Minería de Datos

Welcome to the Machine Learning class

Valentin Barriere

Universidad de Chile – DCC

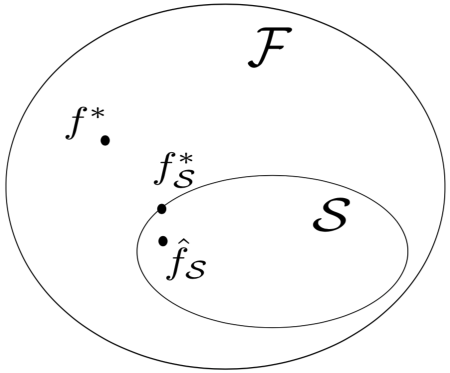
CC5205, Otoño 2024

Sesgos y Causalidades en la Modelizacion

Outline : Generalization and Approximation errors

Complejidad y modelos: Objetivo Riesgo Minimum

- $\mathcal{F} = \{f$:funciones medibles $\mathcal{X} \to \mathcal{Y}\}$
- Mejor solución $f^* = \arg\min_{f \in \mathcal{F}} \mathcal{R}(f)$
- Clase de funciones $\mathcal{S} \subset \mathcal{F}$ utilizadas como modelos
- Objetivo ideal en $\mathcal{S}$: $f_{\mathcal{S}}^* = \arg\min_{f \in \mathcal{S}} \mathcal{R}(f)$
- Estimación obtenida en $\mathcal{S}$: se obtiene $f_{\mathcal{S}}$ tras un entrenamiento



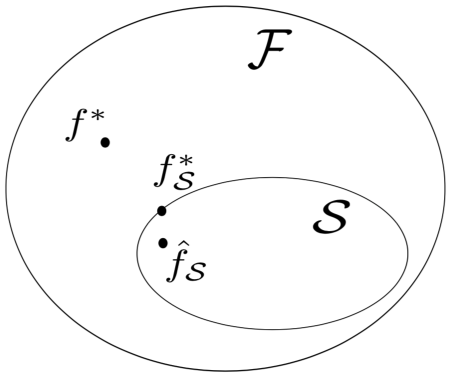
Error de aproximación y error de generalizacion

$$\mathcal{R}(\hat{f}_{\mathcal{S}}) - \mathcal{R}(f^*) = \underbrace{\mathcal{R}(f_{\mathcal{S}}^*) - \mathcal{R}(f^*)}_{\text{error de aproximacion}} + \underbrace{\mathcal{R}(\hat{f}_{\mathcal{S}}) - \mathcal{R}(f_{\mathcal{S}}^*)}_{\text{error de estimacion}}$$

- El error de aproximación puede ser grande si el modelo $\mathcal{S}$ no es adaptado
- El error de estimación puede ser grande si el modelo es complejo
- $\mathcal{R}(f^*)$ viene de la limitacion de los datos

Complejidad y modelos: Objetivo Riesgo Minimum

- $\mathcal{F} = \{f$:funciones medibles $\mathcal{X} \to \mathcal{Y}\}$
- Mejor solución $f^* = \arg\min_{f \in \mathcal{F}} \mathcal{R}(f)$
- Clase de funciones $\mathcal{S} \subset \mathcal{F}$ utilizadas como modelos
- Objetivo ideal en $\mathcal{S}$: $f_{\mathcal{S}}^* = \arg\min_{f \in \mathcal{S}} \mathcal{R}(f)$
- Estimación obtenida en $\mathcal{S}$: se obtiene $f_{\mathcal{S}}$ tras un entrenamiento

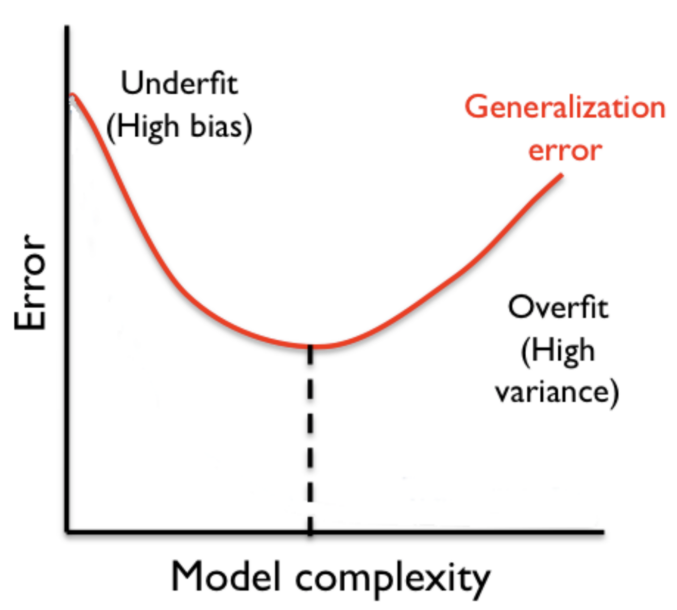


Error de aproximación y error de generalizacion

$$\mathcal{R}(\hat{f}_{\mathcal{S}}) - \mathcal{R}(f^*) = \underbrace{\mathcal{R}(f_{\mathcal{S}}^*) - \mathcal{R}(f^*)}_{\text{error de aproximacion}} + \underbrace{\mathcal{R}(\hat{f}_{\mathcal{S}}) - \mathcal{R}(f_{\mathcal{S}}^*)}_{\text{error de estimacion}}$$

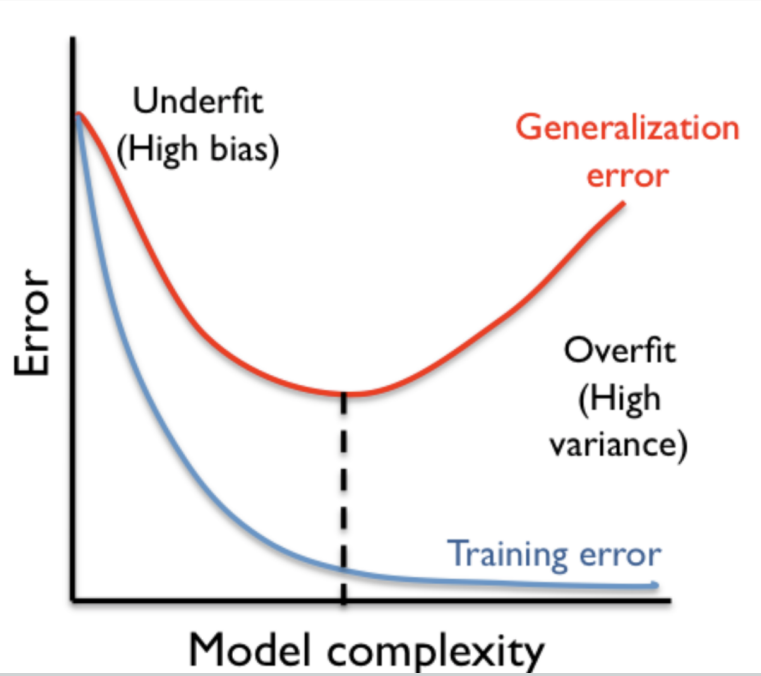
- El error de aproximación puede ser grande si el modelo $\mathcal{S}$ no es adaptado
- El error de estimación puede ser grande si el modelo es complejo
- $\mathcal{R}(f^*)$ viene de la limitacion de los datos

Sobre-aprendizaje y sub-aprendizaje



- Según la complejidad del modelo (por ejemplo, tiempo de entrenamiento, cantiadad de parametros) se observa un comportamiento diferente
- Los modelos poco complejos son aprendidos fácilmente pero el error de aproximación puede ser grande (sub-aprendizaje)
- Los modelos muy complejos pueden tener el objetivo correcto pero un gran error de estimación (sobre-aprendizaje)

