



Agenda



Rune Nielsen, PhD
Data scientist & AI
specialist at SAS Institute

- What is data ethics?
 - What should we be aware of?
- Why does it raise problems?
 - How do we handle it?
- Which opportunities does it create?



What is data ethics?

Data pitfalls

How many pitfalls are directly related to the data selection?

DATA FALLACIES TO AVOID



CHERRY PICKING

Selecting results that fit your claim and excluding those that don't.



DATA DREDGING

Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.



SURVIVORSHIP BIAS

Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.



COBRA EFFECT

Setting an incentive that accidentally produces the opposite result to the one intended. Also known as a Perverse Incentive.



FALSE CAUSALITY

Falsely assuming when two events appear related that one must have caused the other.



GERRYMANDERING

Manipulating the geographical boundaries used to group data in order to change the result.



SAMPLING BIAS

Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.



GAMBLER'S FALLACY

Mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice versa).



HAWTHORNE EFFECT

The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.



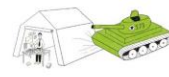
REGRESSION TOWARDS THE MEAN

When something happens that's unusually good or bad, it will revert back towards the average over time.

	2017		2018	
	APPLICANTS	ACCEPTED	APPLICANTS	ACCEPTED
MATHS	100	20	200	40
ENGLISH	100	10	200	20
TOTAL	200	30	400	60

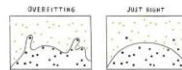
SIMPSON'S PARADOX

When a trend appears in different subsets of data but disappears or reverses when the groups are combined.



MCNAMARA FALLACY

Relying solely on metrics in complex situations and losing sight of the bigger picture.



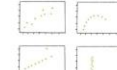
OVERFITTING

Creating a model that's overly tailored to the data you have and not representative of the general trend.



PUBLICATION BIAS

Interesting research findings are more likely to be published, distorting our impression of reality.



DANGER OF SUMMARY METRICS

Only looking at summary metrics and missing big differences in the raw data.

Data pitfalls

How many pitfalls are directly related to the data selection?

DATA FALLACIES TO AVOID

CHERRY PICKING
Selecting results that fit your claim and excluding those that don't.

DATA DREGGING
Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.

SURVIVORSHIP BIAS
Drawing conclusions from an incomplete set of data, because that data has "survived" some selection criteria.

WANTED / AWARDED
COBRA EFFECT
Getting an incentive that accidentally produces the opposite result to the one intended. Also known as a **PERVERSE INCENTIVE**.

FALSE CAUSALITY
Falsely assuming when two events appear related that one must have caused the other.

GERYMANDERING
Manipulating the geographic boundaries used to group data in order to change the result.

SAMPLING BIAS
Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.

GAMBLER'S FALLACY
Mistakenly believing that because something has happened more frequently than usual, it's somewhat likely to happen in future (and vice versa).

HAWTHORNE EFFECT
The act of monitoring someone can affect their behavior, leading to spurious findings. Also known as the **Observer Effect**.

REGRESSION TOWARDS THE MEAN
When something happens that's certainly good or bad, it will revert back towards the average over time.

SIMPSON'S PARADOX
When a trend appears in different subsets of data but disappears or reverses when the groups are combined.

MCMANAMA FALLACY
Relying solely on metrics to compare situations and losing sight of the bigger picture.

OVERFITTING
Creating a model that's overly tailored to the data you know and not representative of the general trend.

PUBLICATION BIAS
Interesting research findings are more likely to be published, distorting our impression of reality.

