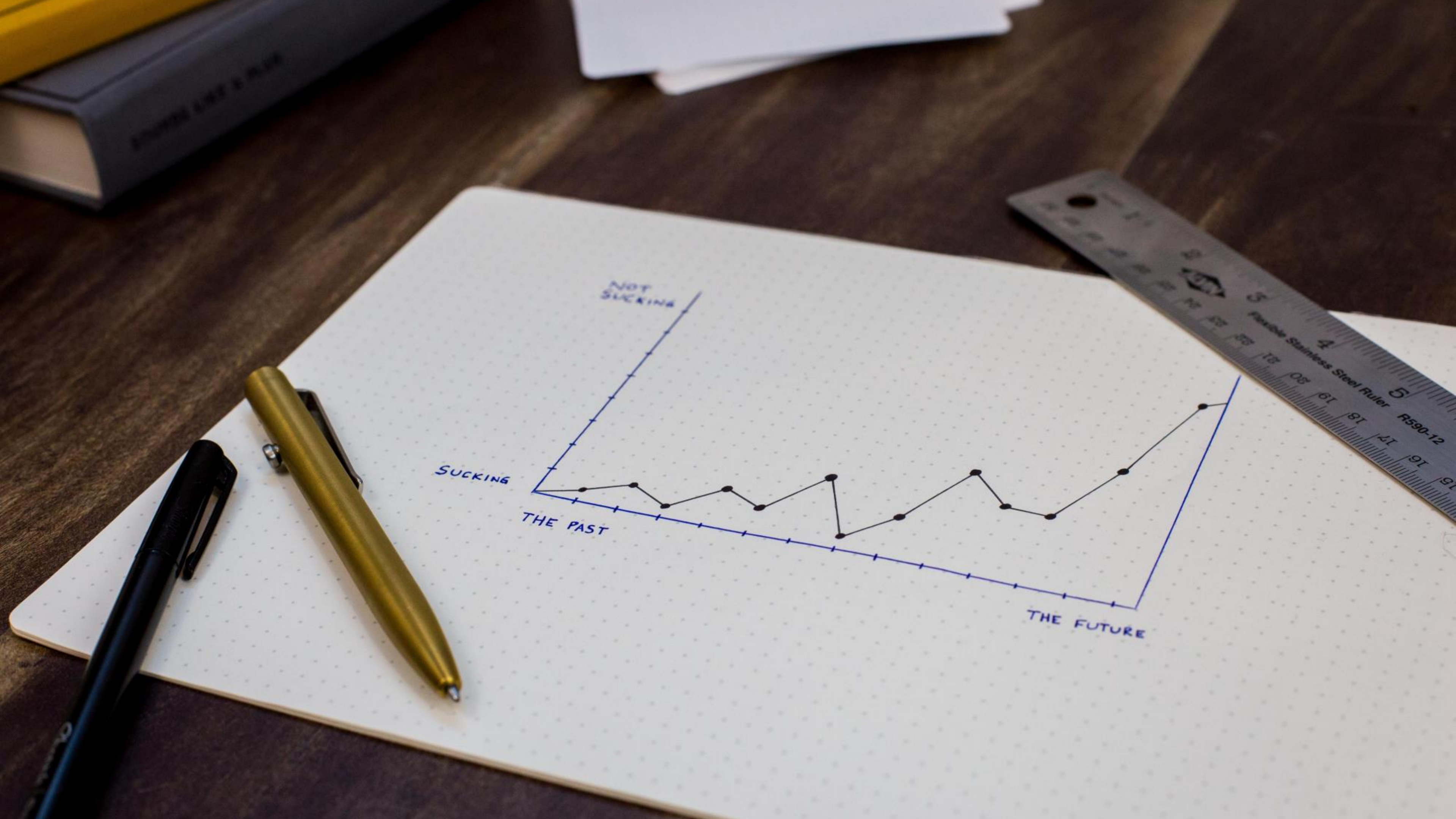
A data scientist’s job according to unsplash

Copyright © SAS Institute Inc. All rights reserved.


