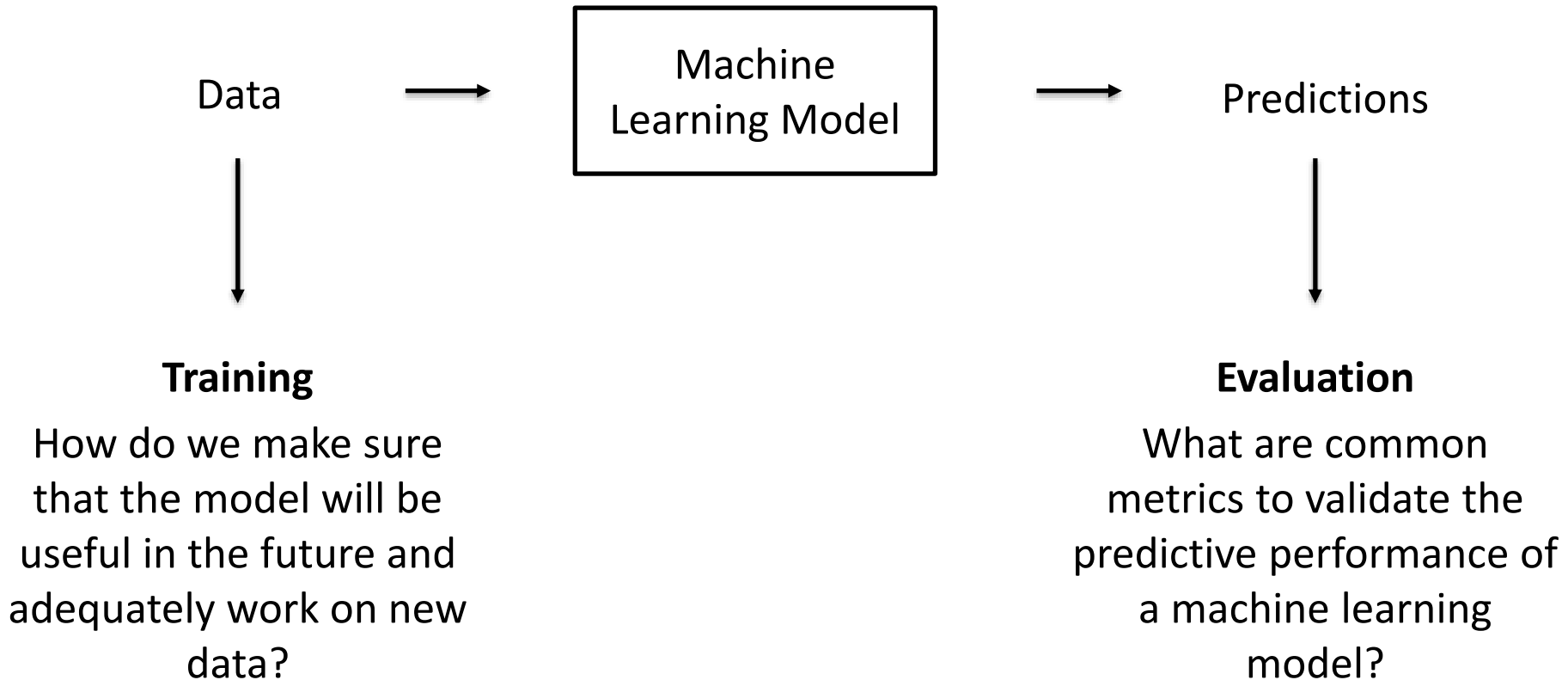
Part 1: Data Science and Main Types of Statistical Machine Learning Models

Part 2: Evaluation of ML models: how do we know if our model is working?