Sobre-aprendizaje: Problema

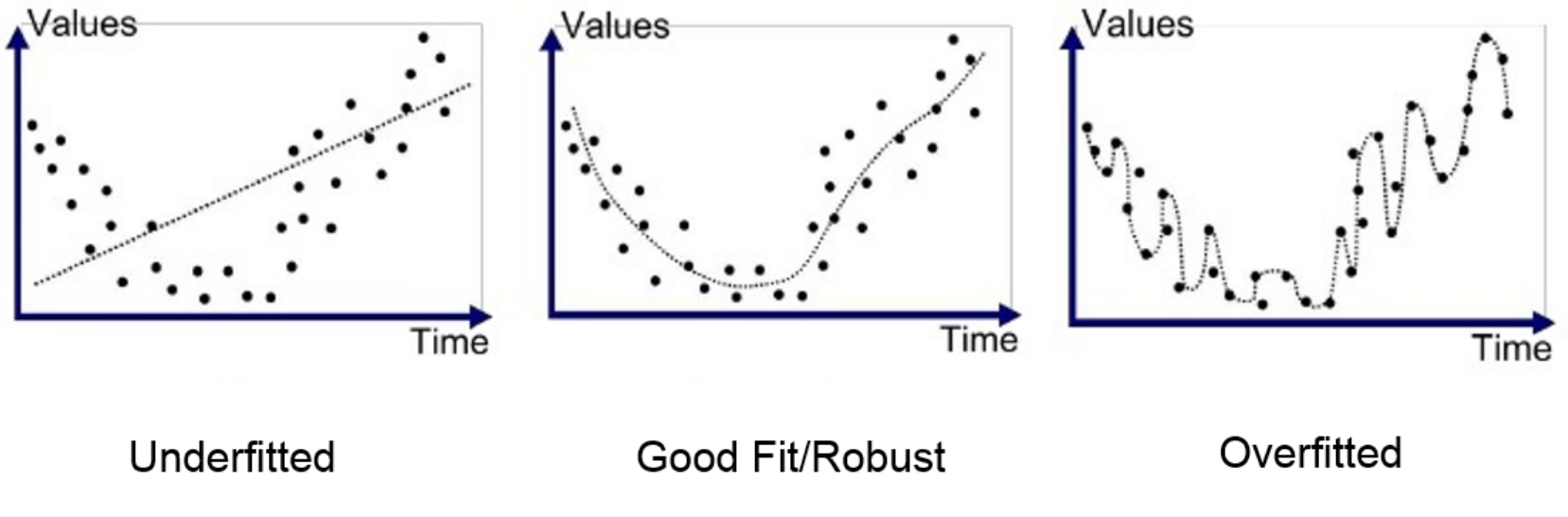


Error y riesgos

- El riesgo empírico (error en el conjunto de entrenamiento) disminuye con el aumento de la complejidad del modelo
- El riesgo real (error en observaciones de un nuevo conjunto) es muy diferente. ¡Tenemos un problema de generalización!

- Sobre-aprendizaje : los parámetros aprendidos son demasiado específicos para el conjunto de entrenamiento
- Se debe usar un criterio diferente al error en el conjunto de entrenamiento

Sobre-aprendizaje: Complejidad



Complejidad

- Si el modelo es demasiado simple, entonces ya no sigue los datos
- Si el modelo es demasiado complejo, el modelo aprende todas las irregularidades del conjunto de datos $\mathcal{D}_n$: el objetivo no es memorizar los datos de entrenamiento, sino generalizar los nuevos datos
- Ejemplo : si el modelo es el de la curva del medio más una componente de ruido no considerada en las variables, el modelo de la derecha aprende ese ruido

Simple case

Let's say we have

- A N-**order polynomial** $X_N = \sum_i^N \alpha_i X^i$ using a
- I want to **predict the variable** Y composed of X_N and some noise ε that we could not measure: $Y = X_N + \varepsilon$
- My **model is a** k-**order polynomial** $\hat{Y} = X_k = \sum_i^k w_i X^i$

With $k \leq N$, noting $Y_k = \sum_1^k \alpha_i X^i$ and $\varepsilon_{N-k} = \sum_{k+1}^N \alpha_i X^i$ we obtain:

$$Y = Y_k + \varepsilon_{N-k} + \varepsilon$$

It is possible to model Y using $\hat{Y}$ up to a certain points because:

- ε_{N-k} cannot be modeled because of our model limitation
- ε cannot be modeled because of our data limitation

Simple case

Let's say we have

- A N-**order polynomial** $X_N = \sum_i^N \alpha_i X^i$ using a
- I want to **predict the variable** Y composed of X_N and some noise ε that we could not measure: $Y = X_N + \varepsilon$
- My **model is a** k-**order polynomial** $\hat{Y} = X_k = \sum_i^k w_i X^i$

With $k \geq N$, noting $\varepsilon_{N_+} = \sum_{N+1}^k w_i X^i$ we obtain:

$$\hat{Y} \sim X_N + \varepsilon_{N_+}$$

It is possible to model Y using $\hat{Y}$ but we will reach other problems:

- ε_{N_+} will focus on trying to learn ε, which is impossible to predict
- ε still cannot be modeled because of our data limitation, creating the inevitable error $\mathcal{R}(f^*)$

Simple case

If we have a low $k << N$

Our model will be lowly complex, and in this case ε_{N-k} can be high. The model is underfitting the data and we will get an **approximation error**.

If we have a $k = N$

Our model will have the perfect complexity to measure the data, and in this case ε_{N-k} can be as close as zero. Still, the model will not be able to model ε as its **information is not contained in the observations**.

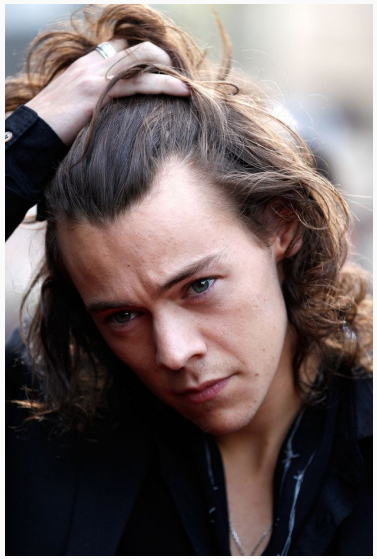
If we have a high $k >> N$

Our model will be overly complex for the task, it will also learn to memorize the ε that cannot be predicted with the observations. It lost its generalization power. The model is overfitting the data and we will get a **generalization error**.

What is this ε ? Missing information from the observations

Let's the simple example of a **model predicting the gender of a person, using hair length** (one boolean variable).

- The model will used the observations to reach the objective
- To minimize the error, it will rely on the **prior probability of hair length with respect to the gender**.
- Predictions will be true for majority groups



ε is not noise

ε will have a non-zero value for females with short hair and the males with long hair, inducing a **negative bias** in the model behavior regarding these two groups.

What if $\varepsilon = 0$? Missing information from the observations

In the perfect case where all the target information is stored in the observations, $\varepsilon = 0$, however other problems arise:

- The perfect model might be difficult to learn: it would need more data (for example difficult examples) or more parameters.
- But if the model has too many input variables and/or parameters, it will be impacted by the noise it brings (cf. curse of dimensionality)[1]
- Especially when the data is skewed, the model will rely on simple heuristics to predict its answer.

During the training process, the parameters will converge to what helps the model the most to predict, even though relying on correlations but not causations.

[1] The Curse of Dimensionality arises when working with high-dimensional data, leading to increased computational complexity, overfitting, and spurious correlations

Let's say $\varepsilon = 0$

We still can have problems!

Outline : Correlation vs Causations

Predictive Power and Spurious Correlations

Optimizing over a classical loss function will get us a model obtaining the highest performances possible, even though it relies on correlations and not causalities.