NOT SUCKING
SUCKING
THE PAST
THE FUTURE
Flexible Stainless Steel Ruler
R590-12

Rune Nielsen, PhD

Data scientist & AI specialist at SAS Institute

Agenda

- 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?

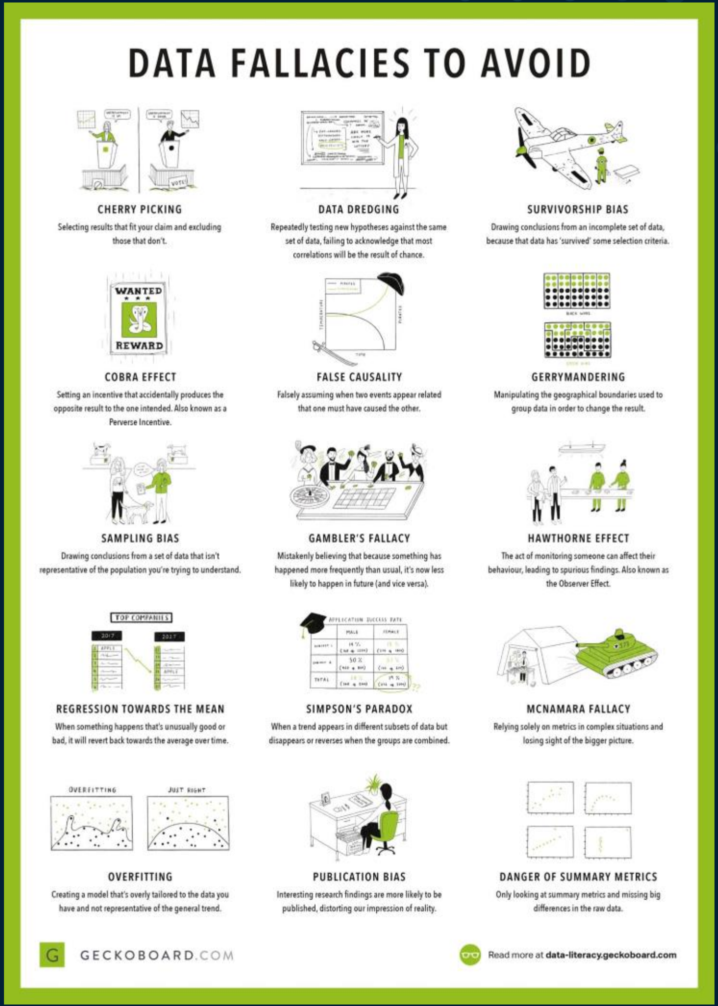
Copyright © SAS Institute Inc. All rights reserved.

What is data ethics?

Copyright © SAS Institute Inc. All rights reserved.

Data pitfalls

How many pitfalls are directly related to the data selection?



Copyright © SAS Institute Inc. All rights reserved.

