Day 3: Artificial intelligence for policy-making

Machine learning 101

Simon Munzert
Hertie School

Table of contents

Types of data-driven research and their role for policy

Hertie School

1. Description

- What is the state of the world?
- What are the trends over time?
- What are the differences between groups?

2. Explanation

- What is the effect of a policy?
- Does the effect vary across groups?
- What are the mechanisms behind the effect?

3. Prediction

- What is the path of an indicator?
- (When) will future events happen?
- What class does this observation most likely belong to?

The value for policy-making

- At the center of **monitoring**
- "How many people consume misinformation online?"
- "How many people are unemployed in a certain district?"
- "How does the distribution of income vary across educational segments of the population?"

The value for policy-making

- At the center of **evaluation**
- "Did the minimum wage increase lead to a decrease in employment?"
- "Did the campaign affect the exposure to misinformation differently across groups?"
- "Why did the intervention not lead to the expected results?"

The value for policy-making

- At the center of **forecasting** but also **targeting** and **measurement**
- "Will there be conflict?"
- "How many people will be unemployed in a certain district next year?"
- "Which individuals are most likely to be affected by a policy?"

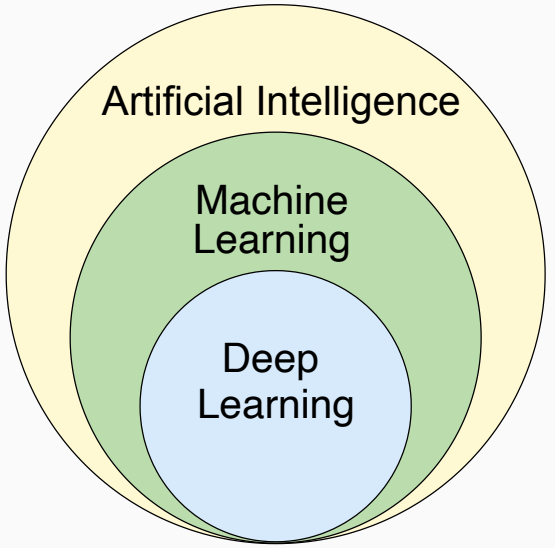
Machine learning, deep learning, AI

What is AI?

Artificial intelligence

> "Artificial intelligence (AI) is intelligence - perceiving, synthesizing, and inferring information - demonstrated by machines, as opposed to intelligence displayed by non-human animals and humans. Example tasks in which this is done include speech recognition, computer vision, translation between (natural) languages, as well as other mappings of inputs."

Wikipedia, *Artificial intelligence*



`Source` Wikipedia

> "The effort to automate intellectual tasks normally performed humans."

Chollet and Allaire, 2018, *Deep Learning with R*

What is AI?

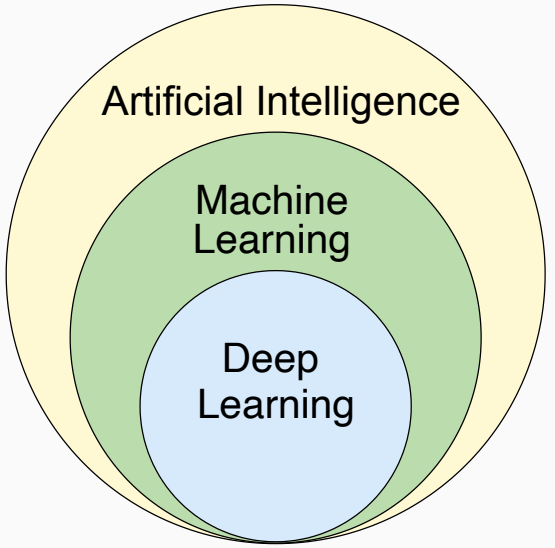
Machine learning

> "Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn' (...) It is seen as a part of artificial intelligence."

Wikipedia, *Machine learning*

> "Machine learning is a specific subfield of AI that aims at automatically developing programs (called models) purely from exposure to training data. This process of turning models data into a program is called learning."

Chollet and Allaire, 2018, *Deep Learning with R*



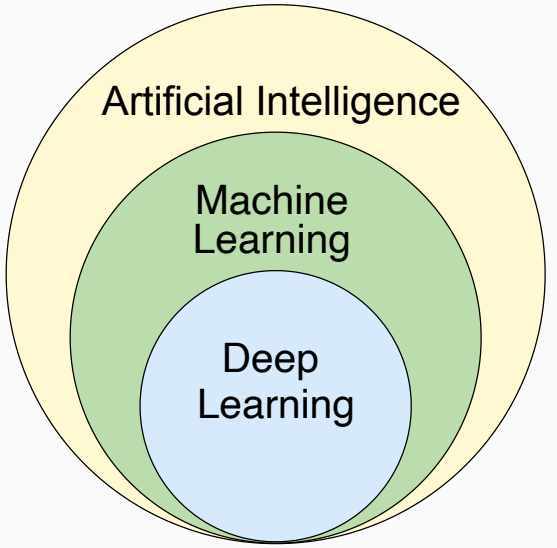
`Source` Wikipedia

What is AI?

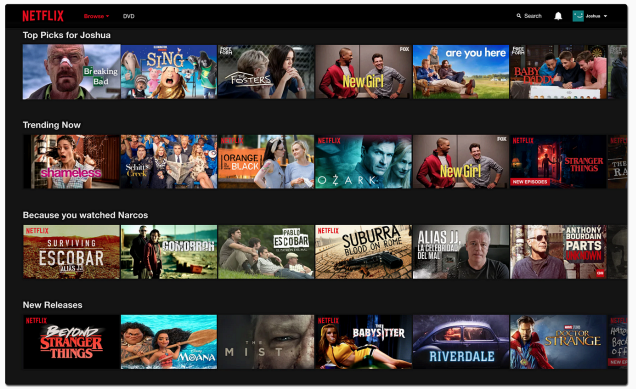
Data mining

> "Application of machine learning methods to large databases is called data mining. The analogy is that a large volume of earth and raw material is extracted from a mine, which when processed leads to a small amount of very precious material; similarly, in data mining, a large volume of data is processed to construct a simple model with valuable use, for example, having high predictive accuracy."

Alpaydin, 2014, *Introduction to Machine Learning*



`Source` Wikipedia

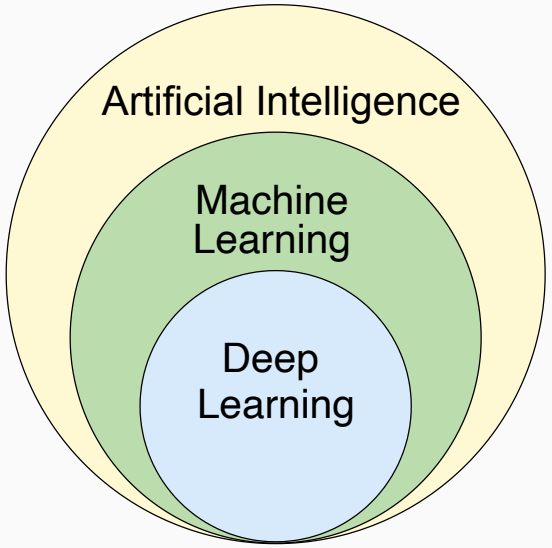


What is AI?

Deep learning

> "Deep learning is the subset of machine learning methods based on neural networks with representation learning. The adjective "deep" refers to the use of multiple layers in the network."

Wikipedia, *Deep learning*



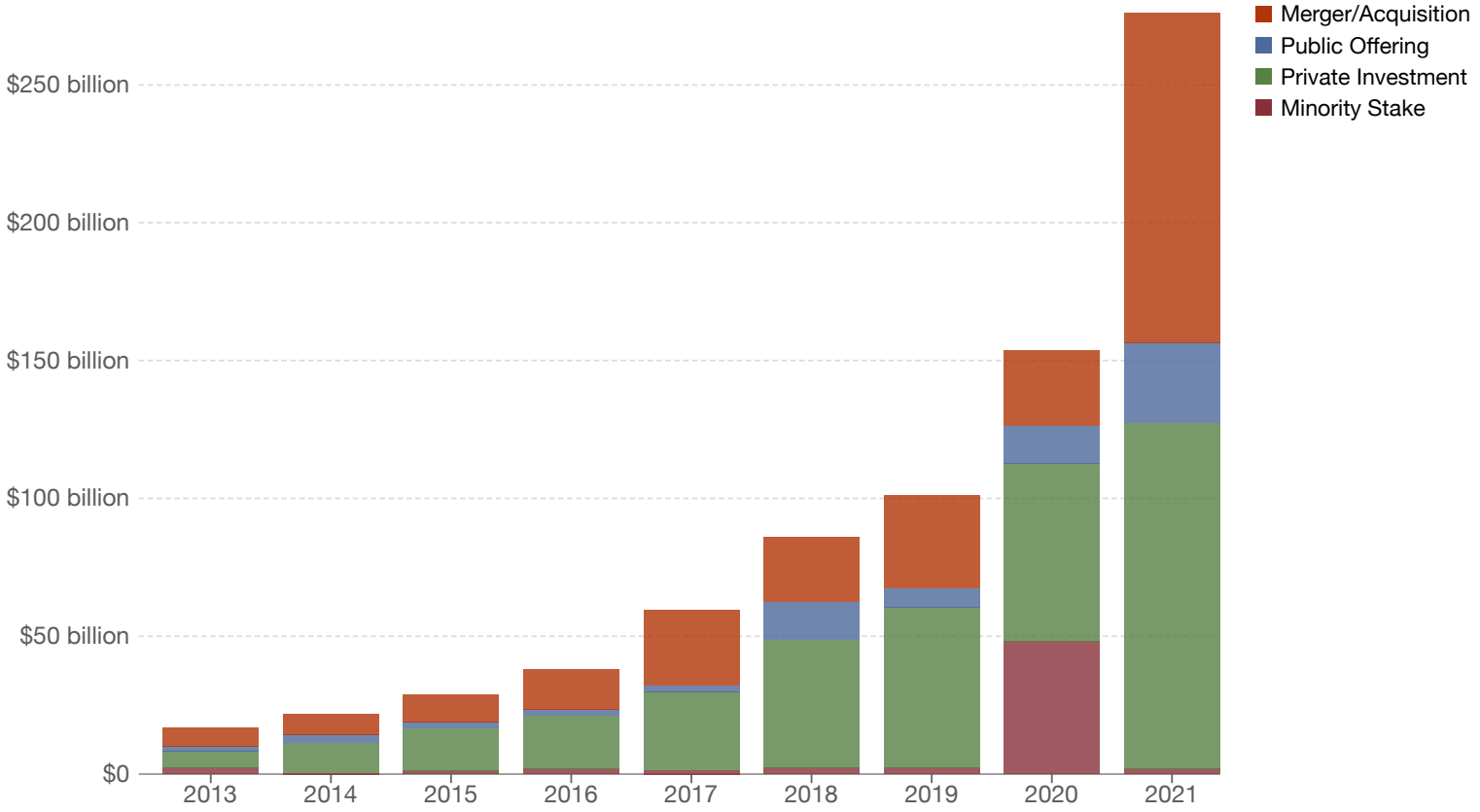
`Source` Wikipedia

Corporate investment in AI

Hertie School

Annual global corporate investment in artificial intelligence, by type

This data is expressed in US dollars, adjusted for inflation.



Data source: NetBase Quid via AI Index Report (2023) OurWorldInData.org/artificial-intelligence | CC BY

Note: Data is expressed in constant 2021 US$. Inflation adjustment is based on the US Consumer Price Index (CPI).

Basic concepts in machine learning

Regression vs. classification

Regression

- Predicts a continuous outcome
- Example: Predicting house prices, GDP growth, temperature

Classification

- Predicts a categorical outcome
- Example: Predicting whether a person will default on a loan, whether an email is spam, whether a patient has a disease

Classifcation problems in the wild

Classification problems occur often, perhaps even more so than regression problems, e.g.:

1. A woman arrives at the emergency room with a set of symptoms. Which condition does she have?
2. An online banking service must be able to determine whether or not a transaction is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.
3. On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Decision-making problems often are classification problems!