Part 3: Explainability, Fairness, and other considerations



Model Validation





Evaluation of Models

Supervised Learning

For supervised learning models, we can compare their performance against the actual output:

$$y_i \text{ vs } \hat{y}_i$$

Regression

In Regression models, this is often done by calculating the Mean-Squared-Error:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Classification

In Classification models, the most basic metric is the error rate:

$$Error\ Rate = \sum_{i=1}^n I(y_i \neq \hat{y}_i) / n$$



How to Define Fairness Metrics in Classification Algorithms

Confusion Matrix
defines common
performance metrics

Accuracy: $\frac{TN+TP}{All\ Predictions}$

Error Rate:
 $1 - Accuracy$

**Area under the curve
(AUC)**

		True Class	
		Negative	Positive
Predicted Class	Negative	True Negative Count (TN)	False Negative Count (FN)
	Positive	False Positive Count (FP)	True Positive Count (TP)

Specificity: $\frac{TN}{TN + FP}$

Sensitivity, Recall: $\frac{TP}{TP + FN}$

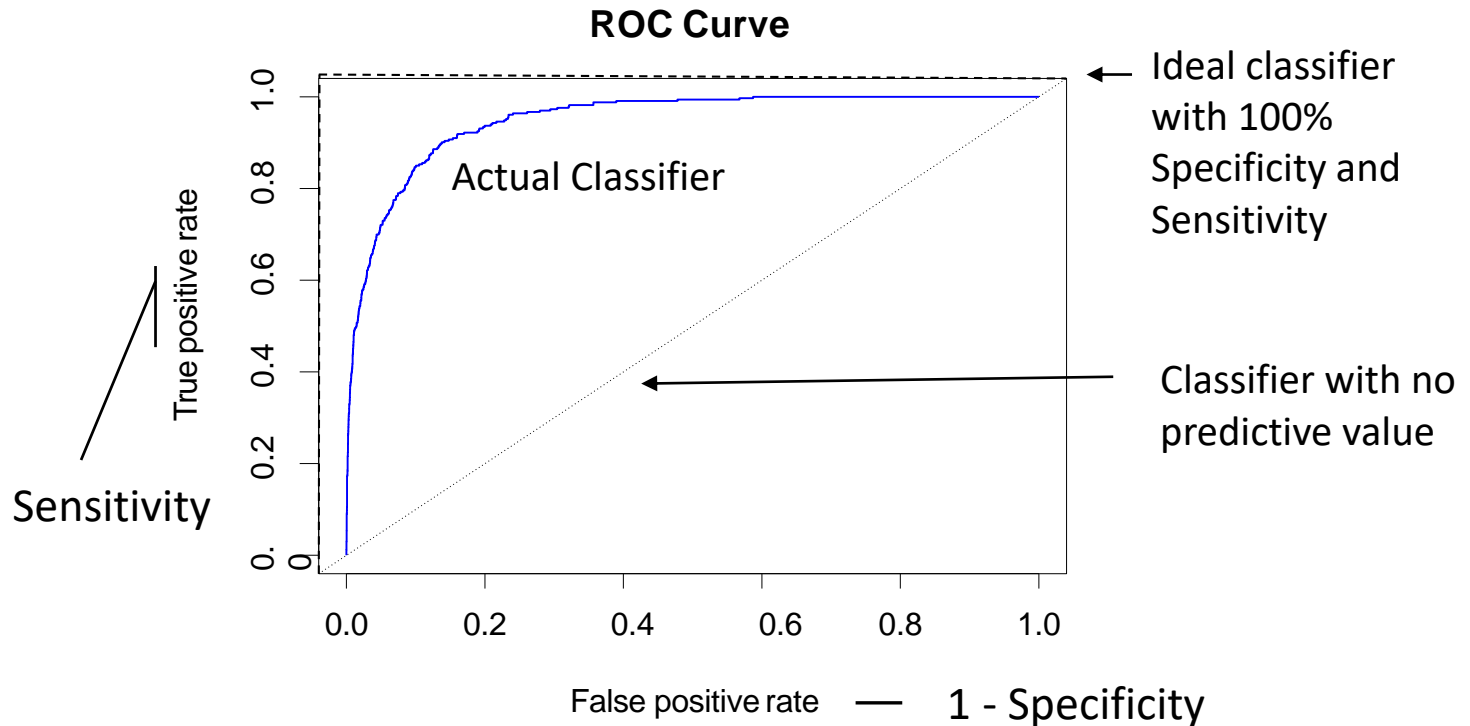
Precision:
 $\frac{TP}{TP + FP}$



AUC and Kappa

AUC

Kappa



Metric how the classifier compares against a purely random classifier
0 indicates random, 1 indicates 'perfect' classification



Classification – Which metric to use?