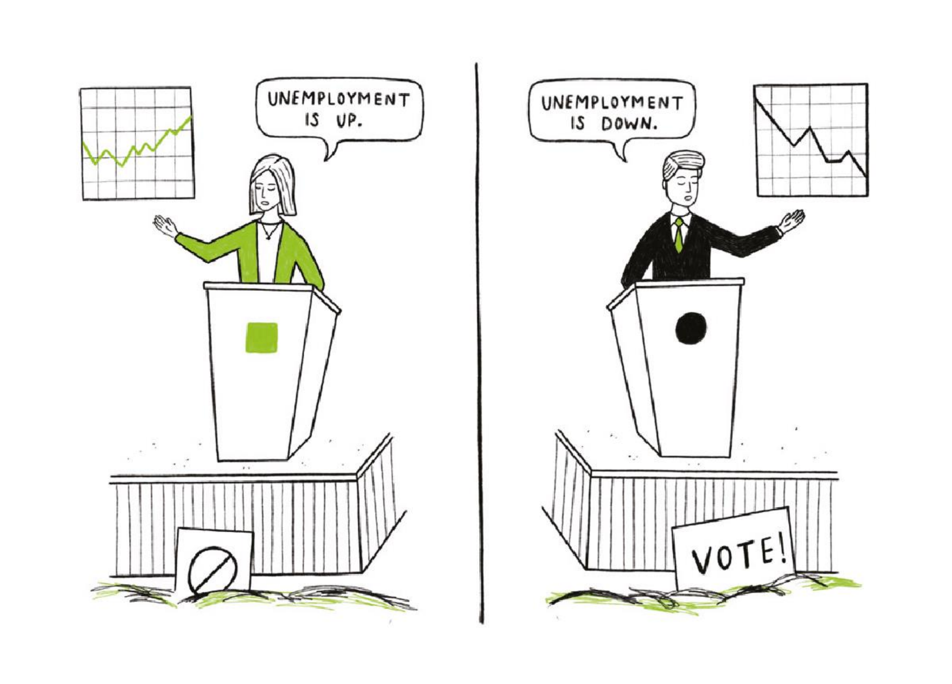
Data pitfalls

How many pitfalls are directly related to the data selection?

Copyright © SAS Institute Inc. All rights reserved.

What is data ethics?

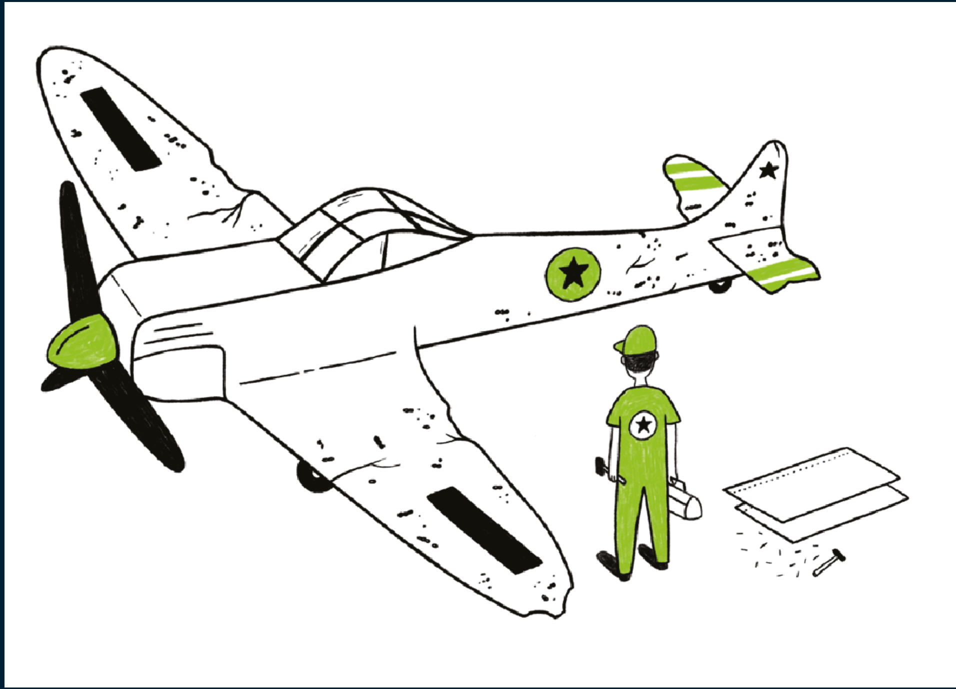
Cherry Picking



Copyright © SAS Institute Inc. All rights reserved.

What is data ethics?

Survivorship bias



Copyright © SAS Institute Inc. All rights reserved.