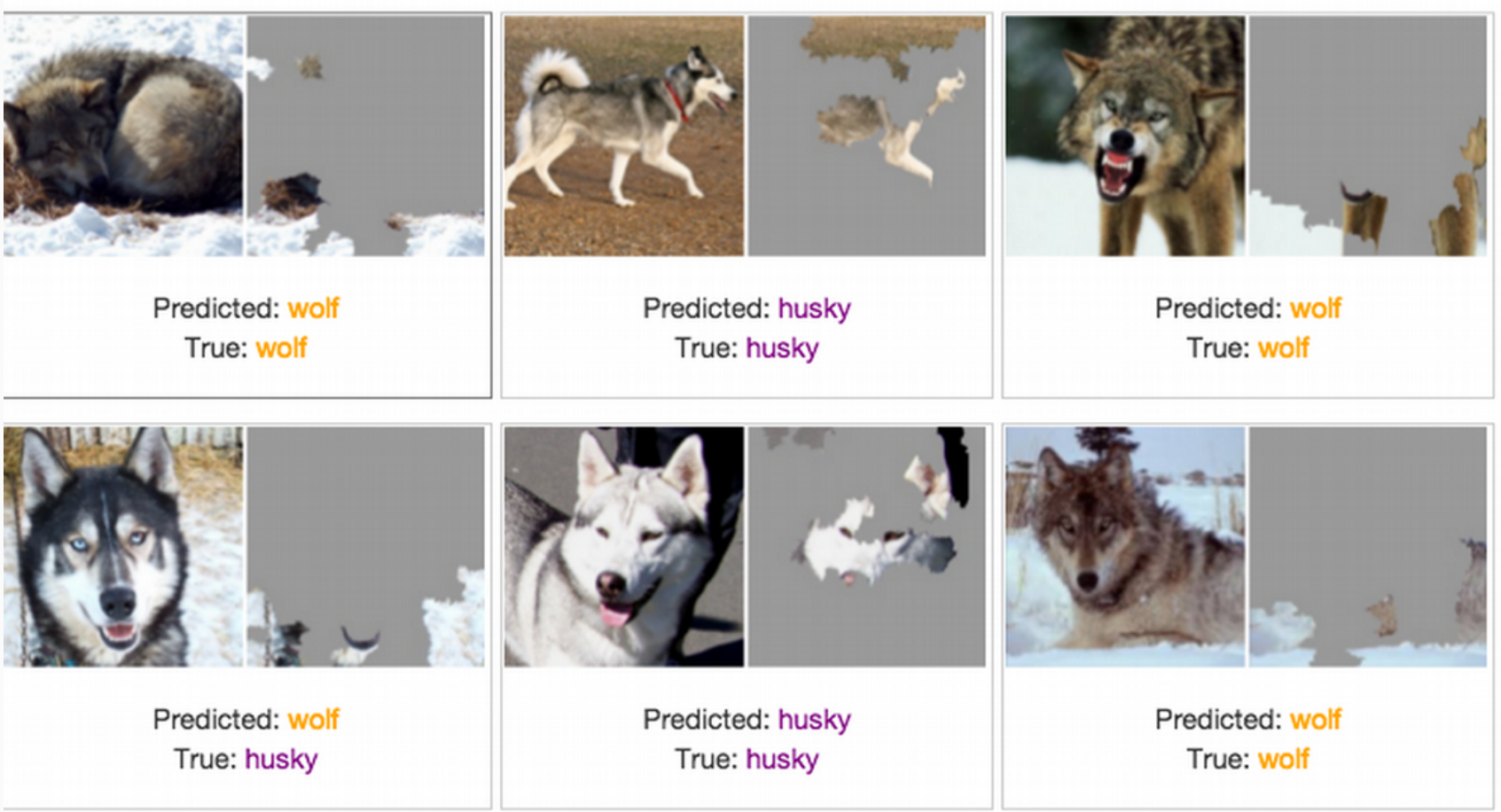


Figure 1: Explanation of the decisions of a Husky/Wolf classifier

Predictive Power and Spurious Correlations

(a) Husky classified as wolf (b) Explanation

Figure 2: Explanation of the decisions of a Husky/Wolf classifier

The algorithm makes mistakes that are detrimental for husky dogs that are living in snowy places! Poor them :'(

Spurious Correlations

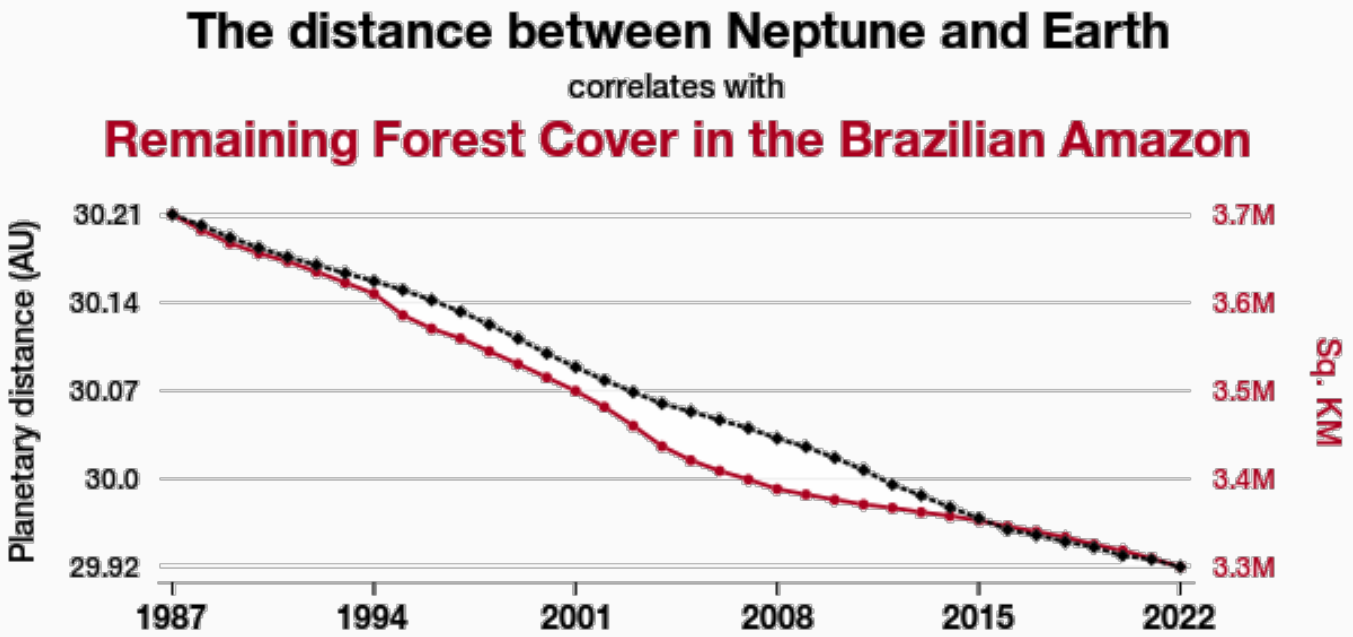


Figure 3: An example of spurious correlation [2]

Spurious Correlations

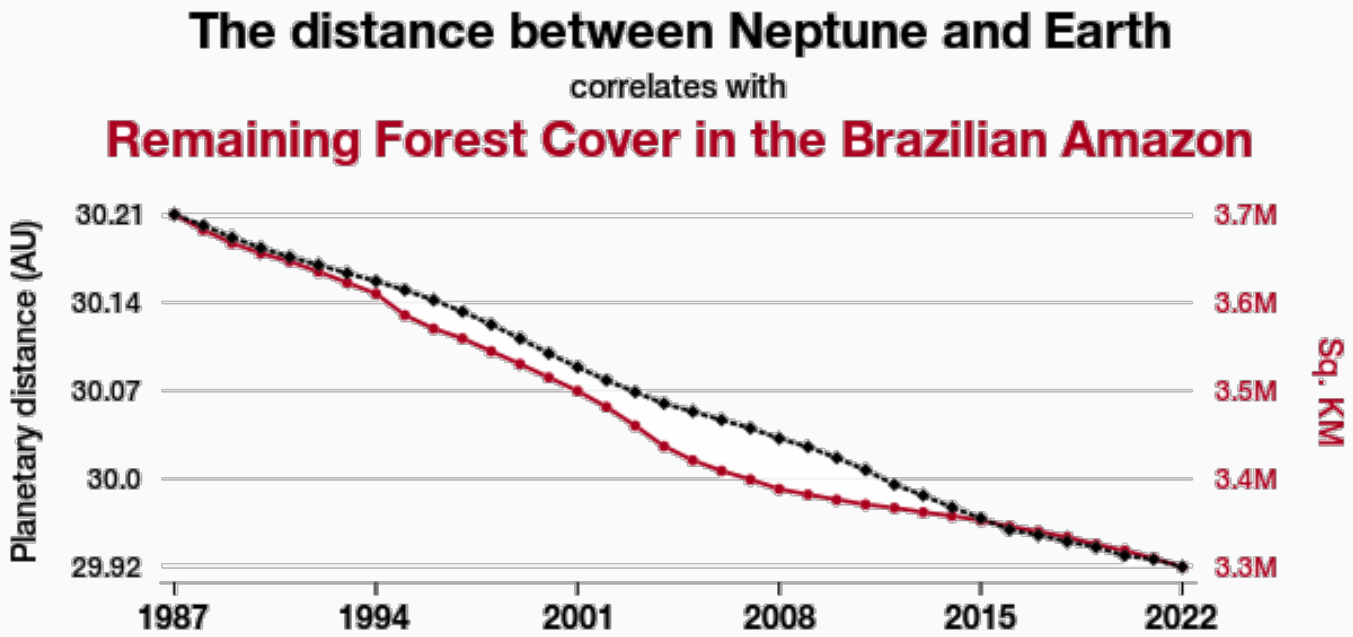


Figure 3: An example of spurious correlation [2] and IA illustrations

Show prompt used to generate this paper

AI academic paper

(Because p < 0.01)

A Cosmic Dance: The Neptunian Distance and Amazonian Resilience

The Journal of Interplanetary Resilience Studies

Jan 2024

Reminder: This paper is AI-generated. Not real!