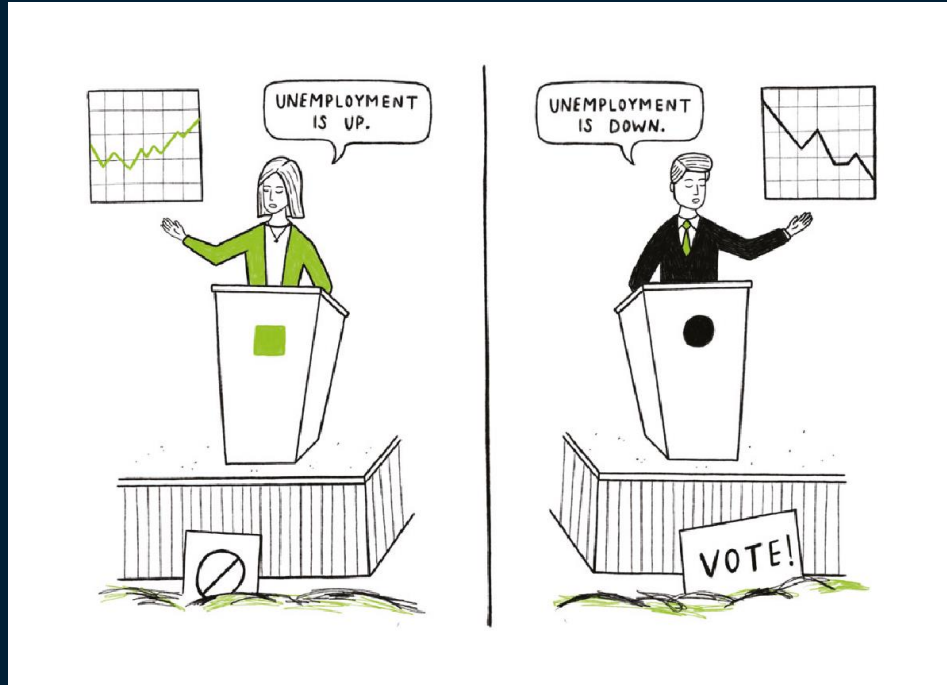
DANGER OF SUMMARY METRICS
Only looking at summary metrics and missing big differences in the raw data.

GECKOBOARD.COM

Reid Murray @data-literacy@geckoboard.com

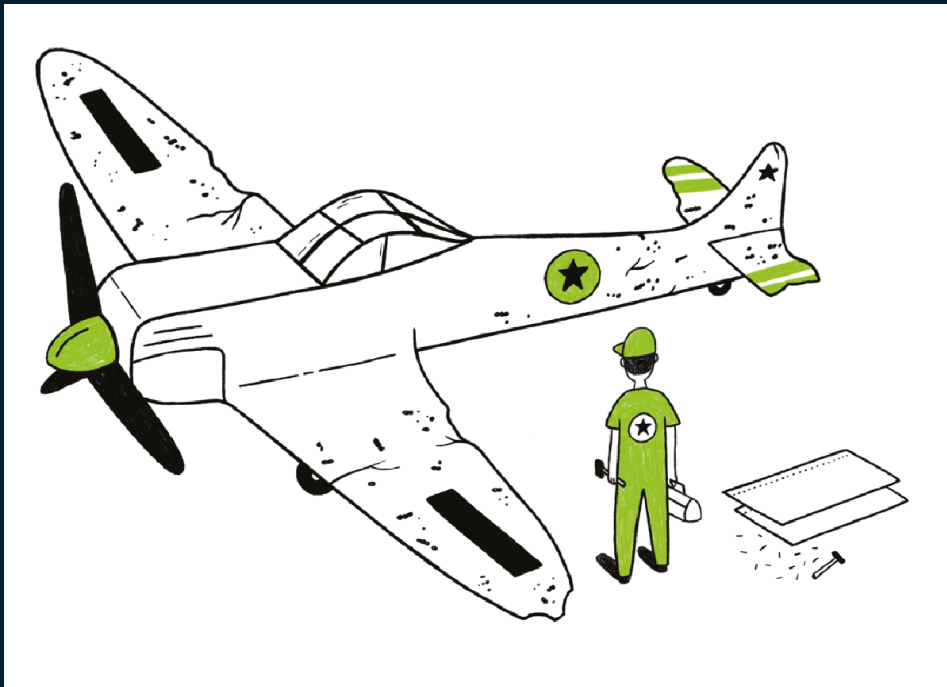
What is data ethics?