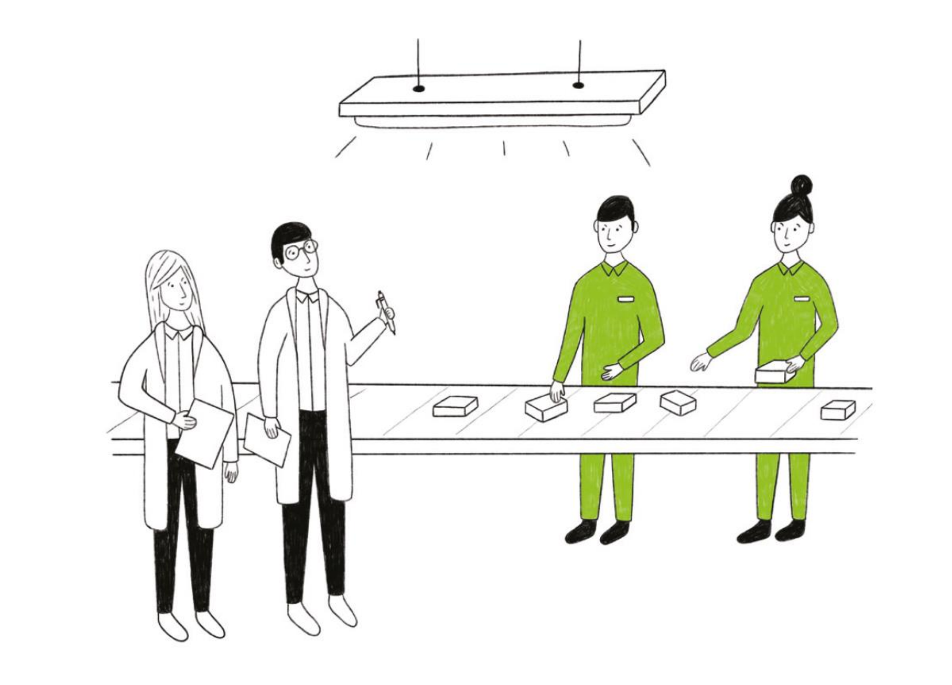
What is data ethics?

Sampling bias

Copyright © SAS Institute Inc. All rights reserved.

What is data ethics?

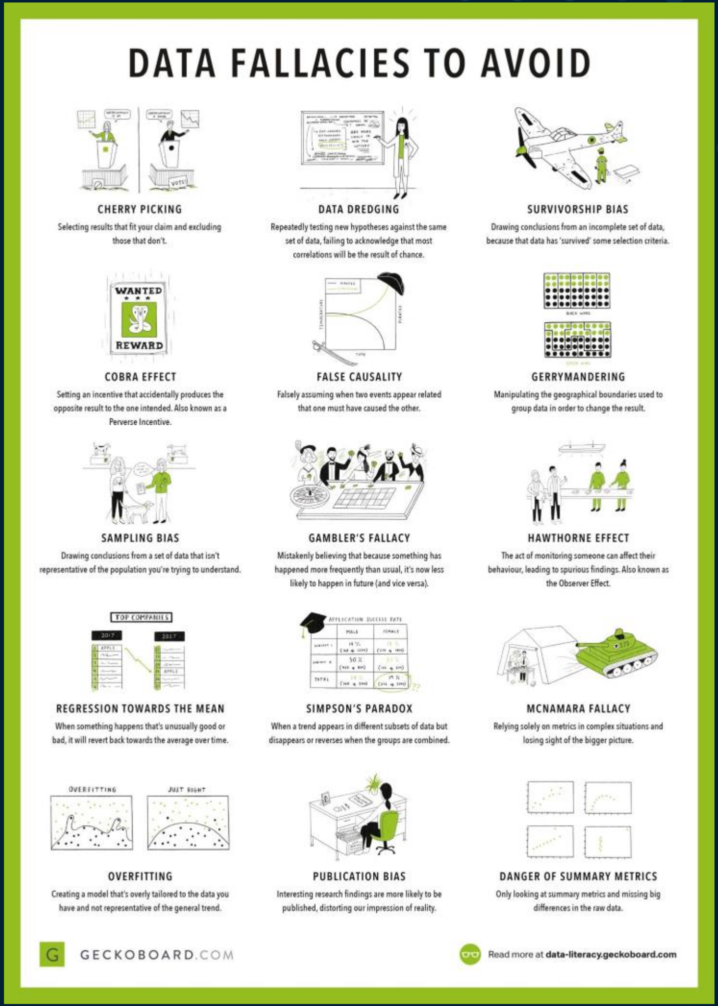
Hawthorne's effect



Copyright © SAS Institute Inc. All rights reserved.