Supervised and unsupervised learning

Supervised learning

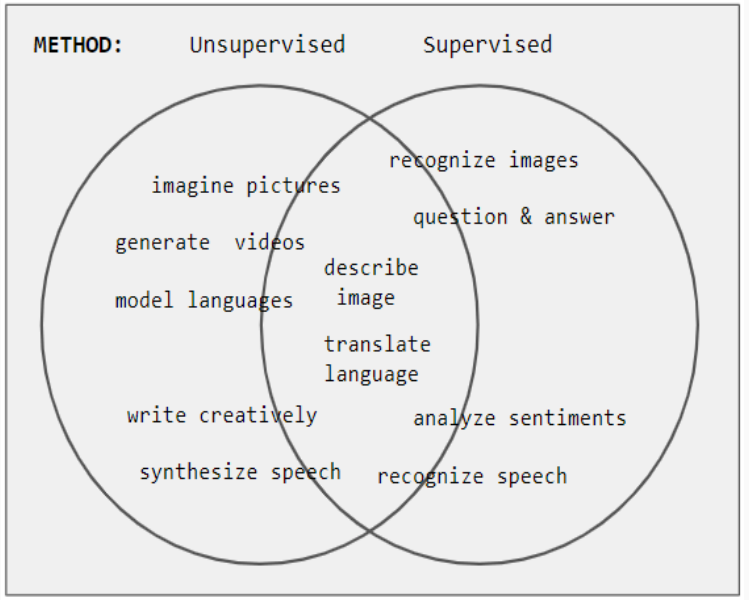
- The algorithm learns from labeled data, i.e., data with known outcomes
- The algorithm is trained on a training dataset and evaluated on a test dataset
- The goal is to the predict unobserved outcomes

Unsupervised learning

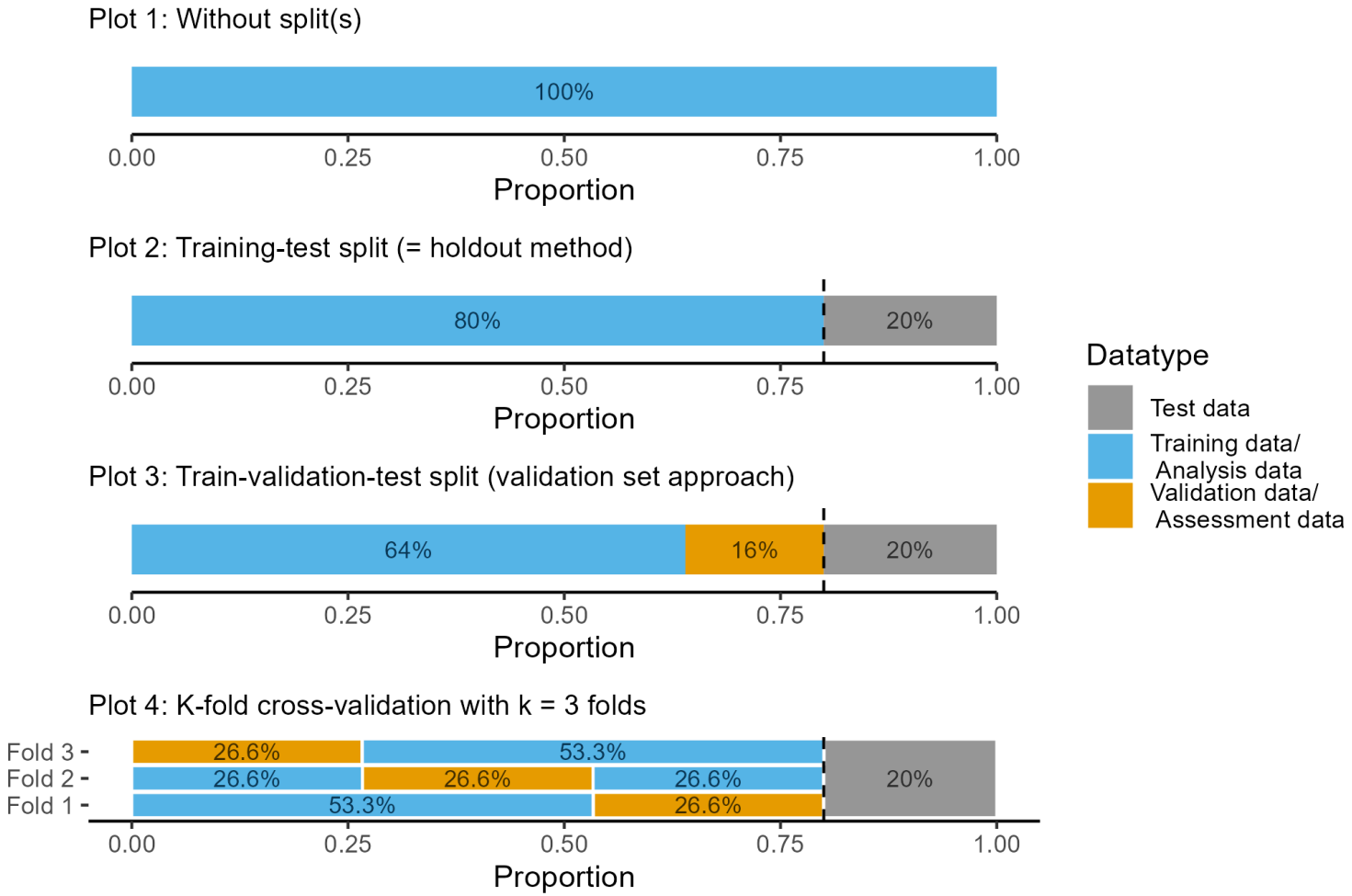
- The algorithm learns from unlabeled data
- There are inputs but no supervising output; we can still learn about relationships and structure from such data

Analogies

- Supervised: Child in school learns math (with teacher's input)
- Unsupervised: Child at home plays with toys (without teacher's input)

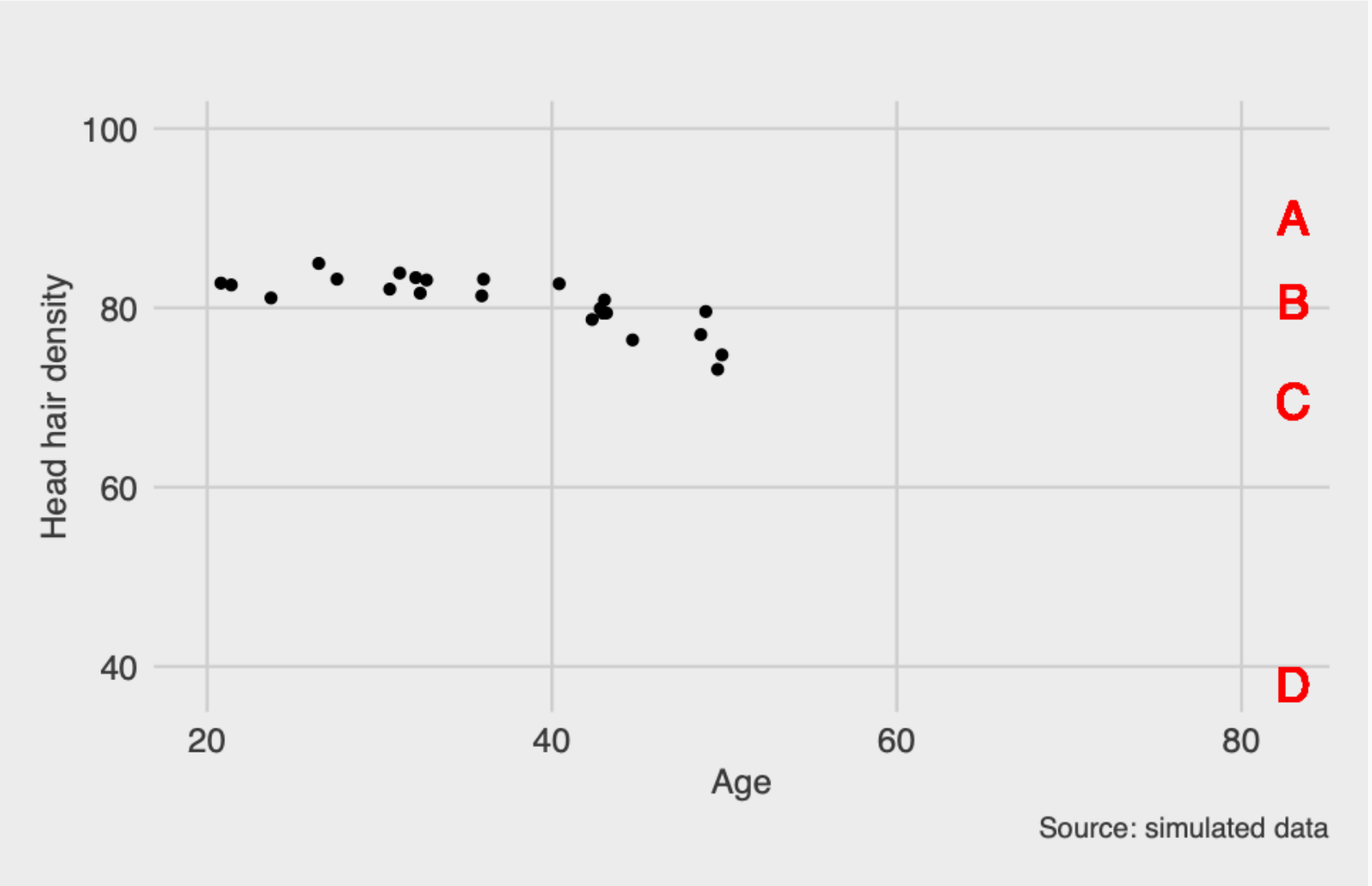


Training, validation and test dataset

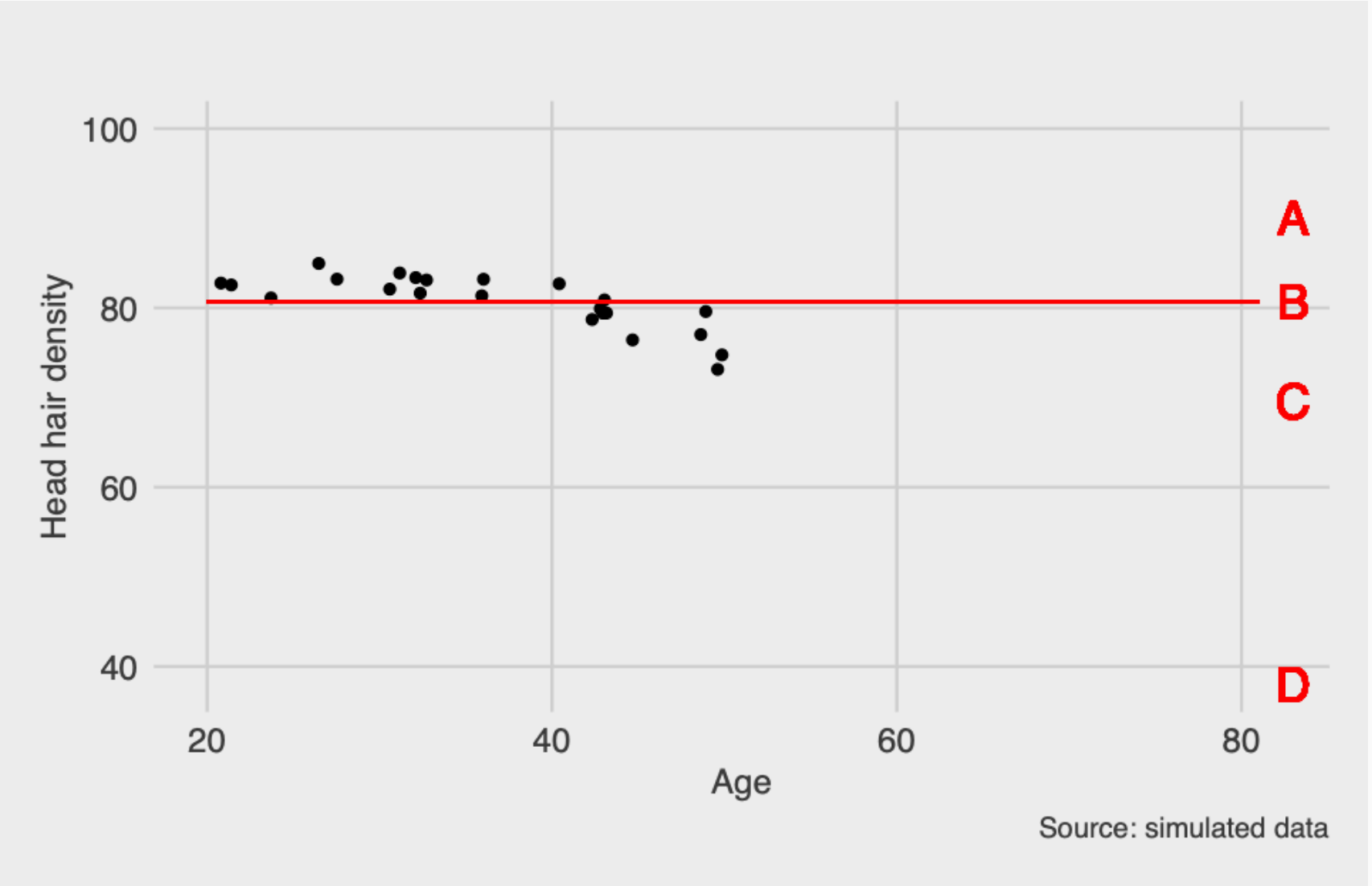


Note: Graph visualizes different evaluation protocols and what percentage of the data is used for training, validation and testing; Plot 3: Assigns 20% to the test data, and 20% of the remaining data (= 16% of overall) to the validation data and 20% of the remaining data to the training data (= 64% of overall); Plot 4: 3-fold cross-validation randomly splits the full training data (here 80%) 3 times, always keeping one third of the data for validation; © Paul C. Bauer