Cherry Picking



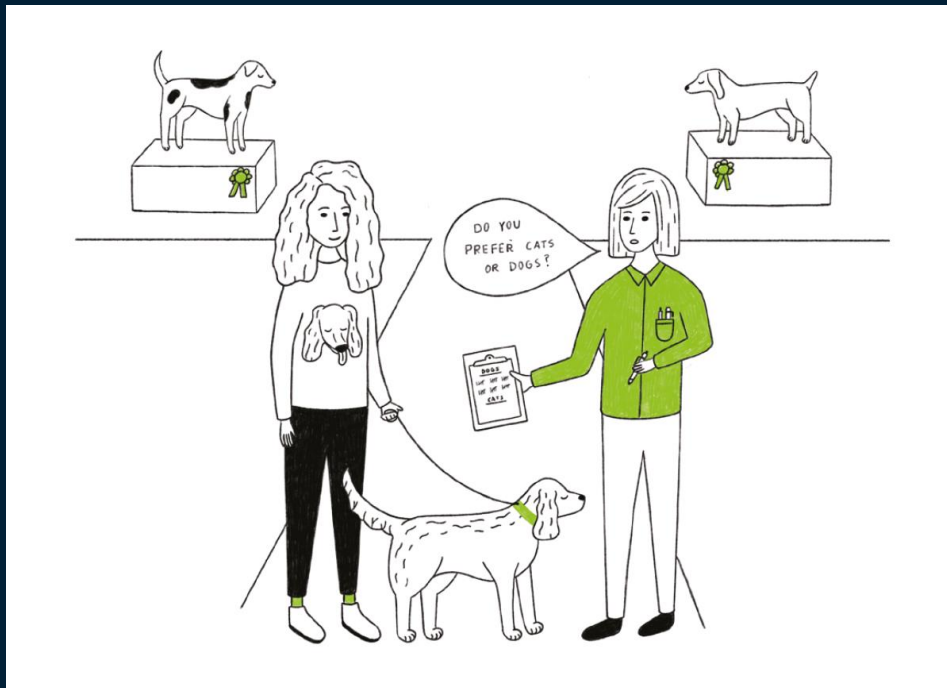
What is data ethics?

Survivorship bias



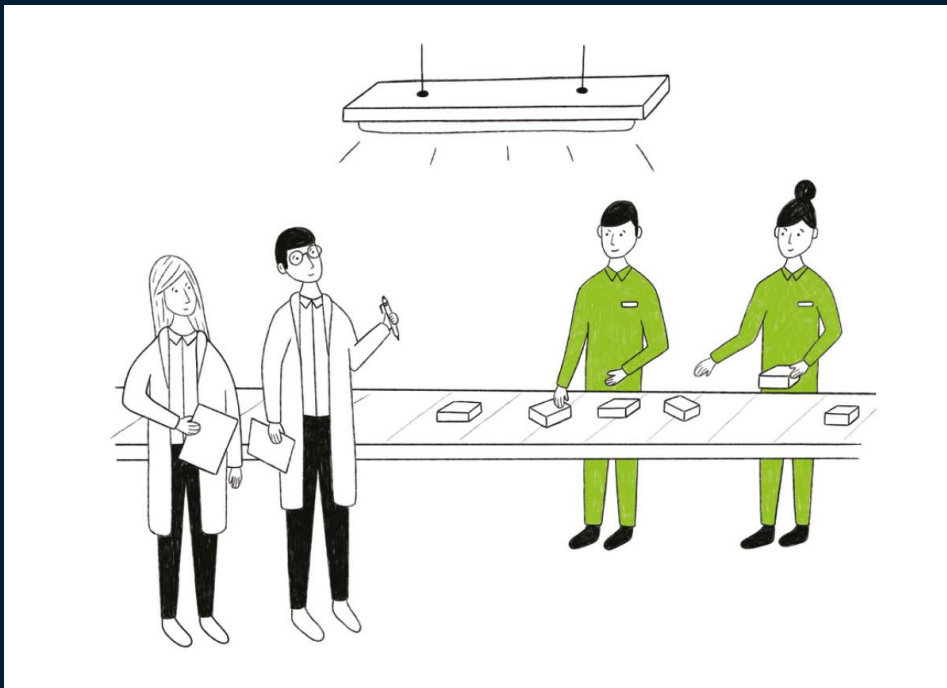
What is data ethics?

Sampling bias



What is data ethics?

Hawthorne's effect



Data pitfalls

How many pitfalls are directly related to how we treat data under the model development?

DATA FALLACIES TO AVOID



CHERRY PICKING

Selecting results that fit your claim and excluding those that don't.



DATA DREDGING

Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.



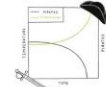
SURVIVORSHIP BIAS

Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.



COBRA EFFECT

Setting an incentive that accidentally produces the opposite result to the one intended. Also known as a Perverse Incentive.