Example

Rare outcomes (e.g., rare disease)

Number of people who don't have the disease: 9990

Number of people who have the disease: 10

Simple Model

Always predict 'No disease'

Accuracy: 99.9%

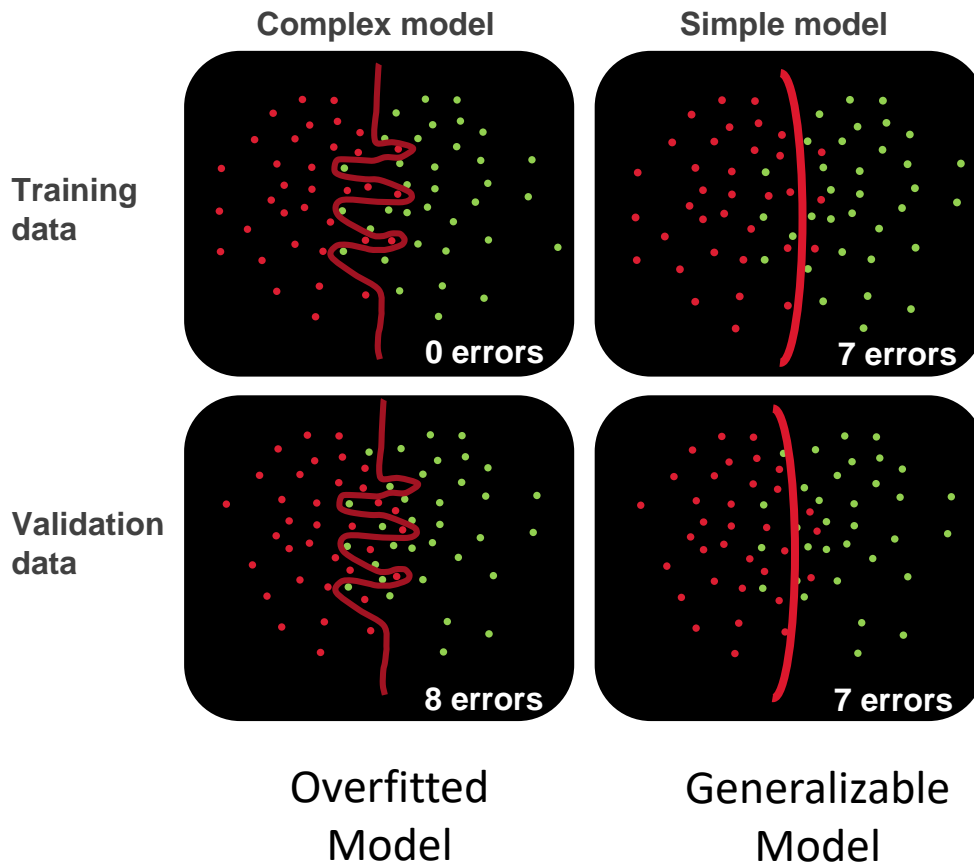
Sensitivity: 0%



In many cases, AUC (Area under the ROC curve), F1 Score (based on sensitivity and specificity), or kappa scores are used instead of accuracy



Using Model Validation to Prevent Overfitting



A complex model can often be 'too flexible' and learn irrelevant relationships from the data

An **overfitted model** is very susceptible to even small changes in the data and thus might perform worse than a simple model



Common Methods for Model Validation

Training and Test Sets

Split the dataset into 2 parts



Build model on training data, evaluate on test data



Training Data

Testing Data

K-fold Cross Validation

Split data into k parts



Use k-1 parts for training, 1 for evaluation.



Repeat k times and average results



Agenda

Part 1: Data Science and Main Types of Statistical Machine Learning Models

Part 2: Evaluation of ML models: how do we know if our model is working?

Part 3: Explainability, Fairness, and other considerations



Feature Selection – Which Variables should be Included in the Model?