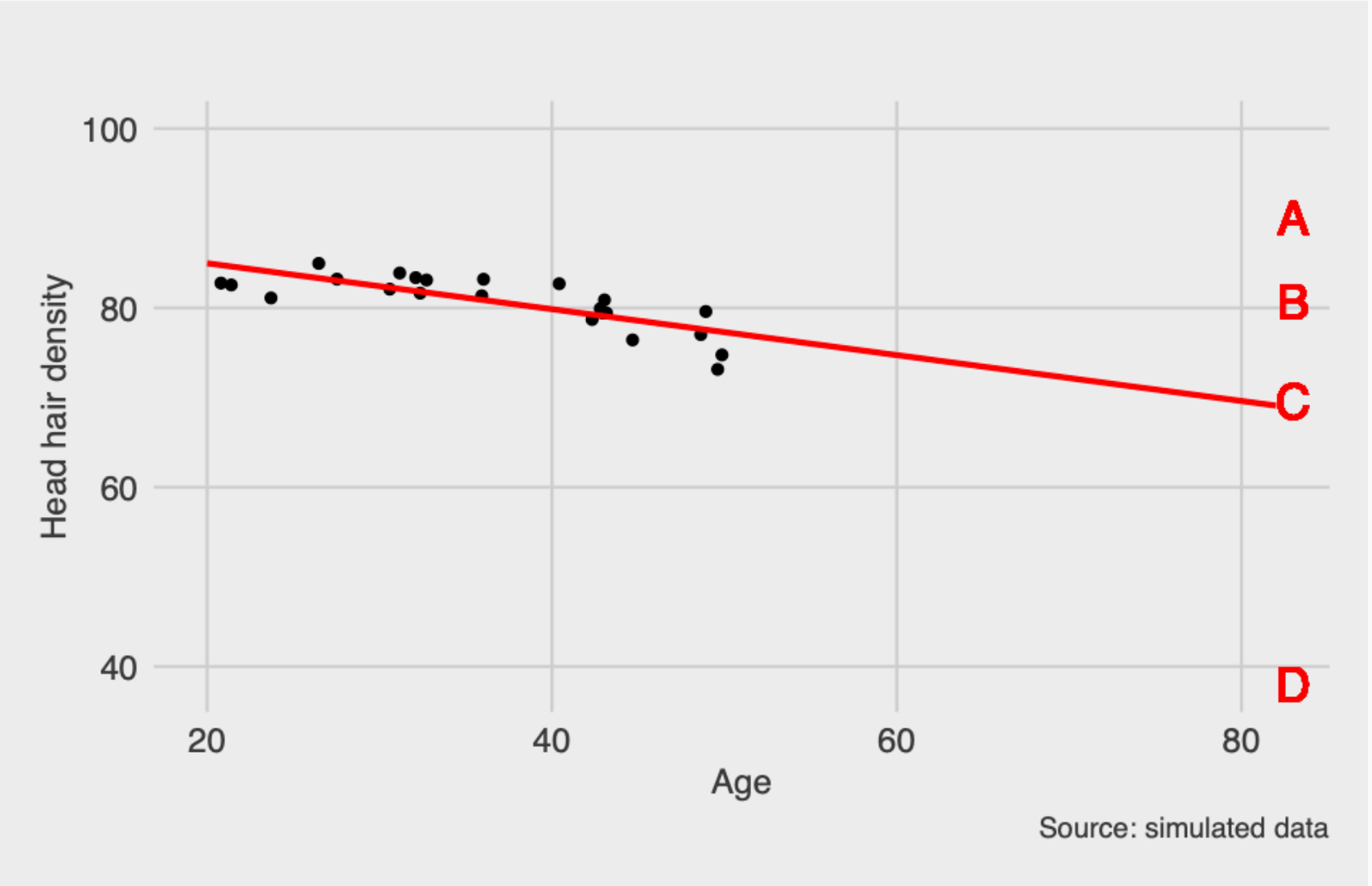
Overfitting



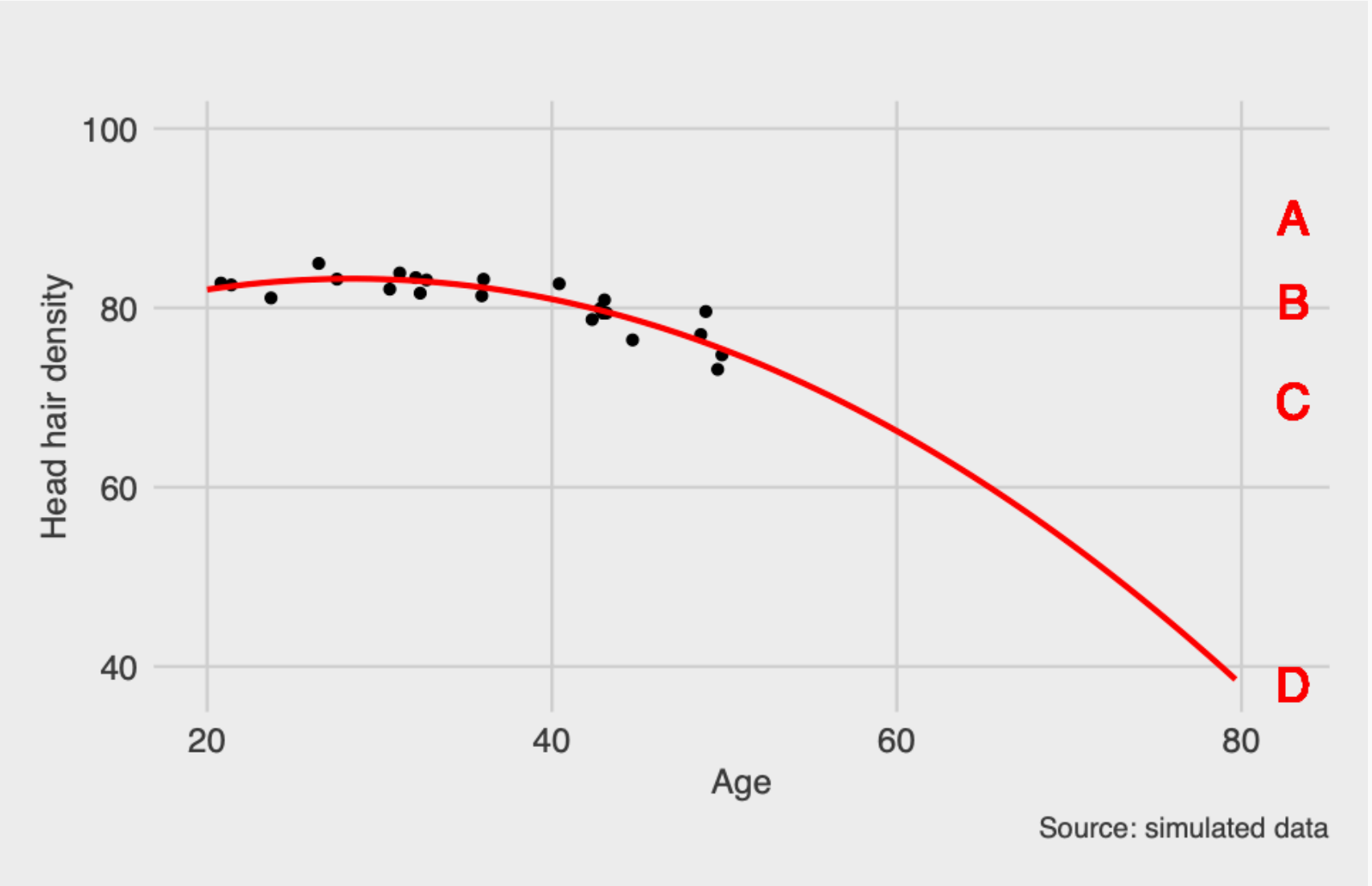
Overfitting



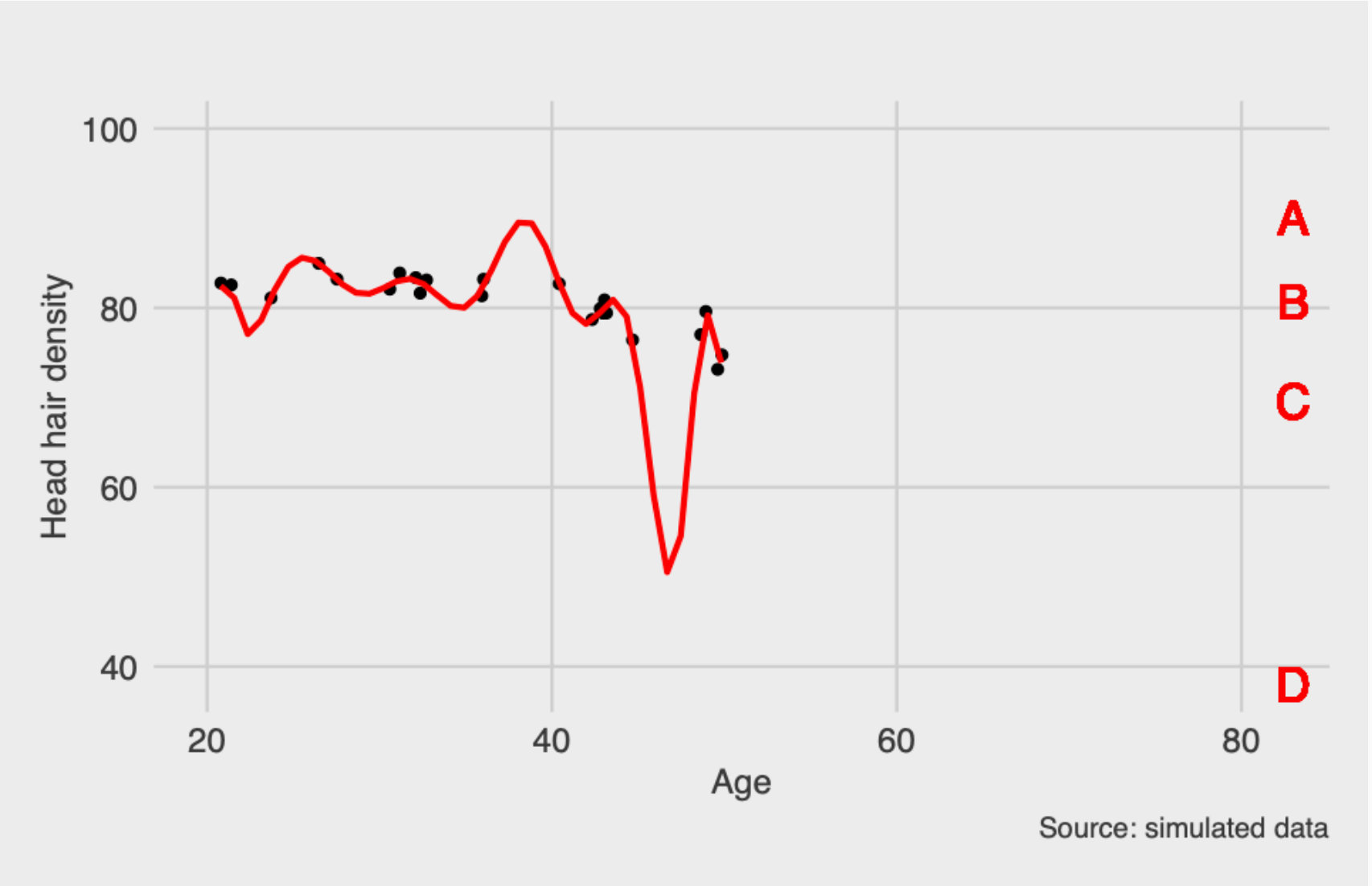
Overfitting



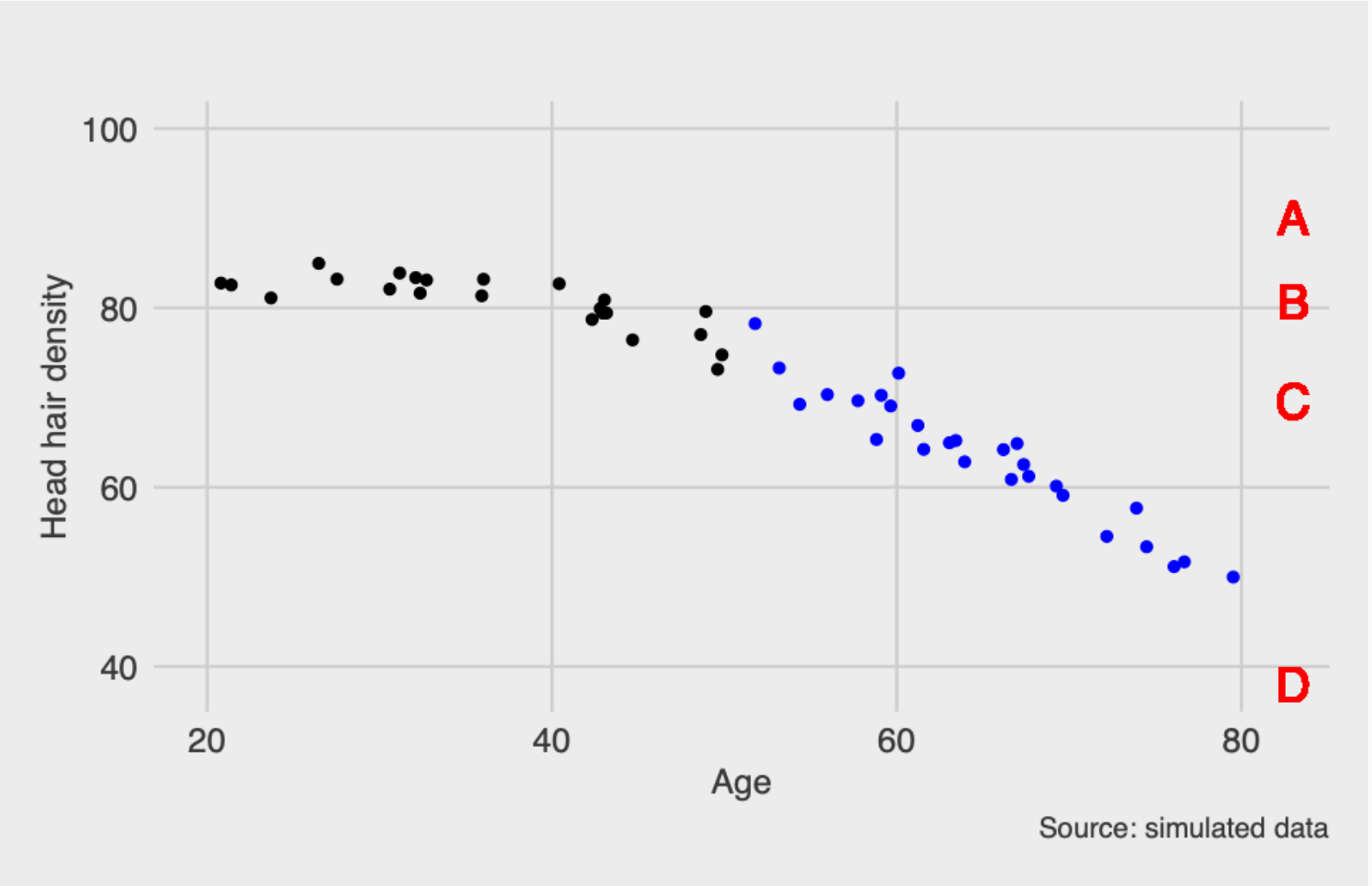
Overfitting



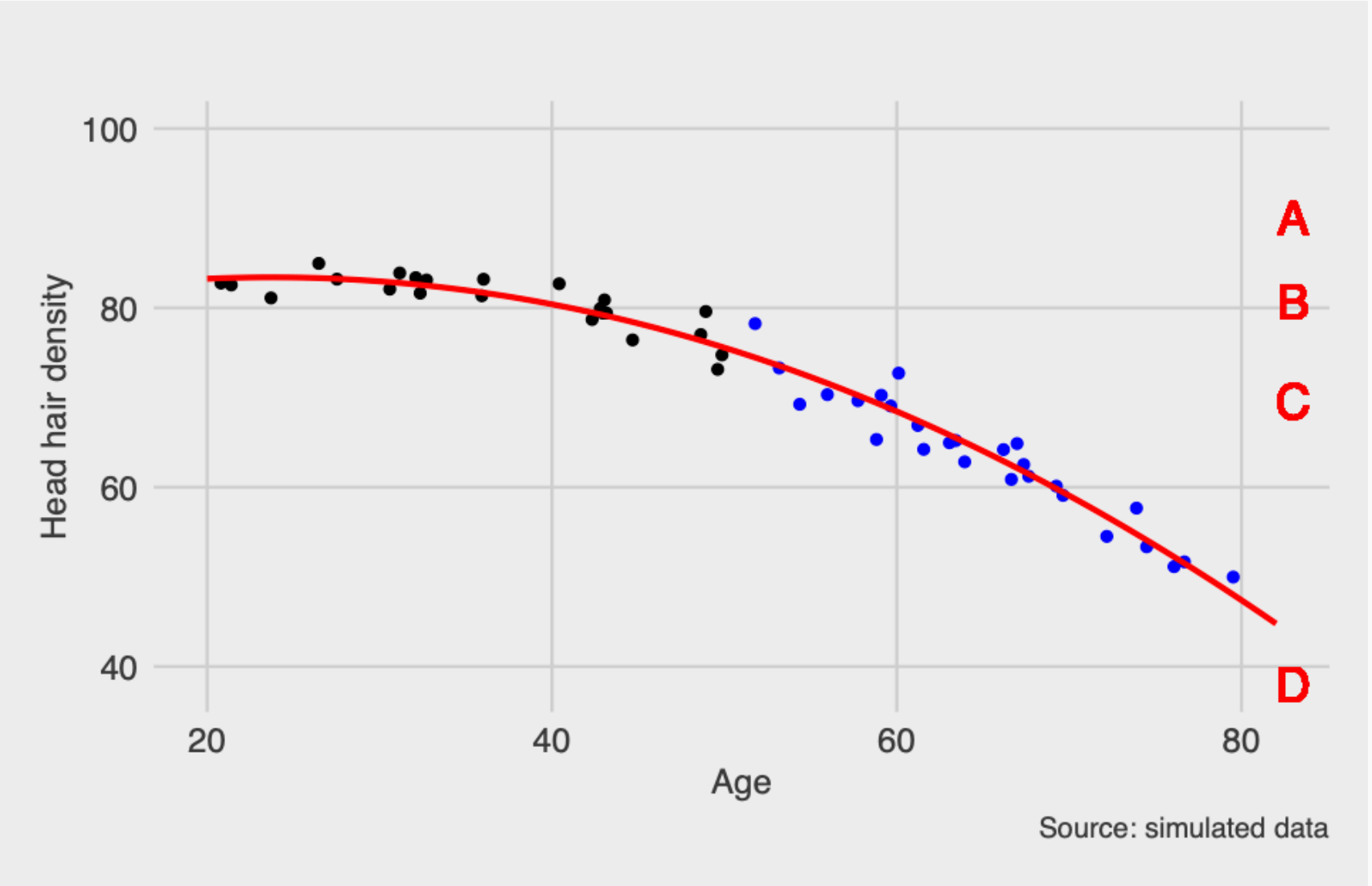
Overfitting



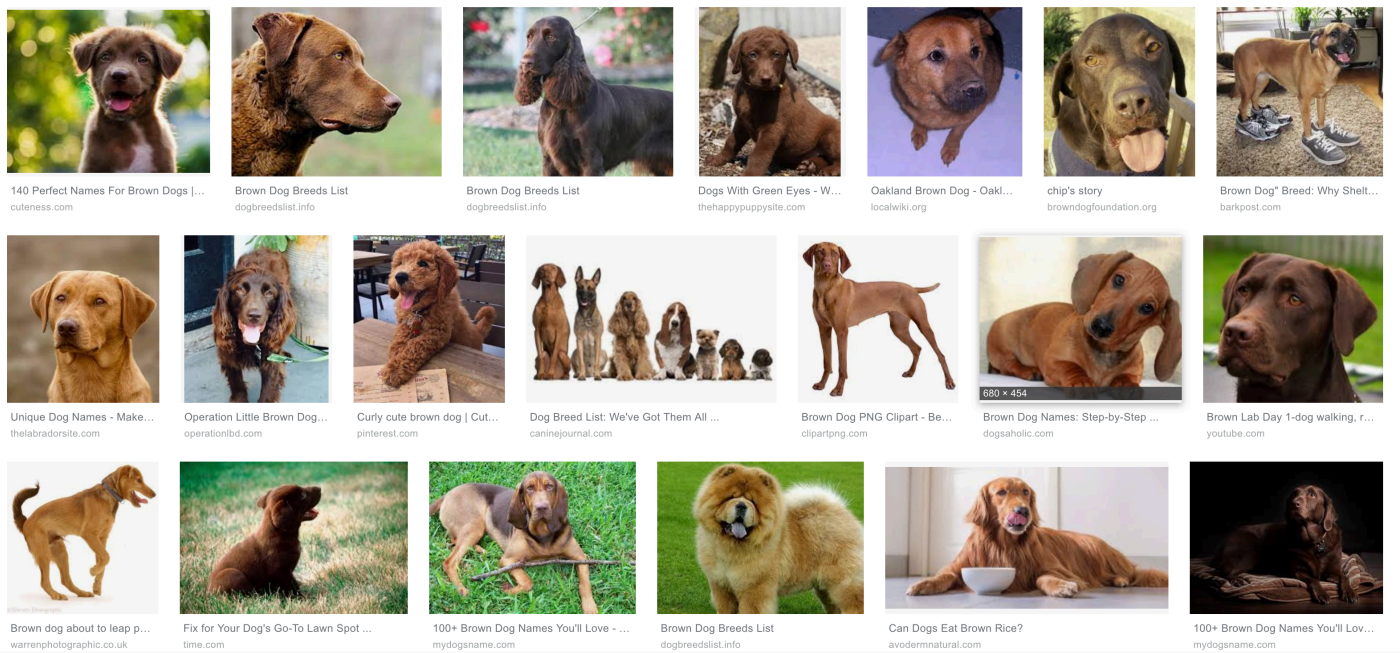
Overfitting