FALSE CAUSALITY

Falsely assuming when two events appear related that one must have caused the other.



GERRYMANDERING

Manipulating the geographical boundaries used to group data in order to change the result.



SAMPLING BIAS

Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.



GAMBLER'S FALLACY

Mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice versa).



HAWTHORNE EFFECT

The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.



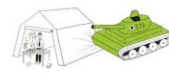
REGRESSION TOWARDS THE MEAN

When something happens that's unusually good or bad, it will revert back towards the average over time.



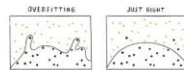
SIMPSON'S PARADOX

When a trend appears in different subsets of data but disappears or reverses when the groups are combined.



MCNAMARA FALLACY

Relying solely on metrics in complex situations and losing sight of the bigger picture.



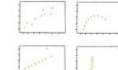
OVERFITTING

Creating a model that's overly tailored to the data you have and not representative of the general trend.



PUBLICATION BIAS

Interesting research findings are more likely to be published, distorting our impression of reality.



DANGER OF SUMMARY METRICS

Only looking at summary metrics and missing big differences in the raw data.

Data pitfalls

How many pitfalls are directly related to how we treat data under the model development?

DATA FALLACIES TO AVOID

CHERRY PICKING
Selecting results that fit your claim and excluding those that don't.

DATA DREDGING
Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.

SURVIVORSHIP BIAS
Drawing conclusions from an incomplete set of data. Airline that data has "survived" some selection criteria.

COOK'S EFFECT
Getting an increase that accidentally produces the opposite result to the randomized data treated as a placebo treatment.

FALSE CAUSALITY
Falsely assuming when two events appear related that one must have caused the other.

GERRYMANDERING
Manipulating the geographical boundaries used to group data in order to change the result.

SAMPLING BIAS
Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.

GAMBLER'S FALLACY
Mistakenly believing that because something has happened more frequently than usual, it's a "win" or "loss" that is more likely to happen in future (and vice versa).

HAWTHORNE EFFECT
The act of measuring someone can affect their behavior, leading to spurious findings. Also known as the Observer Effect.

REGRESSION TOWARDS THE MEAN
When something happens that's certainly good or bad, it will revert back towards the average over time.

SIMPSON'S PARADOX
When a trend appears in different subsets of data but disappears or reverses when the groups are combined.

MCMANAMA FALLACY
Relying solely on metrics to compare situations and losing sight of the bigger picture.

OVERFITTING vs **JUST RIGHT**
OVERFITTING: Creating a model that's overly tailored to the data you have and not representative of the general trend.
JUST RIGHT: A model that fits the data well but is also generalizable.

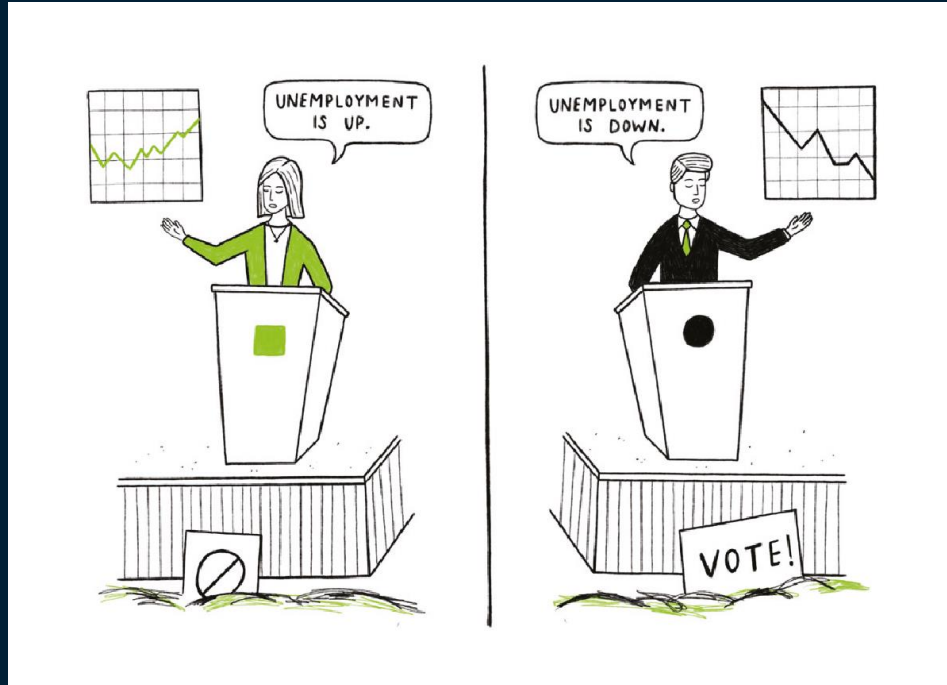
PUBLICATION BIAS
Interesting research findings are more likely to be published, distorting our impression of reality.