Spurious Correlations can be Easy to Rely on

Spurious Correlations

Mathematical relationship in which two or more events or **variables are associated but not causally related**, due to either coincidence or the presence of a certain third, unseen factor (referred to as a "common response variable", "confounding factor", or "lurking variable").

- I certain contexts, **information can be shared** between several variables. For example, in the Husky/Wolf dataset, the variable snow shares a lot of information with the variable animal.
- Obviously, this is a **limitation of the data**, as this is not because the animal is next to snow that it is a wolf.
- Your model might want to use this information to better classify as **it might be easier**.

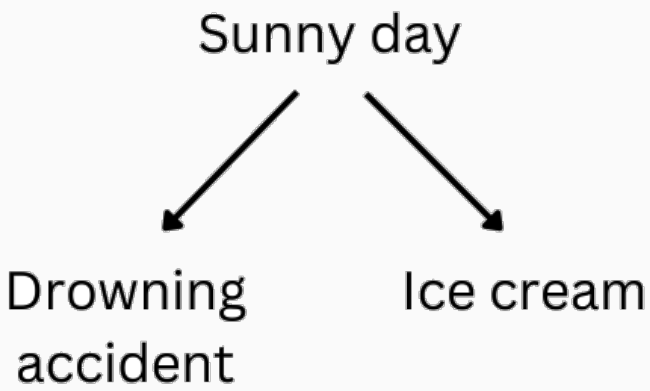
Spurious Correlations can be Easy to Rely on

Spurious Correlations

Mathematical relationship in which two or more events or **variables are associated but not causally related**, due to either coincidence or the presence of a certain third, unseen factor (referred to as a "common response variable", "confounding factor", or "lurking variable").

- I certain contexts, **information can be shared** between several variables. For example, in the Husky/Wolf dataset, the variable snow shares a lot of information with the variable animal.
- Obviously, this is a **limitation of the data**, as this is not because the animal is next to snow that it is a wolf.
- Your model might want to use this information to better classify as **it might be easier**.
- Like when your brain is using a stereotype because it is easier than a real thinking! **It is actually a cognitive bias of humans**.

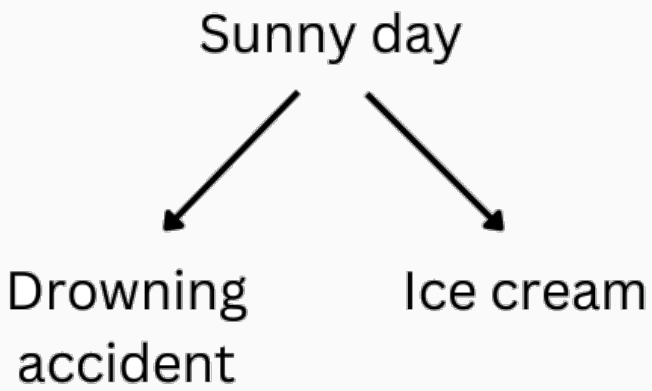
Spurious Correlations or Confounding variables



Case 1: Spurious correlation

- Sunny day impacts both drowning accident and ice cream sales
- With enough variables and samples, the model should work

Spurious Correlations or Confounding variables



Case 1: Spurious correlation

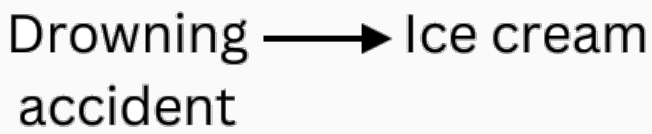
- Sunny day impacts both drowning accident and ice cream sales
- With enough variables and samples, the model should work
- What if we do not have the weather variable?

Spurious Correlations or Confounding variables

Case 1: Spurious correlation

- Sunny day impacts both drowning accident and ice cream sales
- With enough variables and samples, the model should work
- What if we do not have the weather variable?

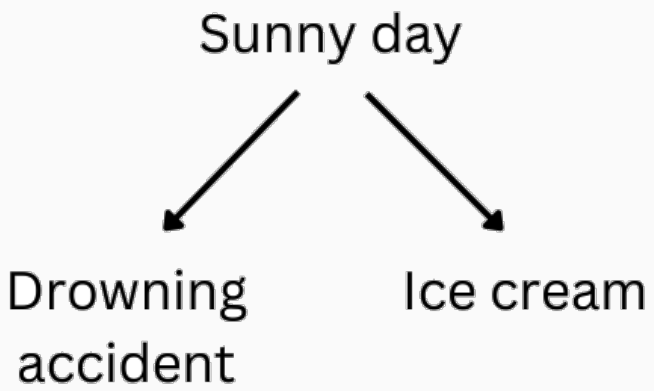
Spurious Correlations or Confounding variables



Case 1: Spurious correlation

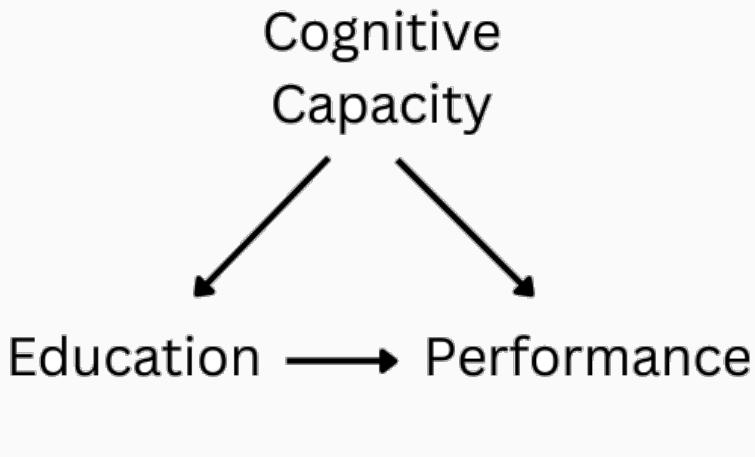
- Sunny day impacts both drowning accident and ice cream sales
- With enough variables and samples, the model should work
- What if we do not have the weather variable?
- Spurious correlation

Spurious Correlations or Confounding variables



Case 1: Spurious correlation

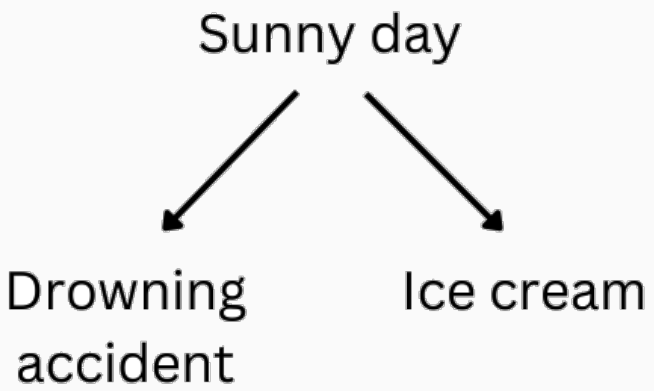
- Sunny day impacts both drowning accident and ice cream sales
- With enough variables and samples, the model should work
- What if we do not have the weather variable?
- Spurious correlation



Case 2: Confounder

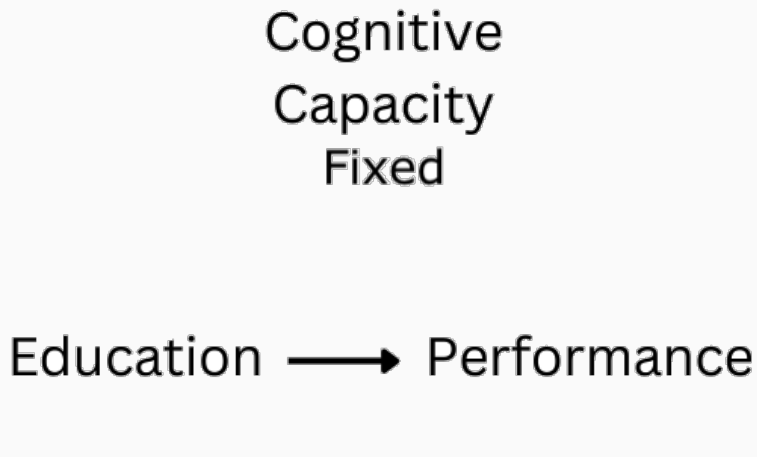
- Confounder are common causes
- Impossible to measure impact of Education over job Performance as a confounder influences both of them behind the scenes

Spurious Correlations or Confounding variables



Case 1: Spurious correlation

- Sunny day impacts both drowning accident and ice cream sales
- With enough variables and samples, the model should work
- What if we do not have the weather variable?
- Spurious correlation