Model

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \end{pmatrix} = f \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \end{pmatrix} + \epsilon_i$$

Not all features might be relevant for the model / improve the model performance

More features also mean more complex models

Filter Methods

Based on characteristics of each predictor, decide if it should be included or not

Examples: variance, correlation

Wrapper Methods

Estimate the effect/ usefulness of subsets of predictors based on outcome

Examples: forward/backward selection

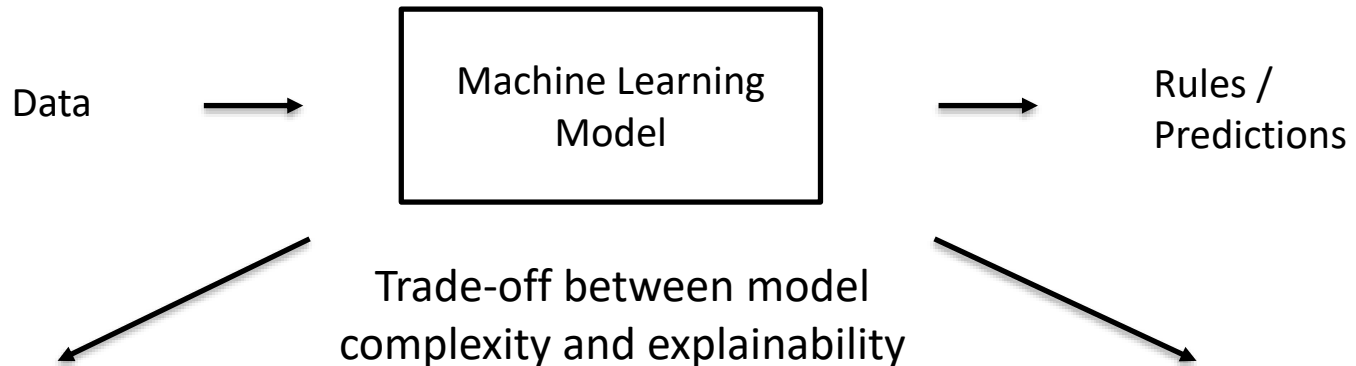
Embedded Methods

Variable selection embedded in the learning procedure, i.e., model-dependent

Example: LASSO



Explainable AI – Can the Model Predictions be understood by (Human) Experts?



Black Box Models



'Explainable' Models

Complex, potentially multi-layered functions

High predictive quality

Learned rules are not easy to understand

Example: Deep Neural Networks

Functions of various complexities

Provide the ability to 'understand' why an observation received a specific prediction