DANGER OF SUMMARY METRICS
Only looking at summary metrics and missing big differences in the raw data.

GECKBOARD.COM
Reid Murray, data-literacy@geckboard.com

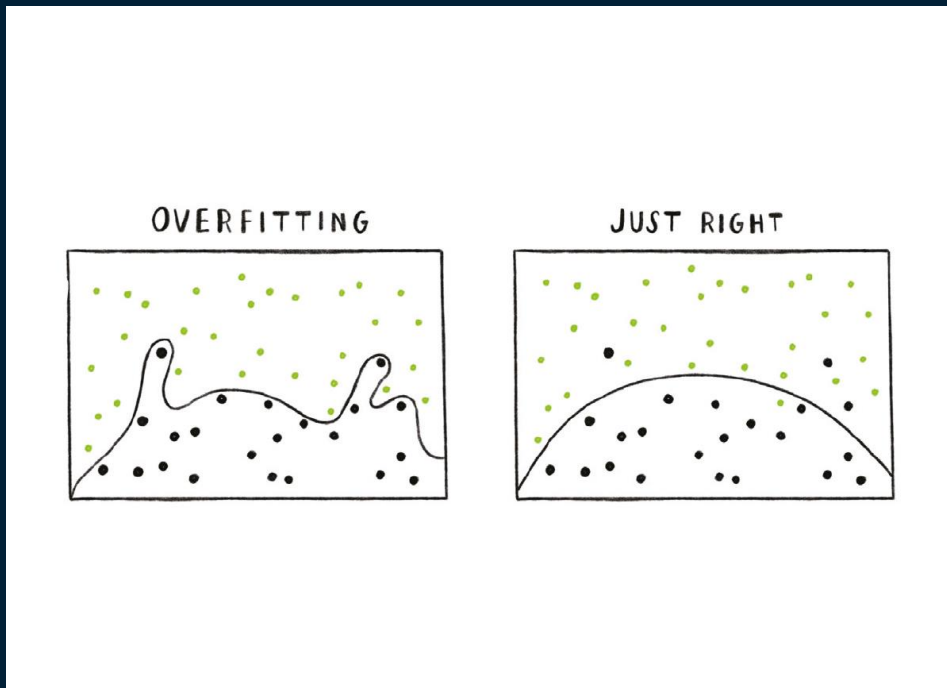
What is data ethics?

Cherry Picking



What is data ethics?

Overfitting



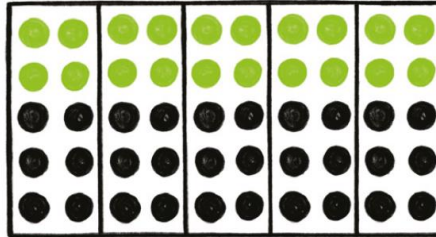
What is data ethics?

Data dredging

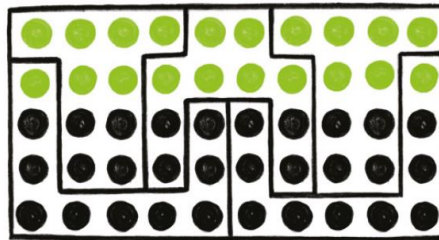


What is data ethics?

Gerrymandering



BLACK WINS



GREEN WINS

What is data ethics?

Simpson's Paradox




APPLICATION SUCCESS RATE

	MALE	FEMALE
SUBJECT 1	14 % (168 of 1200)	15 % (270 of 1800)
SUBJECT 2	50 % (400 of 800)	51 % (102 of 200)
TOTAL	28 % (568 of 2000)	19 % (372 of 2000)


??

Data pitfalls in relation to data treatment.