Case 2: Confounder

- Confounder are common causes
- Impossible to measure impact of Education over job Performance as a confounder influences both of them behind the scenes
- Solution is to fix one variable, so that you can study the other ones

Cognitive Biases and Fast/Slow Thinking [3]

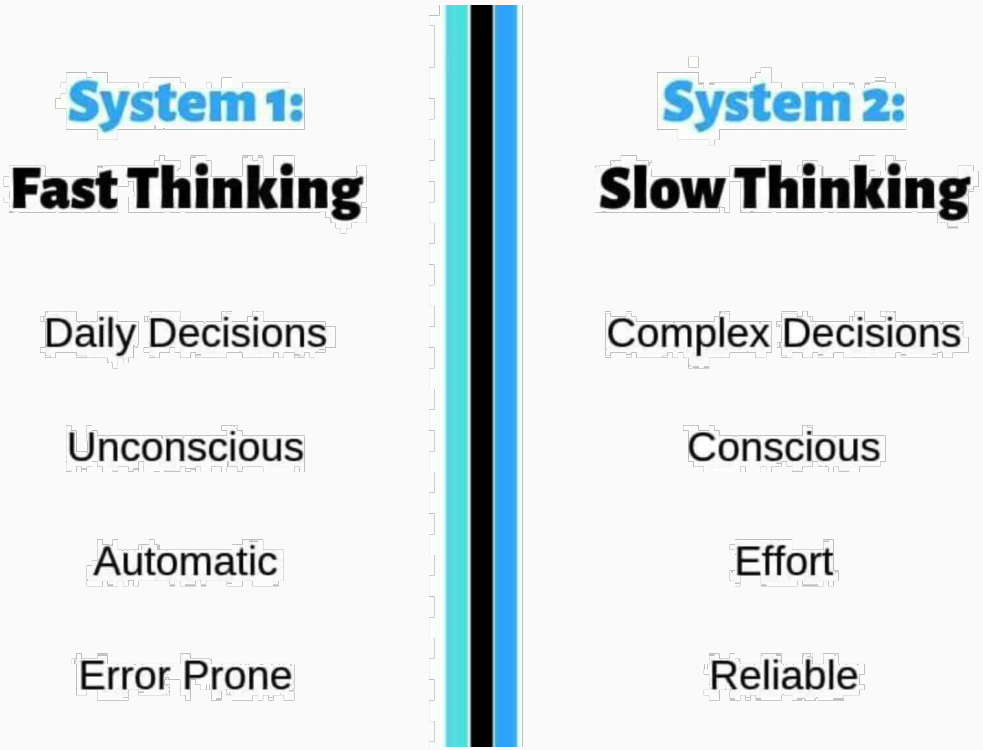


Figure 4: Nobel-winning Daniel Kahneman's book "*Thinking fast and slow*"

- **System 1** is the brain's automatic, fast, and intuitive mode of thinking. It relies on heuristics (mental shortcuts) to make quick judgments and decisions, often based on past experiences or stereotypes
- **System 2** is slower, more deliberative, and analytical. It kicks in when we need to process complex problems, weigh evidence carefully, and revise our beliefs based on reasoning.

Cognitive Biases

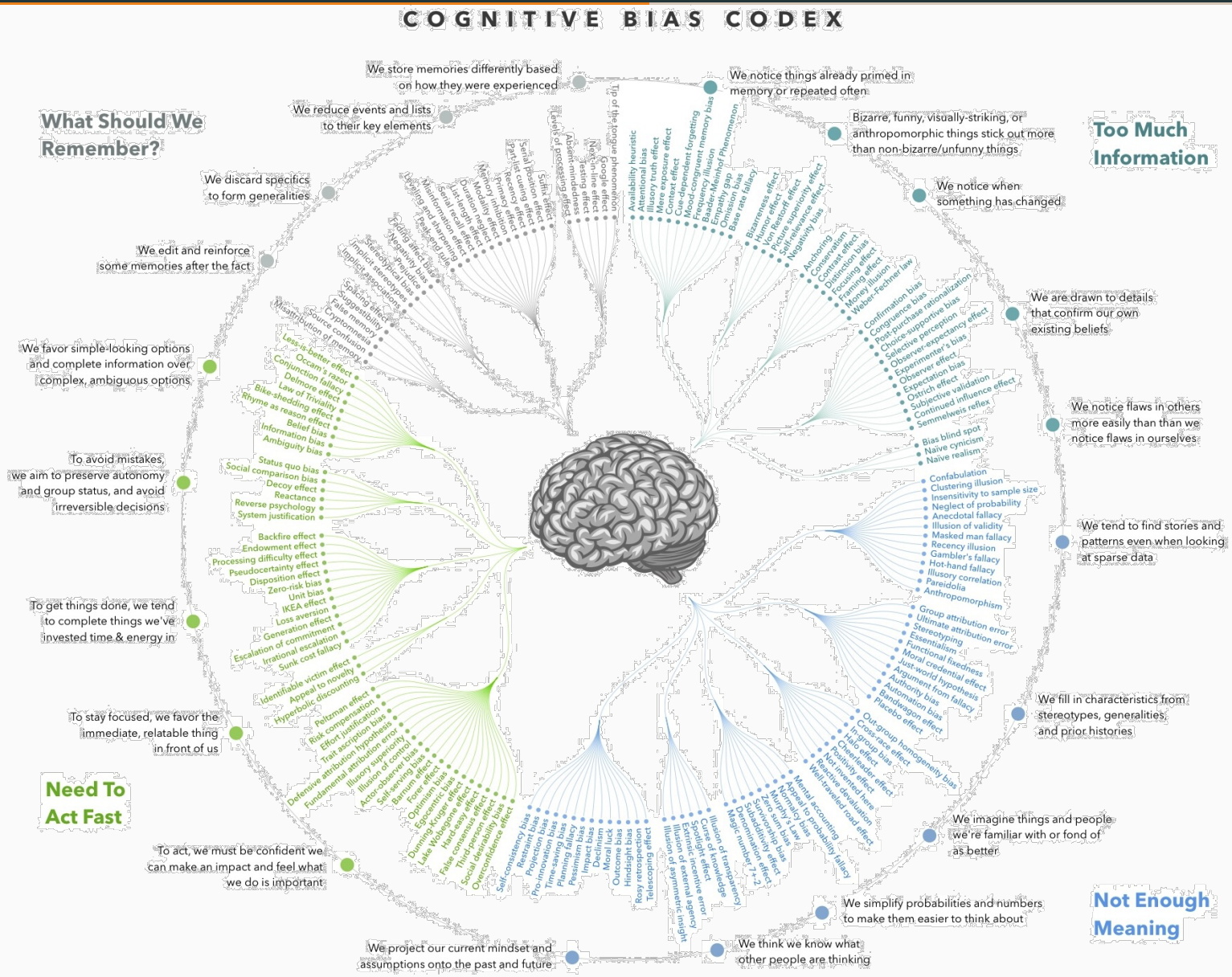


Figure 5: *"Simplify probabilities and numbers to make them easier to think about"* or *"Find stories and pattern even when looking at sparse data"*

Cognitive Biases

Closing this analogy part

- Just as priors in probability can be biased in a dataset, stereotypes are often formed from incomplete, skewed, or generalized information
- This can lead to systematic bias in how individuals process new information: what might be called "confirmation bias" in psychology or a skewed prior in probability theory.
- **ML models are trained on biased data can develop biased priors** leading to unfair or skewed prediction
- Similar to how **individuals may develop and act on biased stereotypes**.

Cognitive Biases

Closing this analogy part

- Just as priors in probability can be biased in a dataset, stereotypes are often formed from incomplete, skewed, or generalized information
- This can lead to systematic bias in how individuals process new information: what might be called "confirmation bias" in psychology or a skewed prior in probability theory.
- **ML models are trained on biased data can develop biased priors** leading to unfair or skewed prediction
- Similar to how **individuals may develop and act on biased stereotypes**.

In conclusion:

- Data can be biased because of spurious correlations due to hazard or confounding variables,
- The model will take advantage of this bias
- Like a human would do

Outline : Biases

Generalization and Approximation errors

Correlation vs Causations

Biases

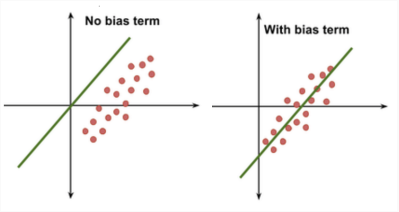
Trade-offs and Dilemmas

Overall

What can be biases?