Overfitting

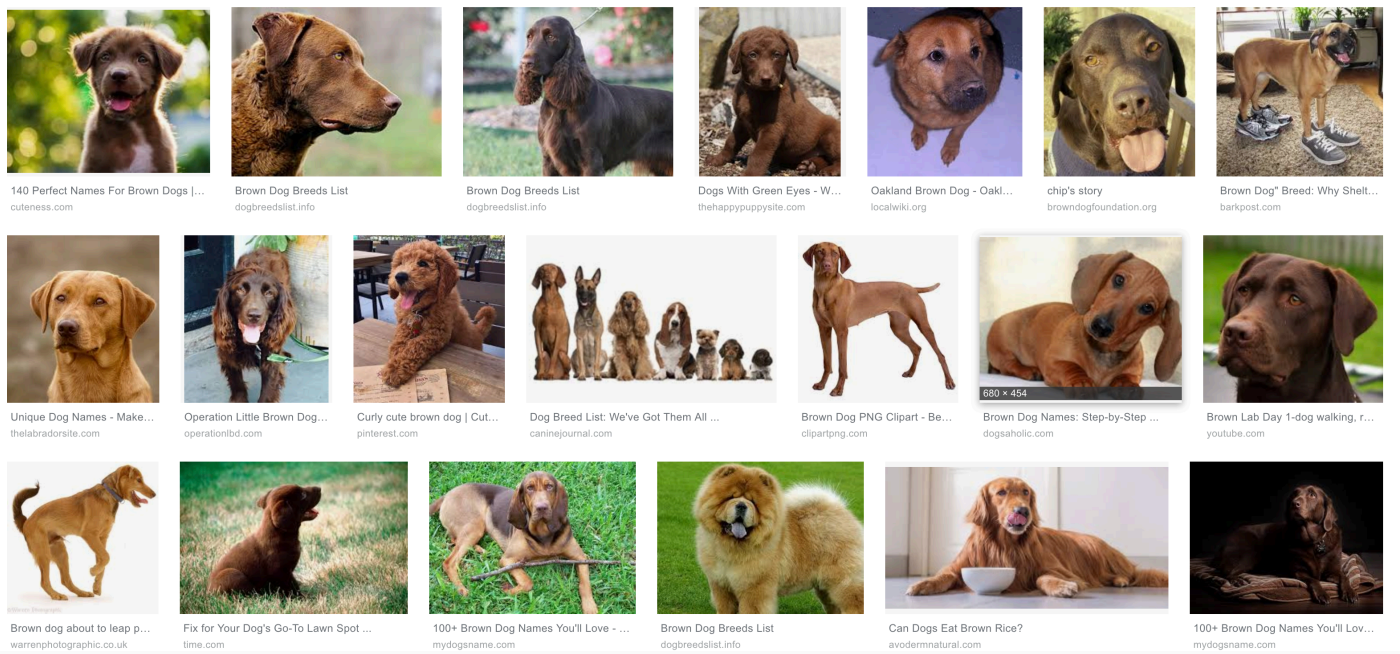


Overfitting in classification

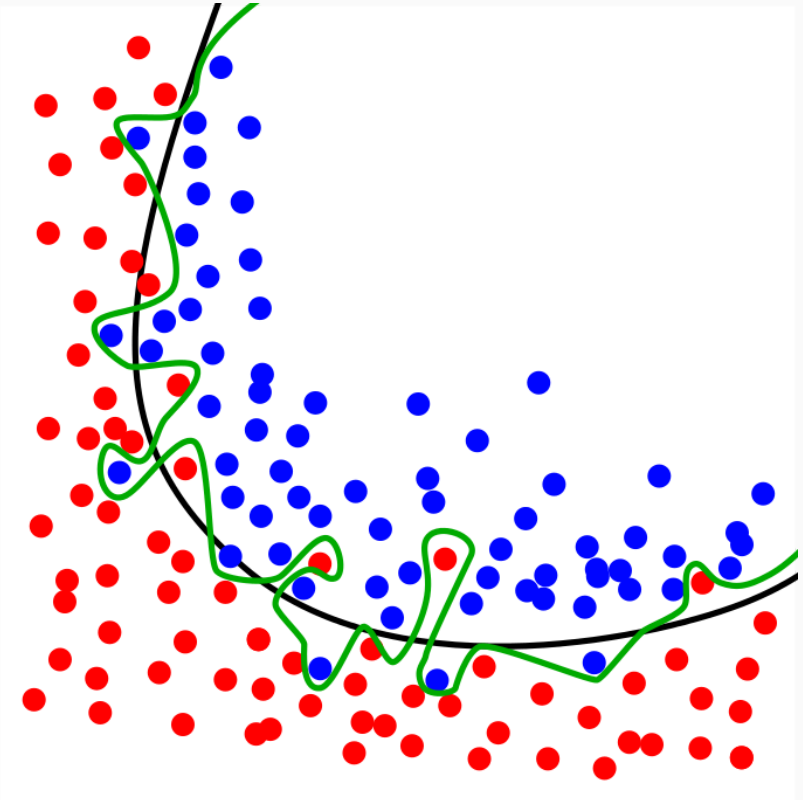


Overfitting in classification

Hertie School



Overfitting in classification

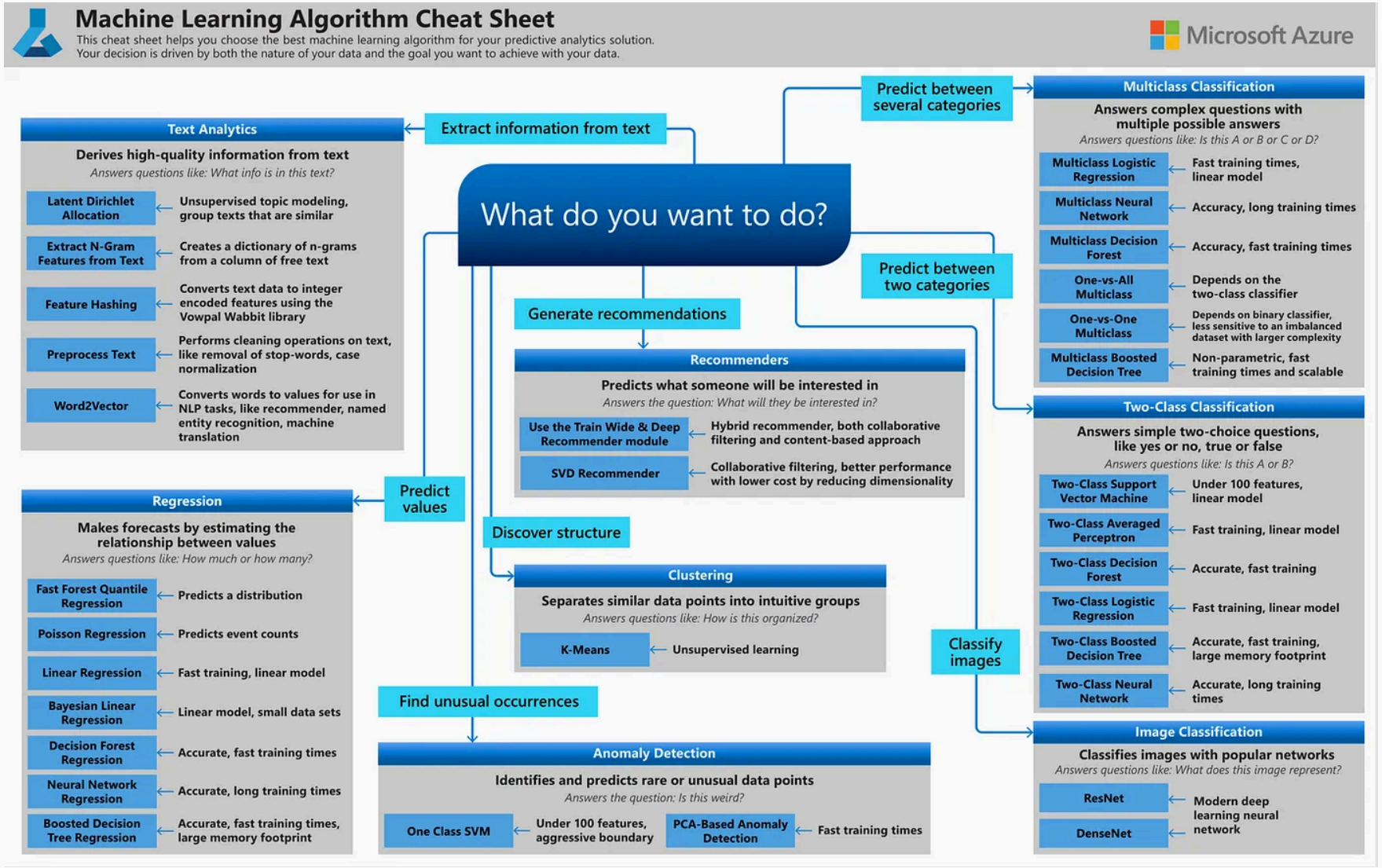


Explained: The green line represents an overfitted model and the black line represents