DATA FALLACIES TO AVOID




CHERRY PICKING
Selecting results that fit your claim and excluding those that don't.




DATA DREDGING
Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.




SURVIVORSHIP BIAS
Drawing conclusions from an incomplete set of data.




COBRA EFFECT
Setting an incentive that accidentally produces the opposite result to the intended data point as a positive incentive.




FALSE CAUSALITY
Falsely assuming when two events appear related that one must have caused the other.




GERRYMANDERING
Manipulating the geographical boundaries used to group data in order to change the result.




SAMPLING BIAS
Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.




GAMBLER'S FALLACY
Mistakenly believing that because something has happened more frequently than usual, it's somewhat likely to happen in future (and vice versa).



HAWTHORNE EFFECT
The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.




REGRESSION TOWARDS THE MEAN
When something happens that's certainly good or bad, it will revert back towards the average over time.




	MALES	FEMALES
ADMITTED	10%	15%
APPLICANTS	100	100
ADMITTED	10%	15%
APPLICANTS	100	100
ADMITTED	10%	15%
APPLICANTS	100	100


SIMPSON'S PARADOX
When a trend appears in different subsets of data but disappears or reverses when the groups are combined.




MCNAMARA FALLACY
Relying solely on metrics in complex situations and losing sight of the bigger picture.




OVERFITTING
Creating a model that's overly tailored to the data you have and not representative of the general trend.



JUST BURY




PUBLICATION BIAS
Interesting research findings are more likely to be published, distorting our impression of reality.



DANGER OF SUMMARY METRICS
Only looking at summary metrics and missing big differences in the raw data.

GECKBOARD.COM

Reid murphy@data-literacy@geckboard.com

A close-up, artistic photograph of a person's face, primarily the nose and eye area, covered in a dense layer of small, multi-colored particles that appear to be glowing or reflecting light. The colors are predominantly blue, purple, and orange. The background is dark, and the overall mood is mysterious and futuristic. A diagonal line separates this image from the text on the right.

What is the cause of the
issue?

What is the cause of the issue?

Ordinary least squares (OLS)

OLS assumptions:

1. Random sample
2. Linear in parameters
3. No multi-collinearity
4. The independent variables are exogenous
5. The error term is homoscedastic
6. The error term is normally distributed with mean 0 and constant variance (Hypothesis test)

Only one point is about data?

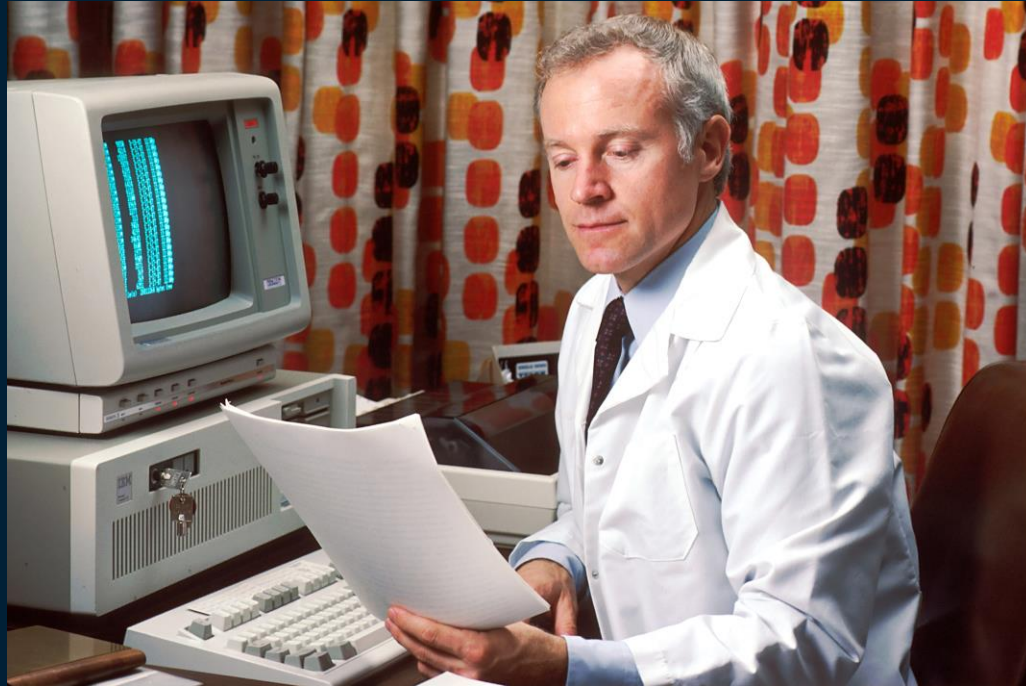
Are you allowed to plot your data before an hypothesis test?