A bias can be a deviation from the norm, the mean, or from the zero:

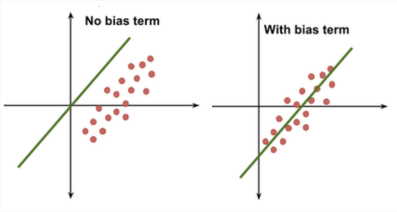
- A bias in a linear model to fit data



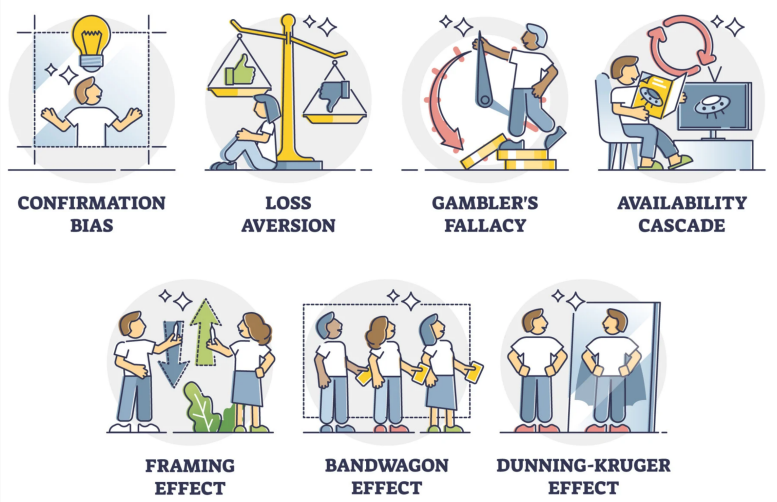
What can be biases?

A bias can be a deviation from the norm, the mean, or from the zero:

- A bias in a linear model to fit data



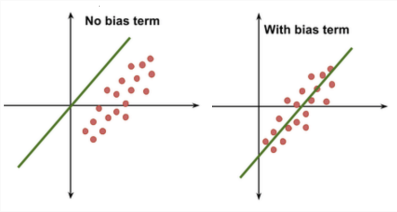
- A cognitive bias: availability bias, confirmation bias, Dunning-Kruger effect, ...



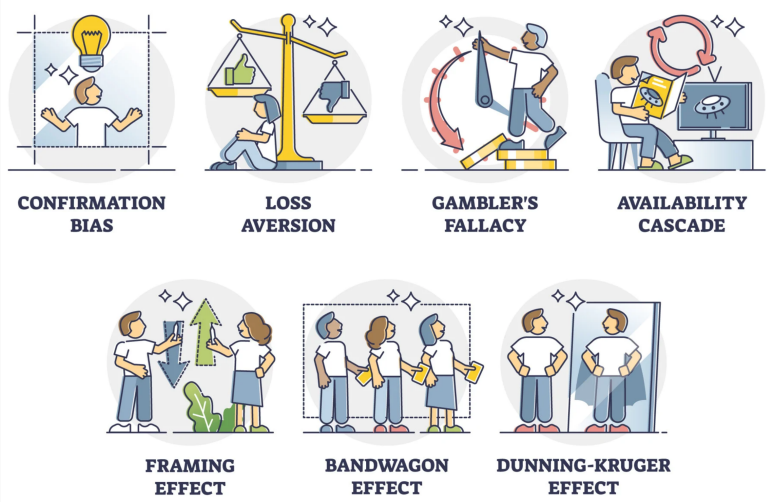
What can be biases?

A bias can be a deviation from the norm, the mean, or from the zero:

- A bias in a linear model to fit data



- A cognitive bias: availability bias, confirmation bias, Dunning-Kruger effect, ...

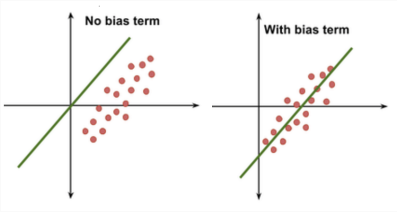


- A social bias like a cultural bias, people have different norms

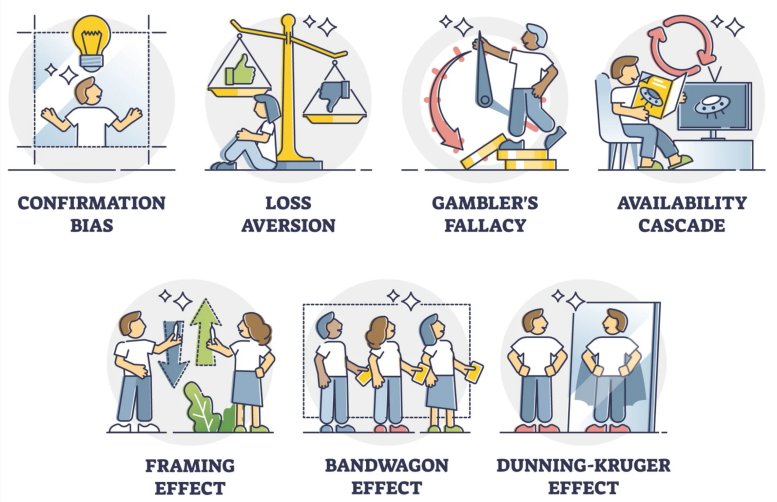
What can be biases?

A bias can be a deviation from the norm, the mean, or from the zero:

- A bias in a linear model to fit data



- A cognitive bias: availability bias, confirmation bias, Dunning-Kruger effect, ...



- A social bias like a cultural bias, people have different norms

In a decision-making process, a bias can be seen as a change of decision actioned by a non-causal variable.

What can be Biases in Predictive Models?

Definition

A deviation from a defined or expected norm. In AI, models often reflect biases present in training data, leading to skewed outputs.

Bias in AI Applications:

- **Loan Approvals**: AI models have been found to unfairly deny loans to women or minorities, reflecting historical lending biases.
- **Hiring Systems**: Biased algorithms favor men over equally qualified women, perpetuating gender inequality in tech fields.
- **Criminal Justice**: AI tools like COMPAS have led to higher incarceration rates for minority groups based on biased risk assessments.

What can be Biases in Predictive Models?

Definition

A deviation from a defined or expected norm. In AI, models often reflect biases present in training data, leading to skewed outputs.

Bias in AI Applications:

- **Loan Approvals**: AI models have been found to unfairly deny loans to women or minorities, reflecting historical lending biases.
- **Hiring Systems**: Biased algorithms favor men over equally qualified women, perpetuating gender inequality in tech fields.
- **Criminal Justice**: AI tools like COMPAS have led to higher incarceration rates for minority groups based on biased risk assessments.

AI models amplify existing societal biases, often reinforcing inequalities instead of mitigating them.

Biases

Fairness

Generally, when talking about unfair models, we are looking for negative biases toward certain target groups.

This can happens in different ways:

- Positiveness/Negativeness over target groups: sentiment more negative for arabic names, recidivism prediction higher for black people, ...
- Heterogeneous performances over target groups: face recognition system that works badly for Asian users, ASR only works for Castilian or Mexican Spanish, ...

Past Bias for Future Predictions

- Predictive model are learning pattern over data to make predictions
- The data reflects the past, but we want to predict the future.

Example

The model has learned that girls have less chance to go get a high-pay job.

Past Bias for Future Predictions

- Predictive model are learning pattern over data to make predictions
- The data reflects the past, but we want to predict the future.

Example

The model has learned that girls have less chance to go get a high-pay job.

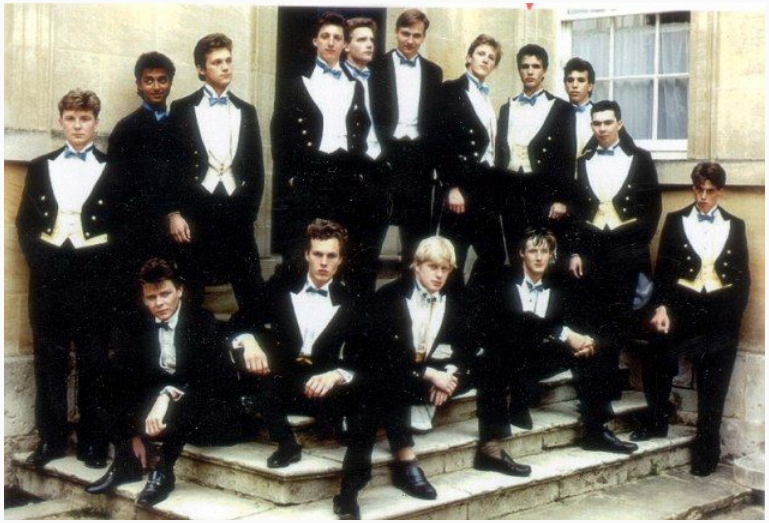
Do we want a model relying on past spurious correlation to predict the future?