a regularized model. While the green line best follows the training data, it is too dependent on that data and it is likely to have a higher error rate on new unseen data, compared to the black line.

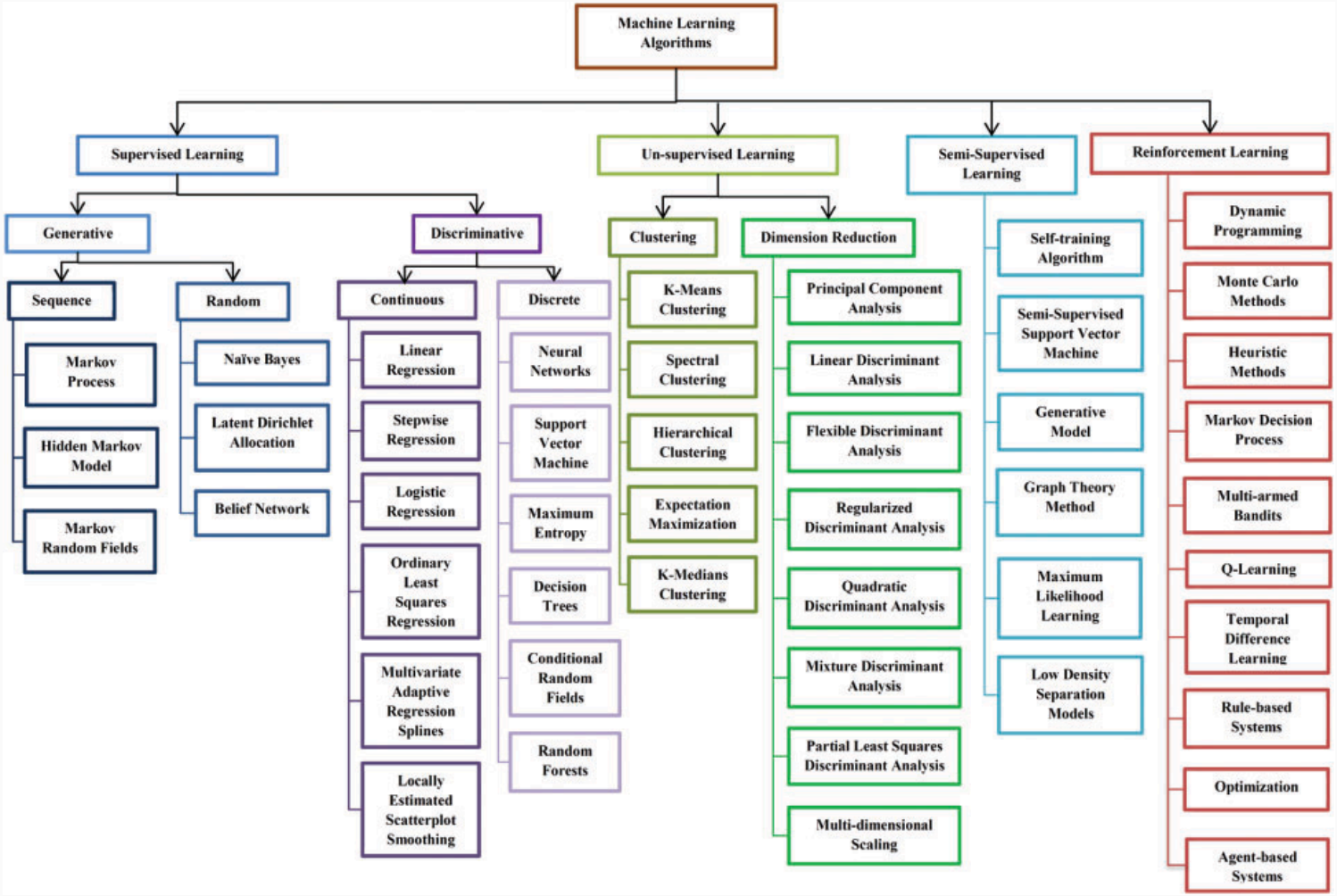
Overfitting, ultimately explained

Overview of ML landscape

The ML landscape (Microsoft.com)