Culture as a part of the modelling toolbox

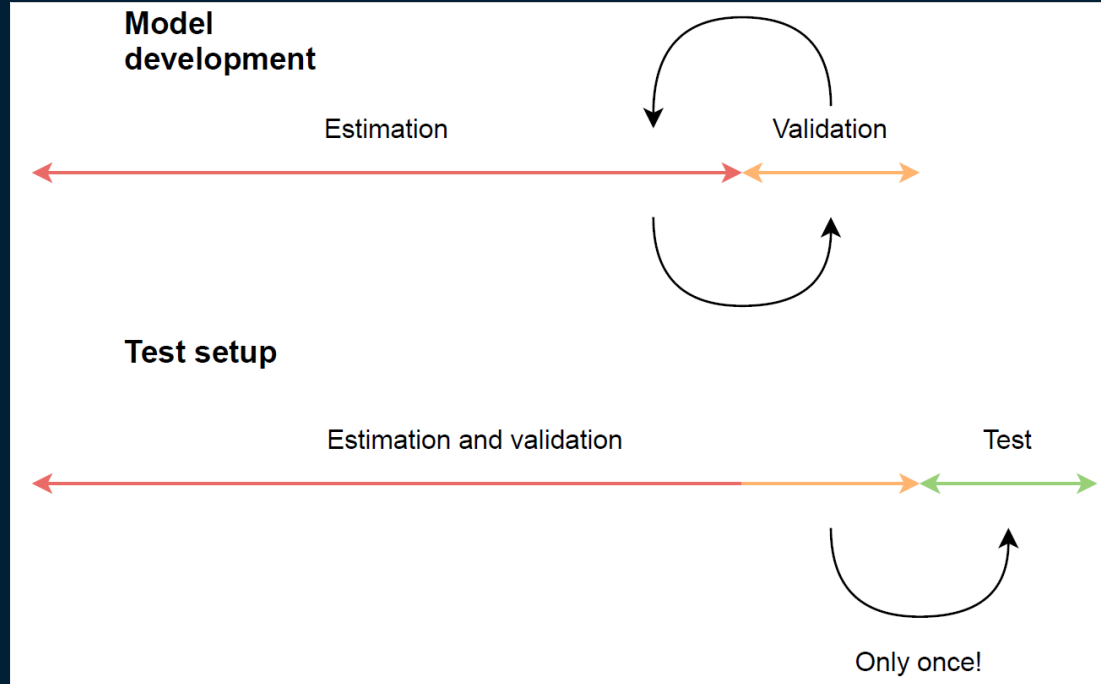
How do we handle the pitfalls?



How do we handle the pitfalls?



How do we handle the pitfalls?



How do we handle the pitfalls?

```
proc forest data=mycas.inData;  
  partition fraction(test=0.25 validate=0.25);  
  ...  
run;
```



Which opportunities does it
create?

Which opportunities does it create?

Ordinary least squares (OLS)

1. Random sample
2. Linear in parameters
3. No multi-collinearity
4. The independent variables are exogenous
5. The error term is homoscedastic
6. The error term is normally distributed with mean 0 and constant variance (Hypothesis test)

Which opportunities does it create?

- Flexible model development
 - Enabling iterative model development
- Business focus
 - Focus on the performance of the model (e.g. forecast precision)
 - Less time on assumptions

References

- Data fallacies to avoid: [Geckoboard](#)
- Billeder fra [Unsplash](#): [Ramón Salinero](#), [Myriam Jessier](#), [Joshua Sortino](#), [David Pupaza](#), [ThisisEngineering RAEng](#), [Firmbee.com](#), [Isaac Smith](#), [C Drying](#), [h heyerlein](#), [Emily Morter](#), [National Cancer Institute](#), [Patrick Weissenberger](#)



Rune Hjorth Nielsen

Providing insights within data science and AI
for SAS customer advisory