Spurious Correlations

- Predictive model are learning pattern over data to make predictions
- Predictive models are relying on whatever they can, and it can be spurious correlations at best, or even non-causal variables

Example

Let's take the extreme case of a 1 feature classifier that predicts whether or not someone will fail its exam based on the color of the skin.

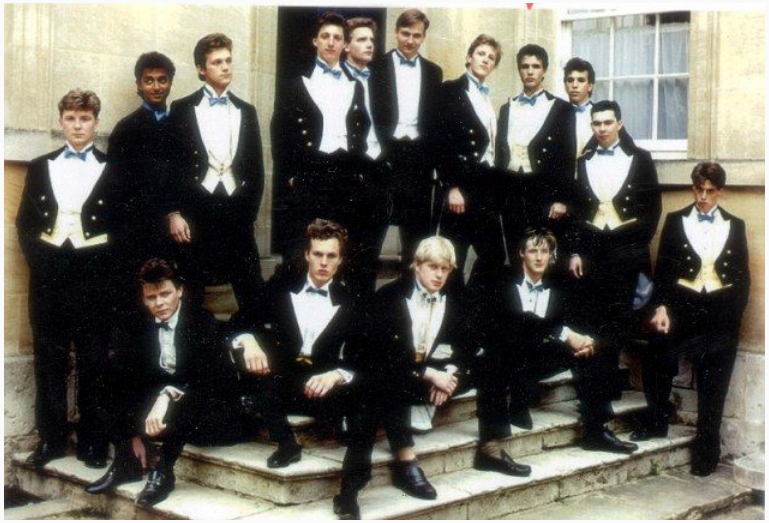


Spurious Correlations

- Predictive model are learning pattern over data to make predictions
- Predictive models are relying on whatever they can, and it can be spurious correlations at best, or even non-causal variables

Example

Let's take the extreme case of a 1 feature classifier that predicts whether or not someone will fail its exam based on the color of the skin.



Should we use this classifier to filter out "probable best students" that should enter to the university in the future?.

Outline : Trade-offs and Dilemmas

Trade-offs

These non-causal variables:

- Can have a strong predictive power, making them very useful for the performances
- However, these models are not robust to new data as they rely on spurious patterns between variables and prediction. They are over-fitting the past data that contains biases.

For example:

Trade-offs

These non-causal variables:

- Can have a strong predictive power, making them very useful for the performances
- However, these models are not robust to new data as they rely on spurious patterns between variables and prediction. They are over-fitting the past data that contains biases.

For example:

- If a model predict the salary that a person would have, it will give reduce salary to women. **The model is more accurate relying on this bias**.
- However, if society stop being misogynist, then women salary would be the same than men, than past model won't work that well. **The model is not robust to change in the data**.

Basic functional trade-offs in Cognition

- **Performance**: The ability of a system to achieve its intended result, meaning how well it accomplishes its main goal.
- **Efficiency**: The ability to perform its function using minimal resources, meaning doing more with less.
- **Robustness**: The system's capacity to maintain performance despite disruptions, meaning it can handle unexpected challenges or failures.
- **Flexibility**: The system's ability to maintain performance across a wide range of inputs, meaning it can adapt to new or unforeseen situations without breaking down.

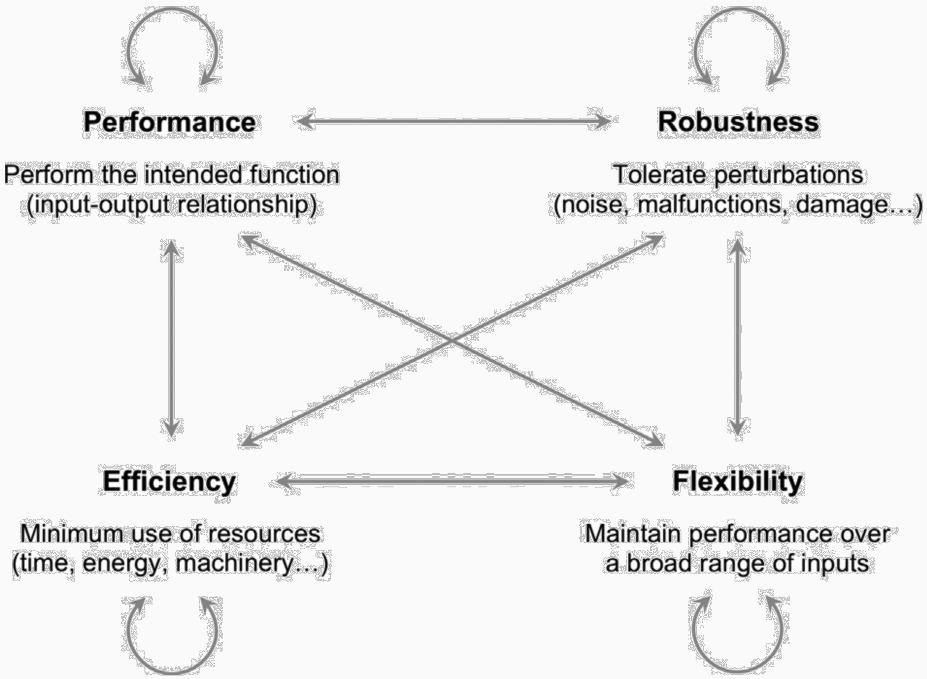


Figure 6: Functional Trade-offs [1]

Less-is-more effects

More Performing but Less Robust

Model that has higher predictive power is generally less robust, and less prone to adapt to new situations where causality is expressed in a different way

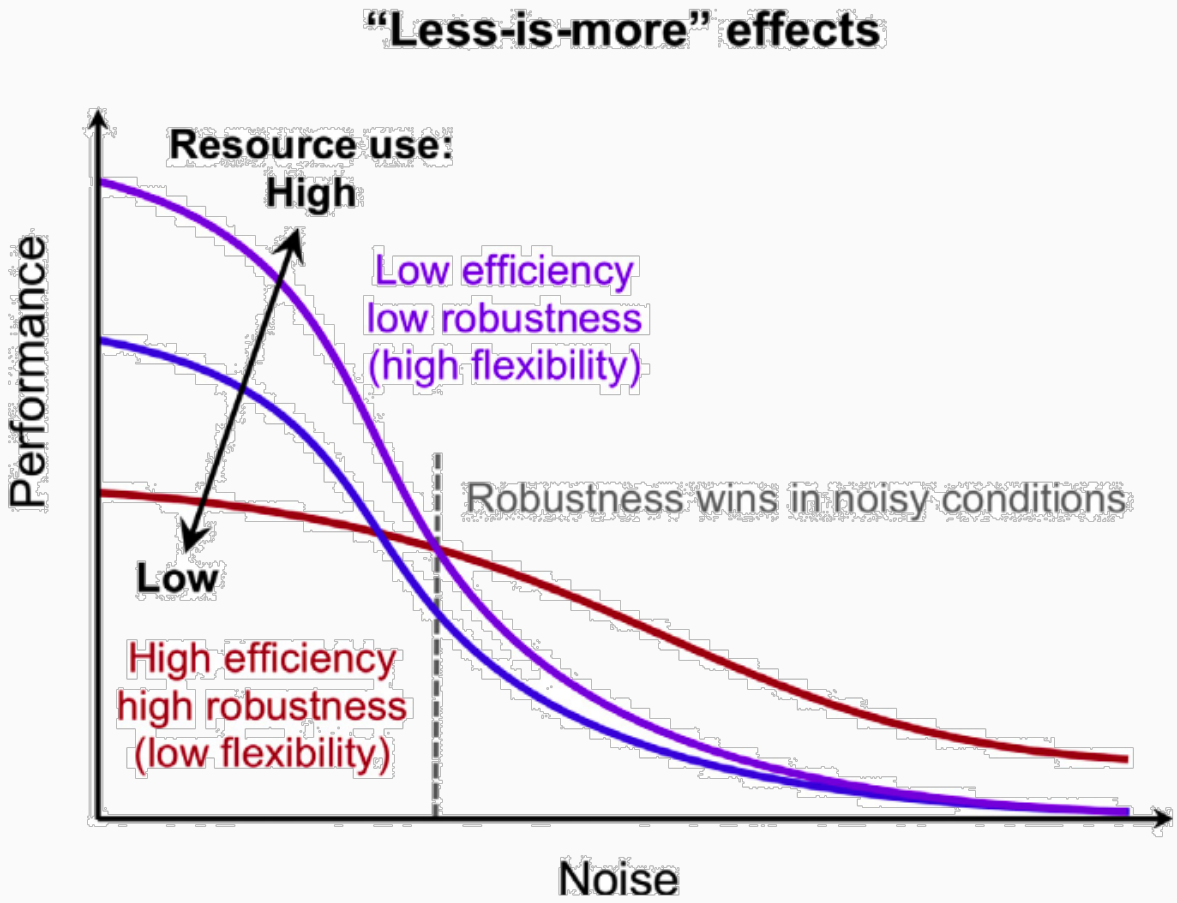


Figure 7: Models using less variables are more robust [1]