Data pitfalls

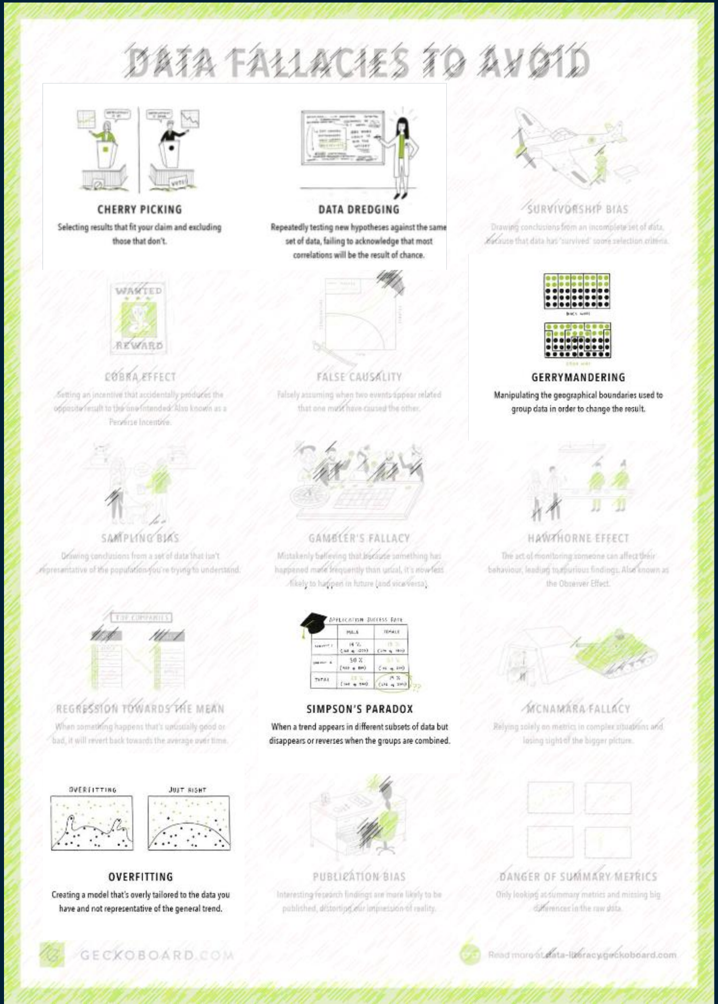
How many pitfalls are directly related to how we treat data under the model development?



Copyright © SAS Institute Inc. All rights reserved.

Data pitfalls

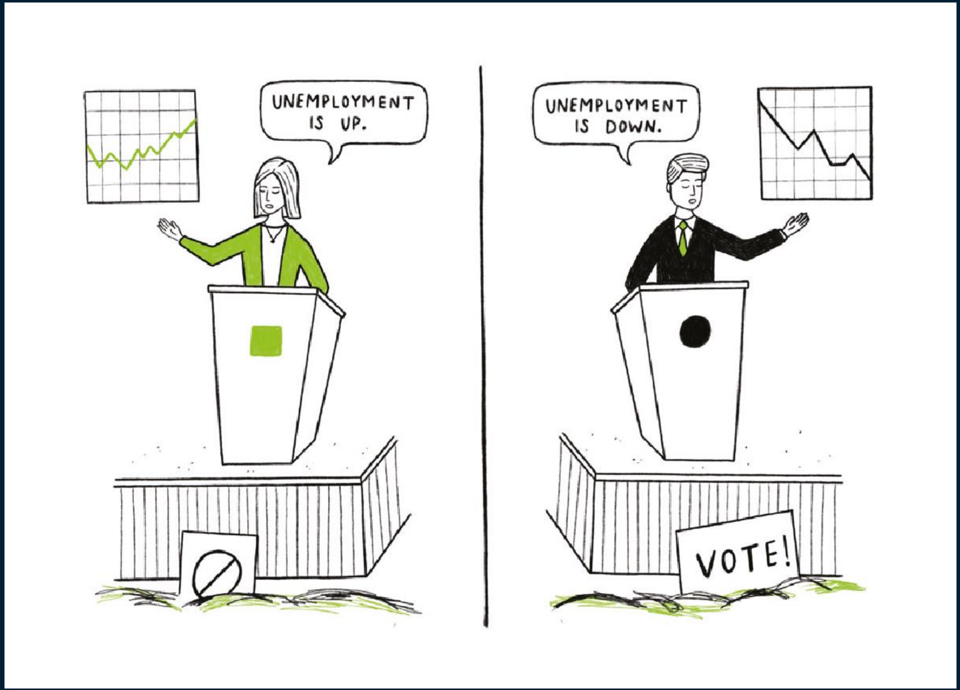
How many pitfalls are directly related to how we treat data under the model development?