ML decision tree



Source Sundararajan et al. 2021

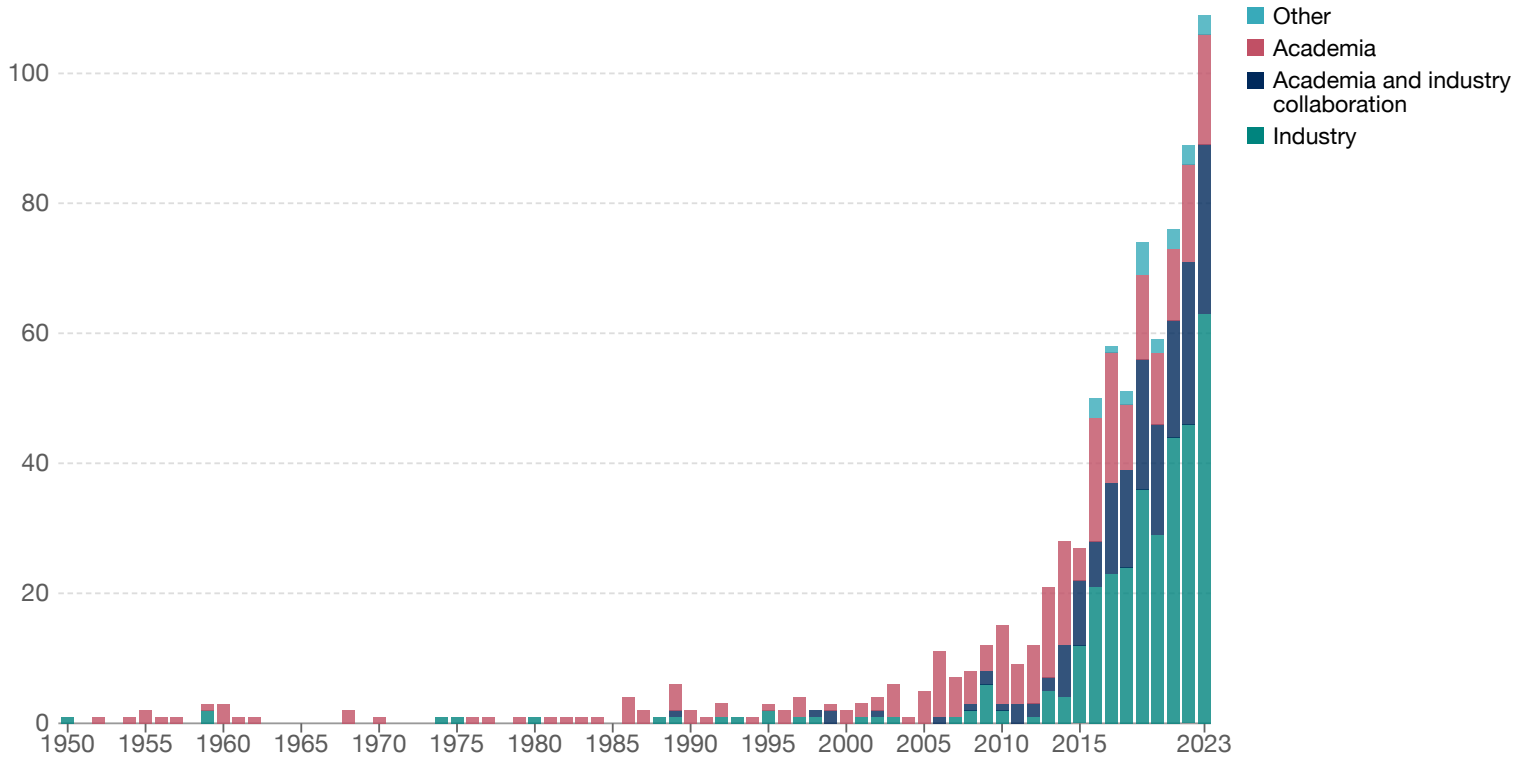
Affiliation of AI researchers

Hertie School

Affiliation of research teams building notable AI systems, by year of publication

Our World in Data

Sector where the authors of an AI system have their primary affiliations.



Data source: Epoch (2024)

OurWorldInData.org/artificial-intelligence | CC BY

Note: A research collective is a group of AI researchers not organized under an academic or industry affiliation. Systems are defined as "notable" by the authors based on several criteria, such as advancing the state of the art or being of historical importance.

Performance metrics

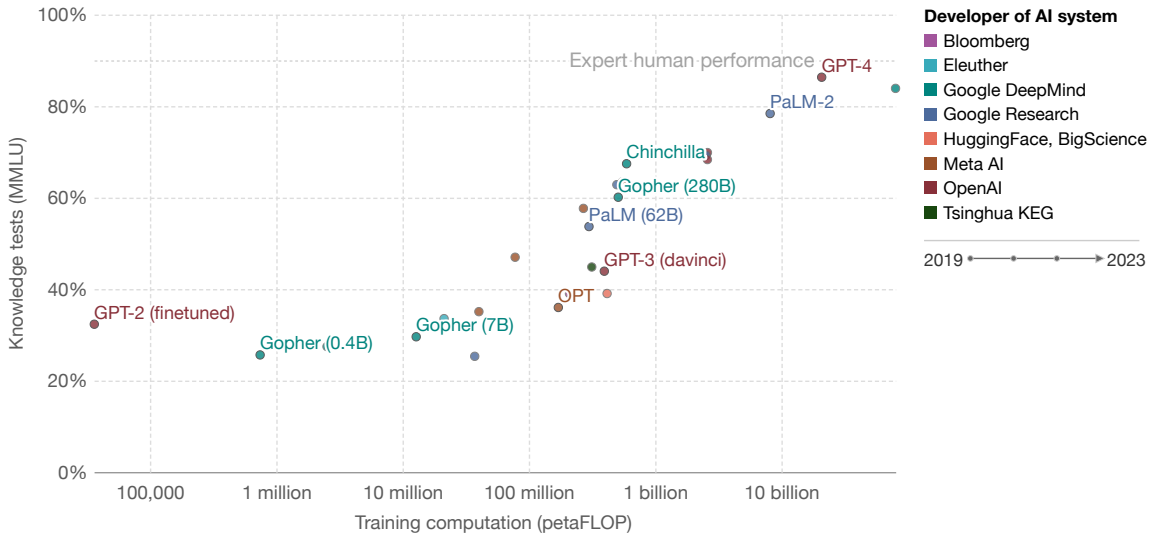
AI performance in knowledge tests

Hertie School

Artificial intelligence: Performance on knowledge tests vs. training computation

Our World in Data

Performance on knowledge tests is measured with the MMLU benchmark[1], here with 5-shot learning, which gauges a model's accuracy after receiving only five examples for each task. Training computation is measured in total petaFLOP, which is 10^{15} floating-point operations[2].



Data source: Epoch (2023)

OurWorldInData.org/artificial-intelligence | CC BY

Note: The values for training computation are estimates and come with some uncertainty, especially for models for which only minimal information has been disclosed, such as GPT-4.

1. MMLU benchmark: The Massive Multitask Language Understanding (MMLU) benchmark mimics a multiple-choice knowledge quiz designed to gauge how proficiently AI systems can comprehend various topics like history, science, or psychology. It has 57 different sections, each one looking at a particular subject. The MMLU test has 15,908 questions in total, which are split up into smaller sets. There are at least 100 questions about each subject. The questions in the test come from many places, like practice tests for big exams or questions from university courses. The difficulty of the questions varies, some are as easy as elementary school level, while others are as hard as what professionals in a field might know. The scores achieved by humans on this test are largely dependent on their level of expertise in the subject matter. Individuals who are not specialists in a given area typically achieve a correctness rate of around 34.5%. However, those with a deep understanding and proficiency in their field, such as doctors sitting for a medical examination, can attain a high score of up to 89.8% on the test.

2. Floating-point operation: A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

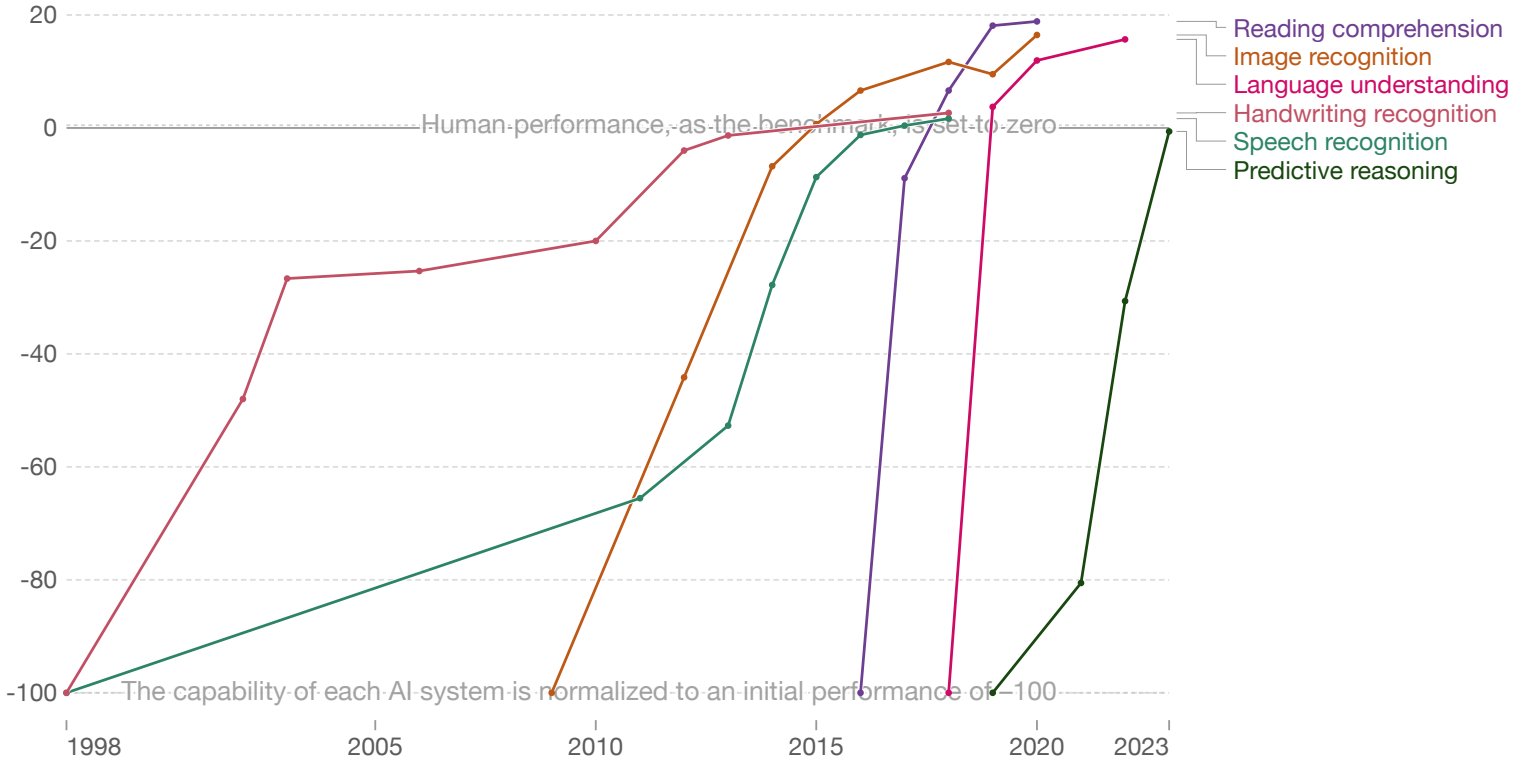
AI capabilities vs. human performance

Hertie School

Test scores of AI systems on various capabilities relative to human performance

Our World in Data

Within each domain, the initial performance of the AI is set to –100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.



Data source: Kiela et al. (2023)

OurWorldInData.org/artificial-intelligence | CC BY

Note: For each capability, the first year always shows a baseline of –100, even if better performance was recorded later that year.