Example: Decision Trees, Logistic Regression

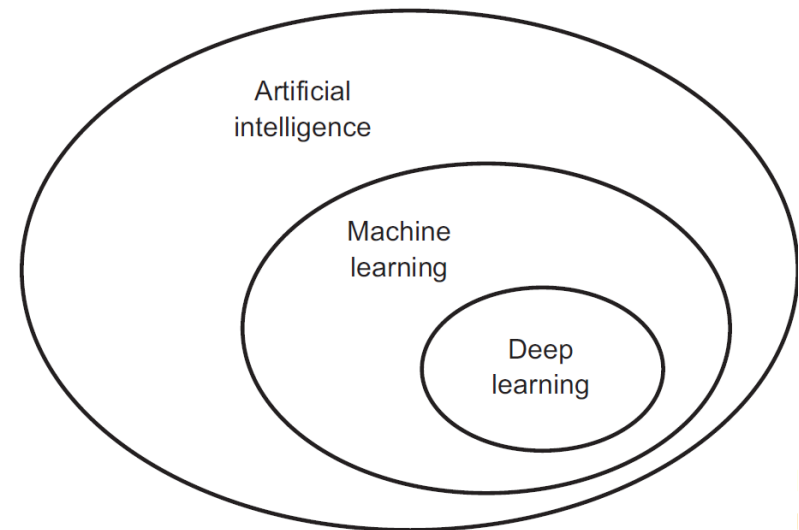


Machine Learning vs Deep Learning

Machine Learning and Deep Learning are closely related

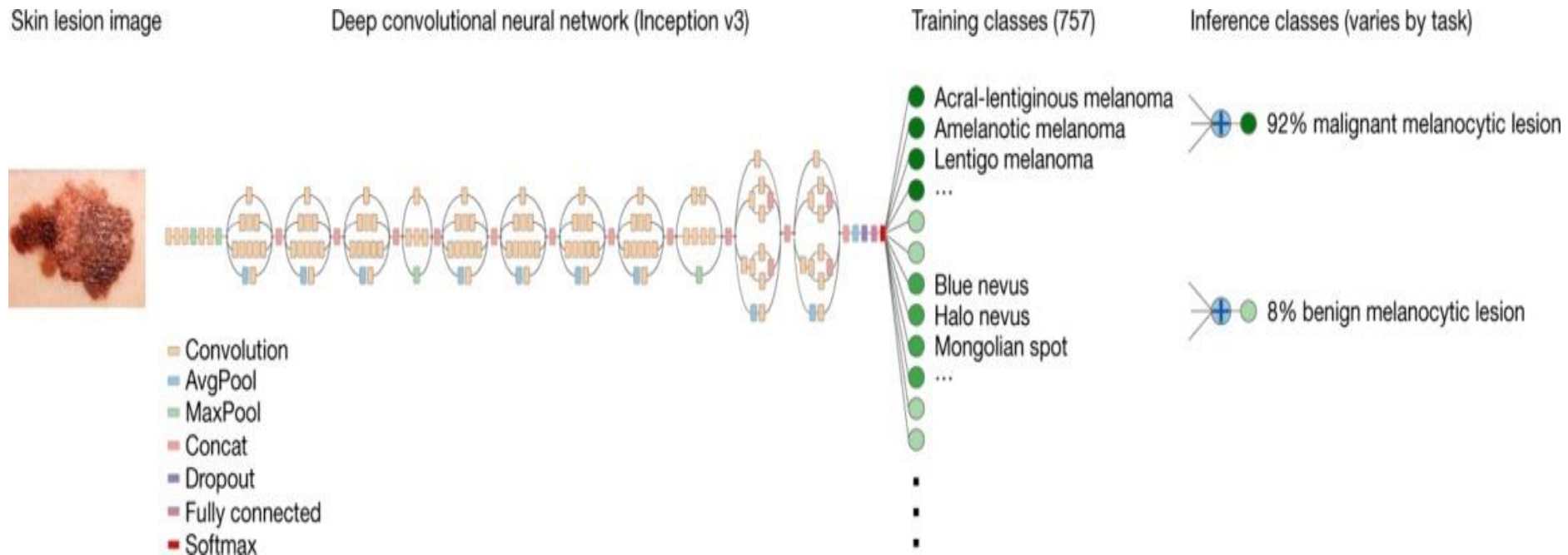
Deep Learning is a subfield of machine learning

‘Deep’ only means that multiple subsequent layers of data representations are used. It does not infer a deeper knowledge or understanding!





A Deep Learning Example





Algorithmic Fairness – Definition and Overview

Algorithms are increasingly used in everyday decision making



With this increased usage, concerns about *potential discrimination* arise



Recent development of open source tools to detect and mitigate biases, e.g. IBM 360 Algorithmic Fairness tool

Algorithmic Fairness studies the definition, identification, mitigation, and prevention of discrimination and bias in algorithm-based decision making



Algorithmic Fairness – Examples

Text Analytics

Word associations:

'she' -> 'nurse'

'he' -> 'doctor'

Criminal Recidivism

Predicted probabilities:

$\Pr(\text{recommit crime} \mid \text{Group 1}) >$

$\Pr(\text{recommit crime} \mid \text{Group 2})$

AI-based Job Recruitment

Selection of resumes:

'male' -> yes

'female' -> no



How to Define Fairness Metrics in Classification Algorithms

Discriminatory,
'Protected' Attribute

User	Gender	Income	...
1	F
2	M
...

One (or more) attributes are used to define two (or more) groups that should be treated fairly

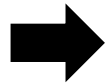
Fairness metrics compare the outcome for the two (or more) groups