Less-is-more effects

More Performing but Less Robust

Model that has higher predictive power is generally less robust, and less prone to adapt to new situations where causality is expressed in a different way $\Rightarrow$ **If future does not reflect the past, model is bad**

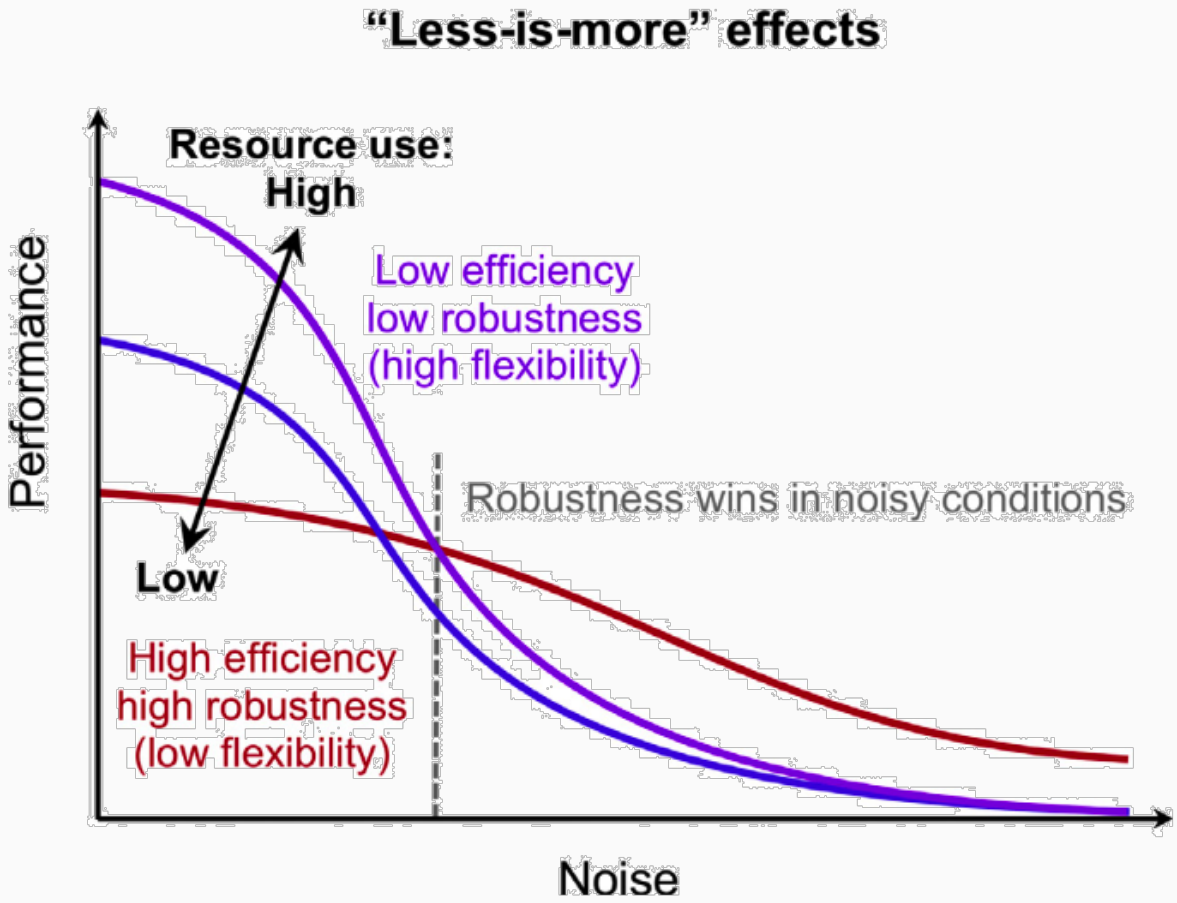


Figure 7: Models using less variables are more robust [1]

Outline : Overall

Techniques for biased data

If you have a dataset composed of causal variables (CV), non causal variables (NCV) and sensitive attributes (SA). You can:

- **Difference in weights**: Train different models using the CV and/or NCV and analyze the differences
- **Using only CV**: Try to learn the most efficient model using only CV
- **Difference in performances**: Test using the CV and/or NCV
- **Fix NCV**: One by one, and look at the differences in performances and in weights of the various models
- **Inspect performances**: Separate the test set using the SA, and look at the difference of performances

Two Visions for One Task

Predictive ML on SAT

I am creating a model to predict the SAT score of a student. What should I do with this model?

Case I. I use it to tag other students outside the data, hypothesising that the distribution would be the same, and that poorer students will fail their exam this year, because it is the reality. I know the reasons why the model predict are wrong but it gives good results!

Case II. I use it to predict the future, and hypothesising that in next year generation the poorer will still get worst results that the richer at their exam

Case I. is statistics-oriented: I am taking a sample of the population and try to predict the value over the whole (polling)

Case II. is ML-oriented: I am trying to predict what will happen next

Example on Job Candidate

Candidate

Suppose that you have to choose between two candidates for a job at your company. Both have equally impressive professional achievements , but one does not have a higher education degree. Which one should you choose?

- **Biased model** will know that the prior of having a higher education degree is higher and that's probably the better candidate.
- **Unbiased model** will choose the one without the higher education because he managed to achieve the same things as the other one but had the odds stacked against him.

In a Nutshell

Spurious Correlations vs. Causal Models

- Spurious Correlations: Improve accuracy but are not robust to real causal changes in data.
- Causal Models: More interpretable and generalizable but may reduce initial accuracy

Ethical Considerations

- Excluding sensitive variables may reduce bias but lower accuracy.
- Including them risks perpetuating inequality and damage target groups (like the snowy huskies or short-haired girls)

Bias detection

- Various techniques can be used for bias detection
- Causal ML techniques can also be used to train trustable models

Ressources

Online content:

- An interesting course on Causal Machine Learning
- A nice blogpost on the difference between correlation and causation

Classes:

- CC5117 class: *Algoritmos, Redes y Equidad: Análisis de Sistemas Sociotécnicos*
- CS109 class from Stanford: *Probability for Computer Scientists*
- CC6104 class: *Statistical Thinking*

Questions?

References i

M. Del Giudice and B. J. Crespi.
Basic functional trade-offs in cognition: An integrative framework.
Cognition, 179(June):56–70, 2018.

T. Given.
Spurious correlation #4,242, 2024.

D. Kahneman.
Thinking, Fast and Slow.
2011.

Predictive Power vs. Ethical Concerns

- **Predictive Power**: Including socio-economic variables may improve the model's accuracy, as there is often a correlation between a student's socio-economic background and their academic performance. Wealthier students might have better access to resources like tutoring, stable learning environments, and higher-quality schools, which can improve their performance.
- **Ethical Considerations**: While these variables may enhance the model's predictive power, using them raises concerns about fairness and bias. Predicting performance based on socio-economic factors can perpetuate systemic inequities, and such models might reinforce harmful stereotypes or lead to discriminatory decisions (e.g., resource allocation, academic tracking).

Spurious Correlations vs Causal Understanding

- **Spurious Correlations**: Socio-economic status (SES) is likely correlated with exam performance, but this relationship may not be directly causal. Other factors like the quality of the educational environment, parental support, or access to learning materials could be the underlying causes. A model that relies on SES as a predictive feature might exploit these correlations without truly addressing the root causes of academic performance.
- **Causal Understanding**: To develop a more robust and ethical model, one would need to identify the true causal drivers behind exam performance, such as cognitive ability, educational support, or motivation. These factors could be more directly linked to performance, though they might be harder to measure or access.

Fairness and Model Bias

Using socio-economic variables might lead to models that are unfair to certain groups. This is a key aspect of algorithmic fairness, where:

- **Fairness Through Unawareness**: One approach would be to exclude socio-economic variables to avoid biasing the model against certain groups, even if it sacrifices some predictive accuracy.
- **Fairness Through Representation**: Alternatively, the inclusion of these variables could be used with the goal of identifying and mitigating disparities (e.g., flagging students who may need extra support), but only if care is taken to interpret the results with caution and to ensure that decisions are made fairly.