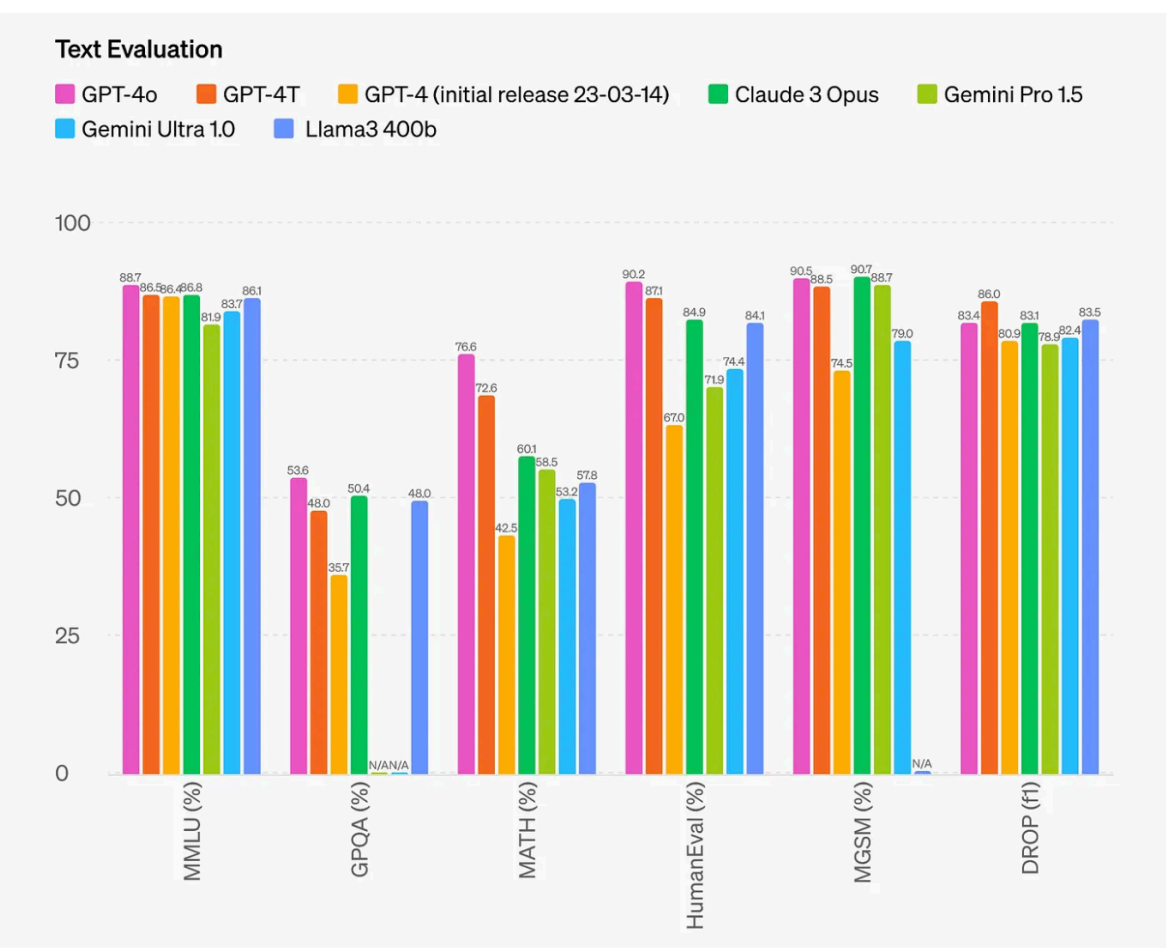
ML performance benchmarking in the wild

Dataset	Moderation Service	ROC AUC	F1	FPR	FNR
ToxiGen	Amazon	70.4%	68.9%	7.2%	52.0%
	Google	62.7%	62.7%	39.1%	35.5%
	OpenAI	70.3%	68.1%	33.2%	56.0%
	Microsoft	59.8%	57.4%	16.4%	64.0%
Jigsaw	Amazon	92.2%	92.2%	7.5%	8.1%
	Google	69.9%	67.2%	58.4%	1.8%
	OpenAI	78.6%	78.6%	17.1%	25.6%
	Microsoft	75.8%	75.7%	20.4%	28.1%
MegaSpeech	Amazon	72.8 %	72.0 %	10.4 %	43.9 %
	Google	73.3 %	72.3 %	41.3 %	12.0 %
	OpenAI	77.1 %	76.7 %	8.4 %	37.3 %
	Microsoft	70.6 %	70.1 %	16.9 %	41.9 %

TABLE 2: Performance metrics by moderation service and dataset. The names of moderation services were abbreviated for better readability. ROC AUC is threshold-invariant, while F1, False Positive Rate and False Negative Rate are threshold-variant. Good performance maximises ROC AUC and F1, and minimises False Positive and False Negative Rates. Per dataset, blue shading signals the best performance, while red shading indicates the worst performance. ToxiGen includes 7,800 observations, while Jigsaw and MegaSpeech each contain 50,000. All datasets are balanced on toxic and non-toxic phrases.

`Source` Oueslati, 2024, Watching the Watchers: A Comparative Audit of Cloud-Based Commercial Content Moderation

ML performance benchmarking in the wild



OpenAI's blog post includes evaluation scores of known datasets, such as MMLU.

`Source` Wiik, 2024, GPT-4o vs. GPT-4 vs. Gemini 1.5 — Performance Analysis (Accuracy).

Assessing classification performance

Accuracy

- Accuracy = $\frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{TP+TN}{TP+TN+FP+FN}$
- Error rate: $1 - \text{Accuracy}$

Usefulness

- Accuracy is a simple and intuitive metric.
- But it can be misleading, especially in imbalanced datasets where the classes are not evenly represented.
- Example: In a dataset with 90% of class A and 10% of class B, a model that predicts all instances as class A will have an accuracy of 90%, but it will not be useful for predicting class B instances.

Example

Outcome recidivism:
1 = recidivate, 0 = not recidivate

	Predicted values		
	0	1	Total
True/actual values			
0	540	224	764
1	255	424	679
Total	795	648	1443

What is the **accuracy** of our recidivism classifier?

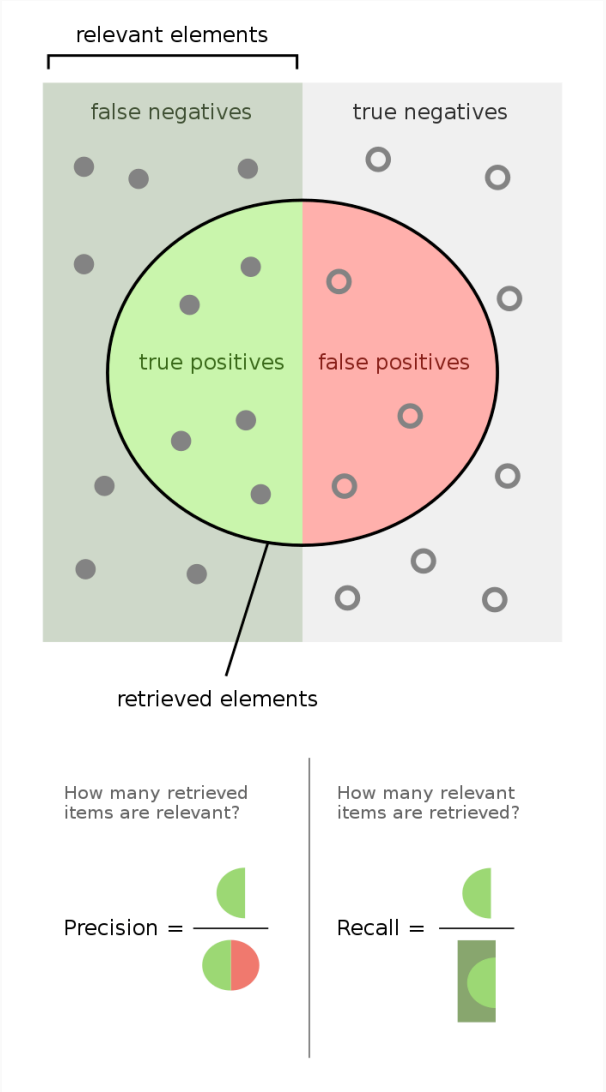
Assessing classification performance

Precision

- Precision = $\frac{\text{Number of true positive predictions}}{\text{Number of positive predictions}} = \frac{TP}{TP+FP}$

Usefulness

- Precision focuses on the accuracy of positive predictions and is useful when the cost of false positives is high.



Assessing classification performance

Precision

- Precision = $\frac{\text{Number of true positive predictions}}{\text{Number of positive predictions}} = \frac{TP}{TP+FP}$

Usefulness

- Precision focuses on the accuracy of positive predictions and is useful when the cost of false positives is high.

Example

Outcome recidivism: 1 = recidivate, 0 = not recidivate			
	Predicted values		
	0	1	Total
True/actual values			
0	540	224	764
1	255	424	679
Total	795	648	1443

What is the **precision** of our recidivism classifier?

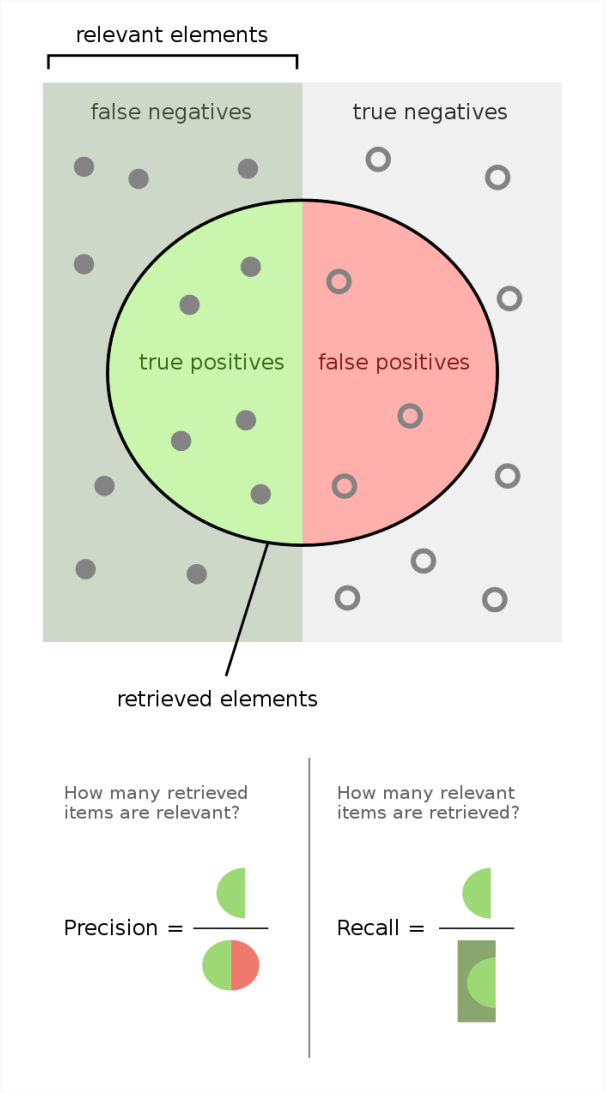
Assessing classification performance

Recall (Sensitivity)

- Recall = $\frac{\text{Number of true positive predictions}}{\text{Number of true positives}} = \frac{TP}{TP+FN}$
- "True positive rate"

Usefulness

- Recall focuses on capturing all positive instances and is important when the cost of false negatives is high.
- Example: In a medical diagnosis, recall is important to ensure that all patients with a disease are correctly identified.
- The complementary measure is specificity (true negative rate; e.g. how many healthy people are identified as not having the condition)



Assessing classification performance

Recall (Sensitivity)

- Recall = $\frac{\text{Number of true positive predictions}}{\text{Number of true positives}} = \frac{TP}{TP+FN}$
- "True positive rate"

Usefulness

- Recall focuses on capturing all positive instances and is important when the cost of false negatives is high.
- Example: In a medical diagnosis, recall is important to ensure that all patients with a disease are correctly identified.
- The complementary measure is specificity (true negative rate; e.g. how many healthy people are identified as not having the condition)

Example

Outcome recidivism: 1 = recidivate, 0 = not recidivate			
	Predicted values		
	0	1	Total
True/actual values			
0	540	224	764
1	255	424	679
Total	795	648	1443

What is the **recall** of our recidivism classifier?