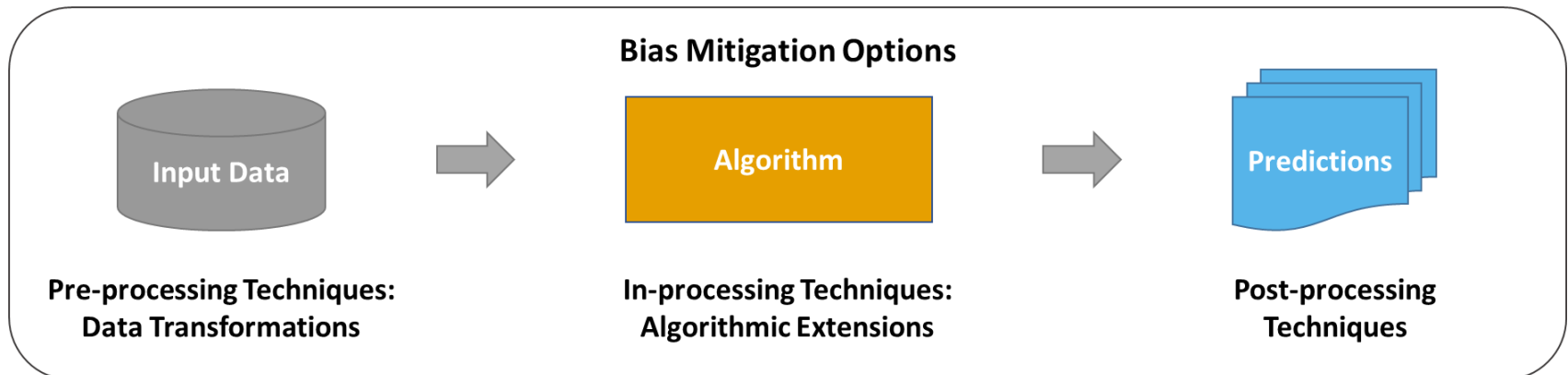
Bias Mitigation Strategies – How do we address the biases?

Simple Approach

Why not simply deleting / not using protected attribute(s) in prediction models?



Research shows that this is not sufficient and does not prevent discrimination





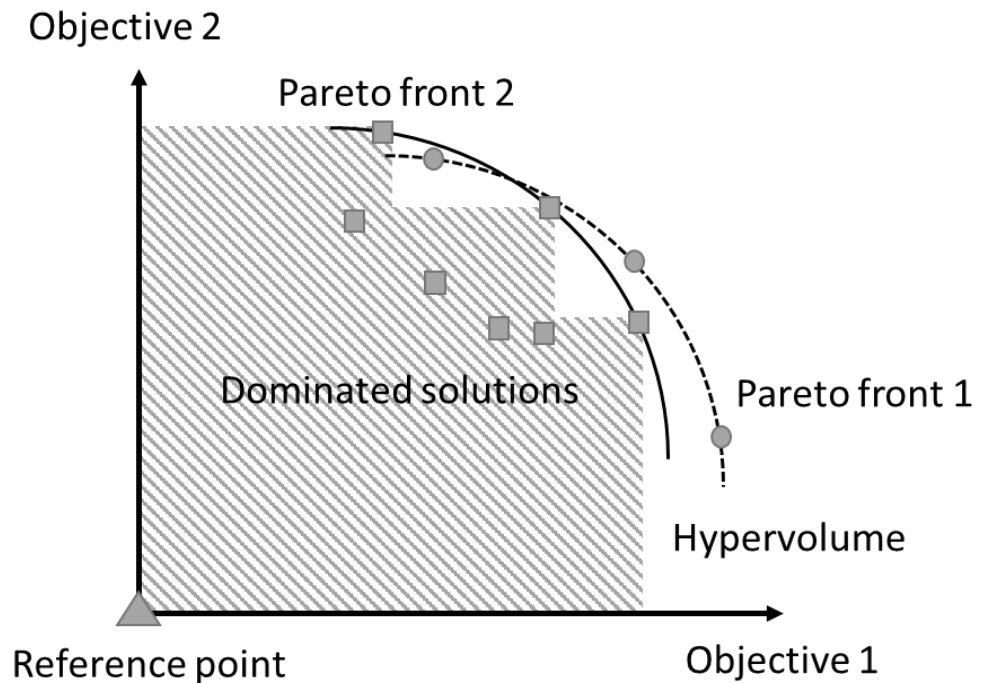
Analyzing the Effect of Fairness on Classification Performance

Effect of fairness improvements

What happens if we try to increase the fairness of a solution?

Specific *trade-offs* between fairness and 'performance' depend on algorithm, bias mitigation strategy, fairness metric, etc.

Captured by *Pareto fronts*



Pareto front based Trade-off Analysis
[Haas, 2019]



Thank you!