Copyright © SAS Institute Inc. All rights reserved.

What is data ethics?

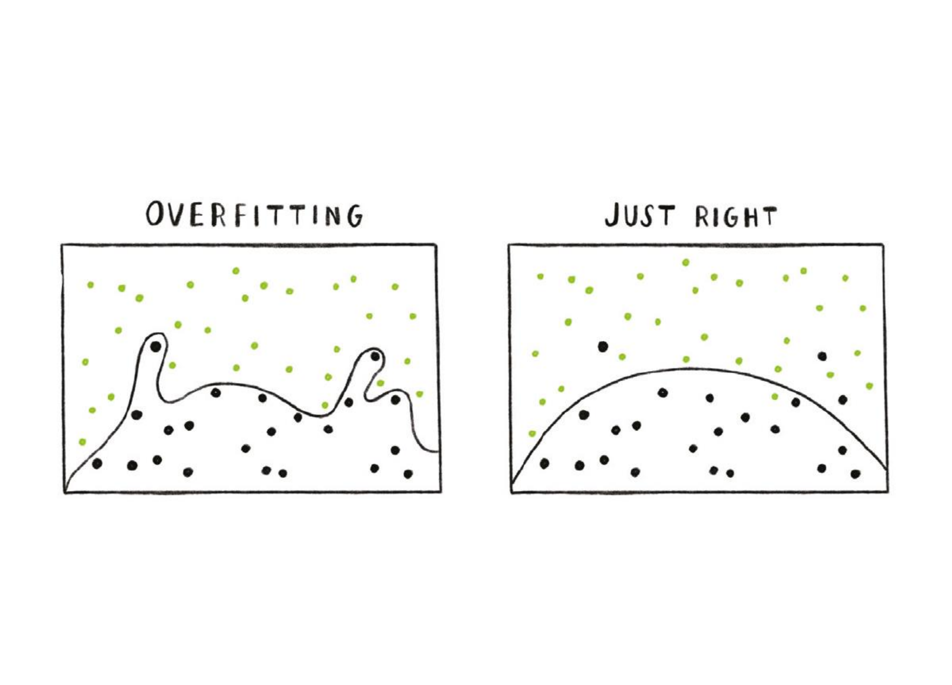
Cherry Picking



Copyright © SAS Institute Inc. All rights reserved.

What is data ethics?

Overfitting



Copyright © SAS Institute Inc. All rights reserved.