Assessing classification performance

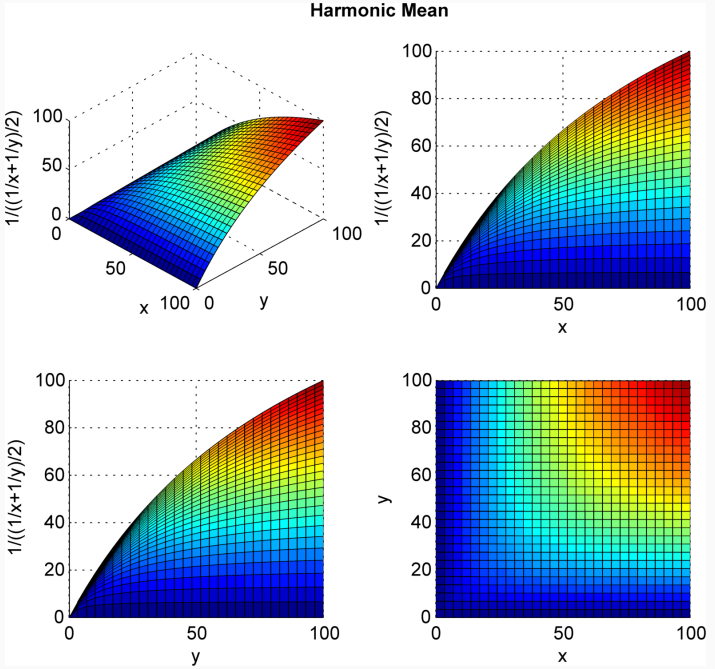
F1 score

- F1 score = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$
- F1 score is the harmonic mean of precision and recall.

Usefulness

- It provides a balance between precision and recall, especially when there is an imbalance between the classes.
- F1 score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates poor performance.

Illustration



Normalised harmonic mean plot where x is precision, y is recall and the vertical axis is F1 score, in % points

`Source` Andong87

Assessing classification performance

F1 score

- F1 score = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$
- F1 score is the harmonic mean of precision and recall.

Usefulness

- It provides a balance between precision and recall, especially when there is an imbalance between the classes.
- F1 score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates poor performance.

Example

Outcome recidivism: 1 = recidivate, 0 = not recidivate			
	Predicted values		
	0	1	Total
True/actual values			
0	540	224	764
1	255	424	679
Total	795	648	1443

What is the **F1 score** of our recidivism classifier?

Why performance metrics can matter

Scenario

- **Outcome:** Recidivism where individual recidivates (1) or not (0)
- **False Positive (FP):** Model predicts an individual will recidivate when they actually do not.
- **False Negative (FN):** Model predicts an individual will not recidivate when they actually do.
  - This could result in individuals who are at risk, being released without proper intervention, potentially leading to repeat offenses.

Assigning costs

- What are downstream costs of FP and FN?
- At which level do the costs apply - individual, societal, ...?

Ethical and economic reasoning

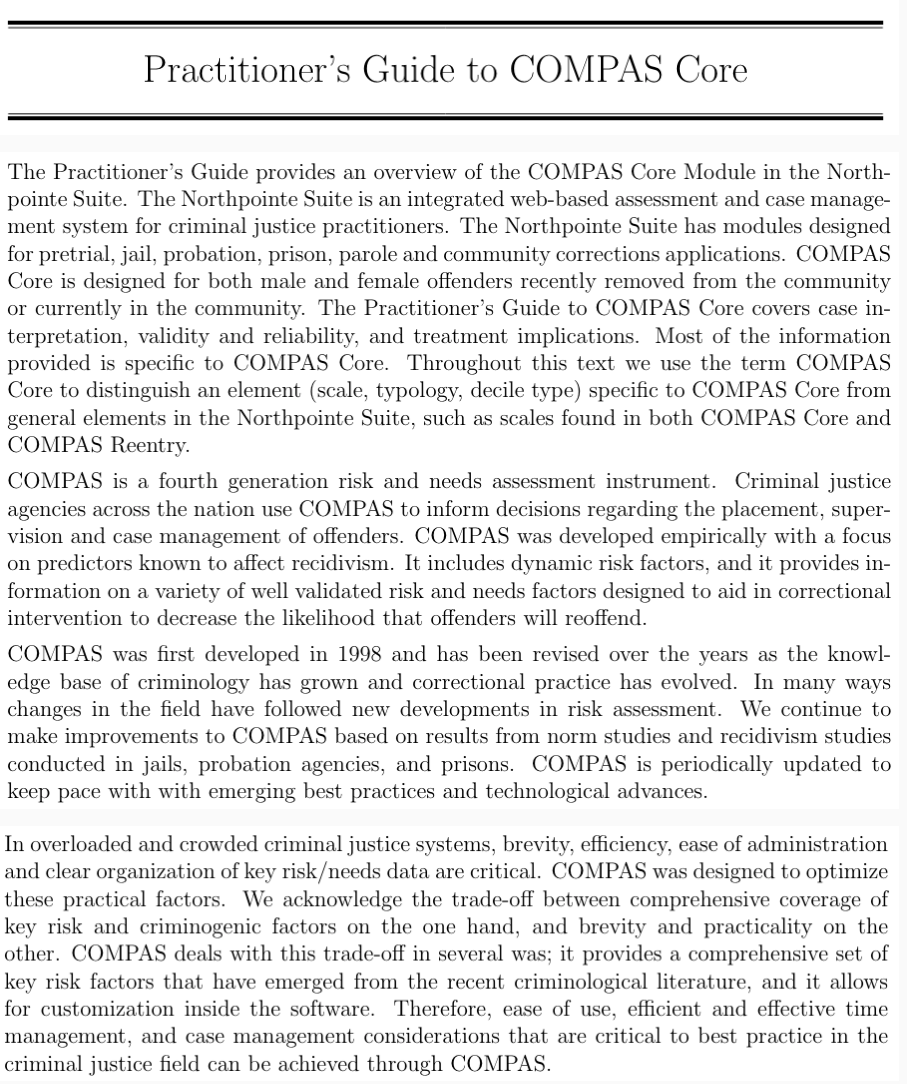
- How should we weigh the costs of FP and FN?
- What should we prioritize in our model - reducing FP, FN, or balancing both?

AI for public policy

The COMPAS algorithm to predict criminals' recidivism

Background

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a decision support tool developed by Northpointe (now Equivant) used by U.S. courts to **assess the likelihood of recidivism**
- Produced several scales (Pretrial release risk, General recidivism, Violent recidivism) based on factors such as age, criminal history, and substance abuse
- The algorithm is proprietary and its inner workings are not public



Practitioner's Guide to COMPAS Core

The Practitioner's Guide provides an overview of the COMPAS Core Module in the Northpointe Suite. The Northpointe Suite is an integrated web-based assessment and case management system for criminal justice practitioners. The Northpointe Suite has modules designed for pretrial, jail, probation, prison, parole and community corrections applications. COMPAS Core is designed for both male and female offenders recently removed from the community or currently in the community. The Practitioner's Guide to COMPAS Core covers case interpretation, validity and reliability, and treatment implications. Most of the information provided is specific to COMPAS Core. Throughout this text we use the term COMPAS Core to distinguish an element (scale, typology, decile type) specific to COMPAS Core from general elements in the Northpointe Suite, such as scales found in both COMPAS Core and COMPAS Reentry.

COMPAS is a fourth generation risk and needs assessment instrument. Criminal justice agencies across the nation use COMPAS to inform decisions regarding the placement, supervision and case management of offenders. COMPAS was developed empirically with a focus on predictors known to affect recidivism. It includes dynamic risk factors, and it provides information on a variety of well validated risk and needs factors designed to aid in correctional intervention to decrease the likelihood that offenders will reoffend.

COMPAS was first developed in 1998 and has been revised over the years as the knowledge base of criminology has grown and correctional practice has evolved. In many ways changes in the field have followed new developments in risk assessment. We continue to make improvements to COMPAS based on results from norm studies and recidivism studies conducted in jails, probation agencies, and prisons. COMPAS is periodically updated to keep pace with with emerging best practices and technological advances.

In overloaded and crowded criminal justice systems, brevity, efficiency, ease of administration and clear organization of key risk/needs data are critical. COMPAS was designed to optimize these practical factors. We acknowledge the trade-off between comprehensive coverage of key risk and criminogenic factors on the one hand, and brevity and practicality on the other. COMPAS deals with this trade-off in several was; it provides a comprehensive set of key risk factors that have emerged from the recent criminological literature, and it allows for customization inside the software. Therefore, ease of use, efficient and effective time management, and case management considerations that are critical to best practice in the criminal justice field can be achieved through COMPAS.

`Source` Practitioner's Guide to COMPAS Core

The COMPAS algorithm to predict criminals' recidivism

The ProPublica and other investigations

- In 2016, ProPublica published an investigation showing that COMPAS was **biased against African Americans**
- **Bias:** The algorithm was more likely for African Americans to wrongly predict that defendants would re-offend.
- **Accuracy:** only 20% of people predicted to commit violent crimes actually went on to do so (in a later study estimated with 65%, still worse than a group of humans with little expertise)

Source ProPublica 2016

Machine Bias*

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner

There's software used across the country to predict future criminals. And it's biased against blacks.

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances—which belonged to a 6-year-old boy—a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late—a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for shoplifting $86.35 worth of tools from a nearby Home Depot store (Figure 6.1.1).

Prater was the more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery charge. Borden had a record, too, but it was for misdemeanors committed when she was a juvenile.

Yet something odd happened when Borden and Prater were booked into jail: A computer program spat out a score predicting the likelihood of each committing a future crime. Borden—who is black—was rated a high risk. Prater—who is white—was rated a low risk.

Two years later, we know the computer algorithm got it exactly backward. Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars' worth of electronics.

Scores like this—known as risk assessments—are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts—as is the case in Fort Lauderdale—to

* Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, "Machine Bias," *ProPublica* (May 23, 2016). Reprinted with permission.

DOI: 10.1201/9781003278290-37

The COMPAS algorithm to predict criminals' recidivism

The ProPublica and other investigations

- In 2016, ProPublica published an investigation showing that COMPAS was **biased against African Americans**
- **Bias:** The algorithm was more likely for African Americans to wrongly predict that defendants would re-offend.
- **Accuracy:** only 20% of people predicted to commit violent crimes actually went on to do so (in a later study estimated with 65%, still worse than a group of humans with little expertise)

SCIENCE ADVANCES | RESEARCH ARTICLE

RESEARCH METHODS

The accuracy, fairness, and limits of predicting recidivism

Julia Dressel and Hany Farid*

Algorithms for predicting recidivism are commonly used to assess a criminal defendant's likelihood of committing a crime. These predictions are used in pretrial, parole, and sentencing decisions. Proponents of these systems argue that big data and advanced machine learning make these analyses more accurate and less biased than humans. We show, however, that the widely used commercial risk assessment software COMPAS is no more accurate or fair than predictions made by people with little or no criminal justice expertise. In addition, despite COMPAS's collection of 137 features, the same accuracy can be achieved with a simple linear predictor with only two features.

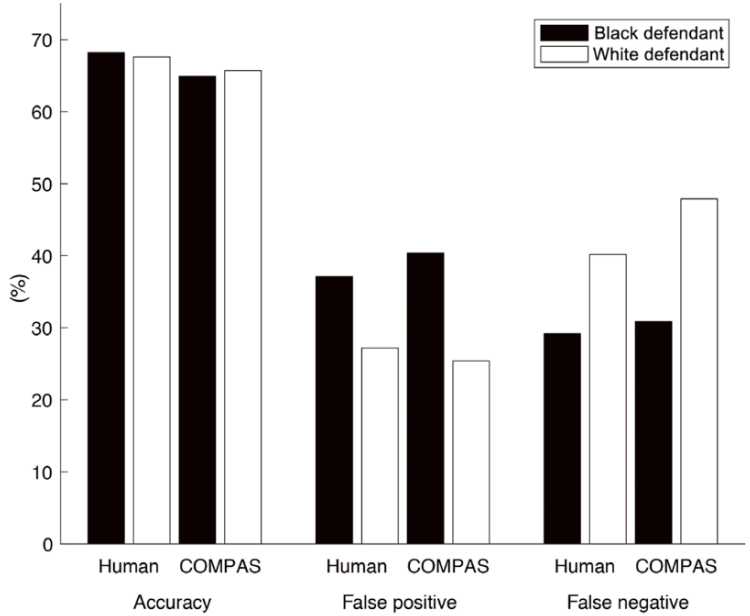


Fig. 1. Human (no-race condition) versus COMPAS algorithmic predictions (see also Table 1).

`Source` Dressel and Fair, 2018, Science Advances

Discussion

1. Where do you see **potential** for AI in public policy-making?
2. Are there **applications of AI in Georgian government** that you are aware of?
3. What role does **AI** play **in your personal (professional) life**?