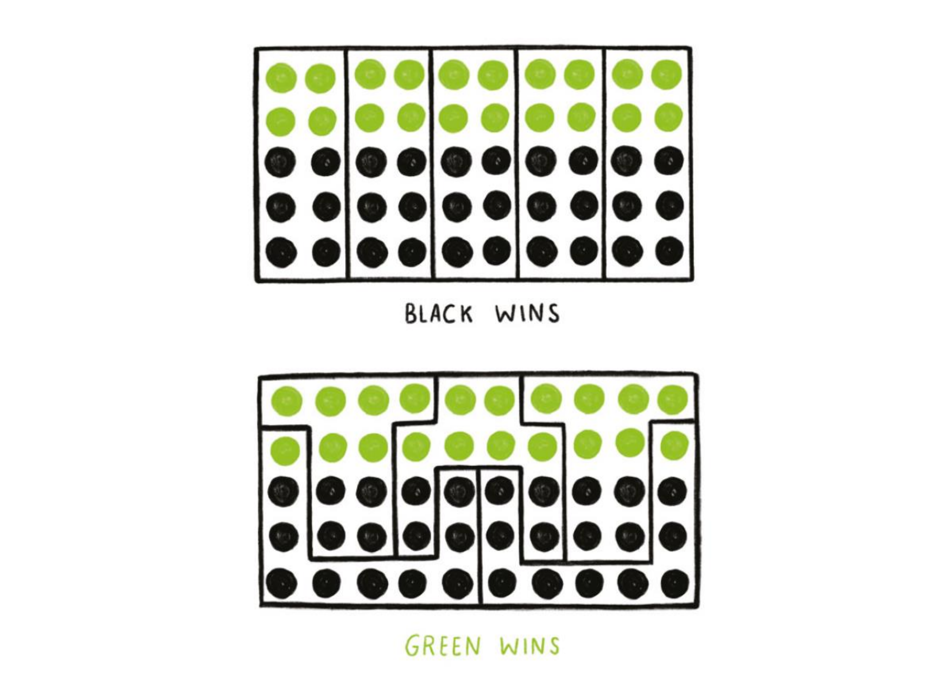
What is data ethics?

Data dredging

Copyright © SAS Institute Inc. All rights reserved.

What is data ethics?

Gerrymandering



Copyright © SAS Institute Inc. All rights reserved.

What is data ethics?

Simpson's Paradox

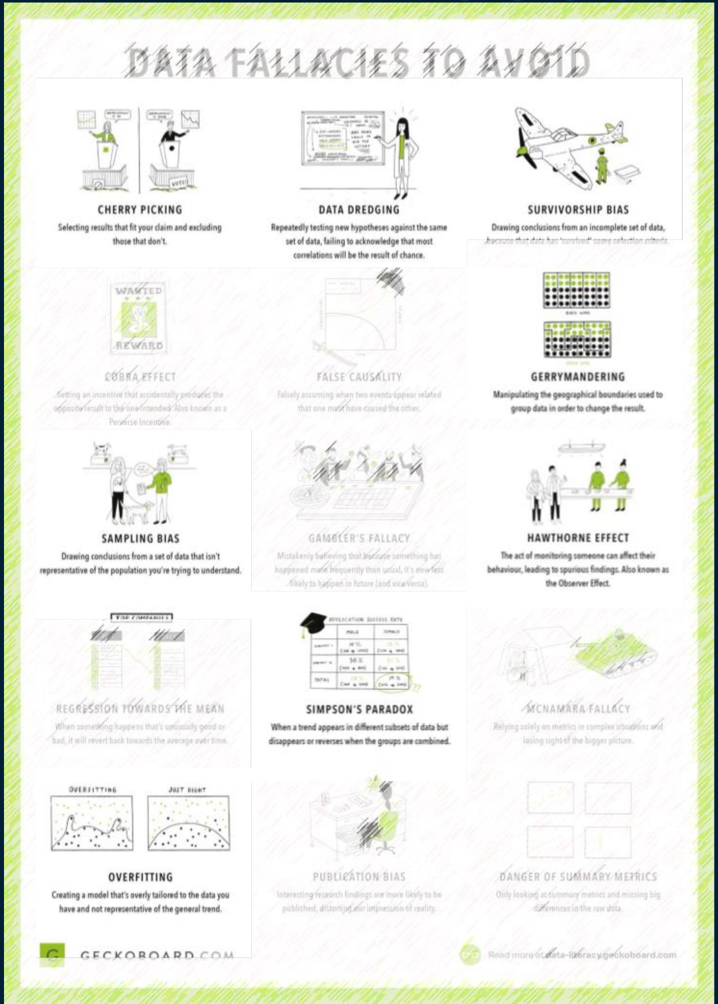
APPLICATION SUCCESS RATE

	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)

??

Copyright © SAS Institute Inc. All rights reserved.

Data pitfalls in relation to data treatment.



Copyright © SAS Institute Inc. All rights reserved.

What is the cause of the issue?

Copyright © SAS Institute Inc. All rights reserved.

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?

Copyright © SAS Institute Inc. All rights reserved.

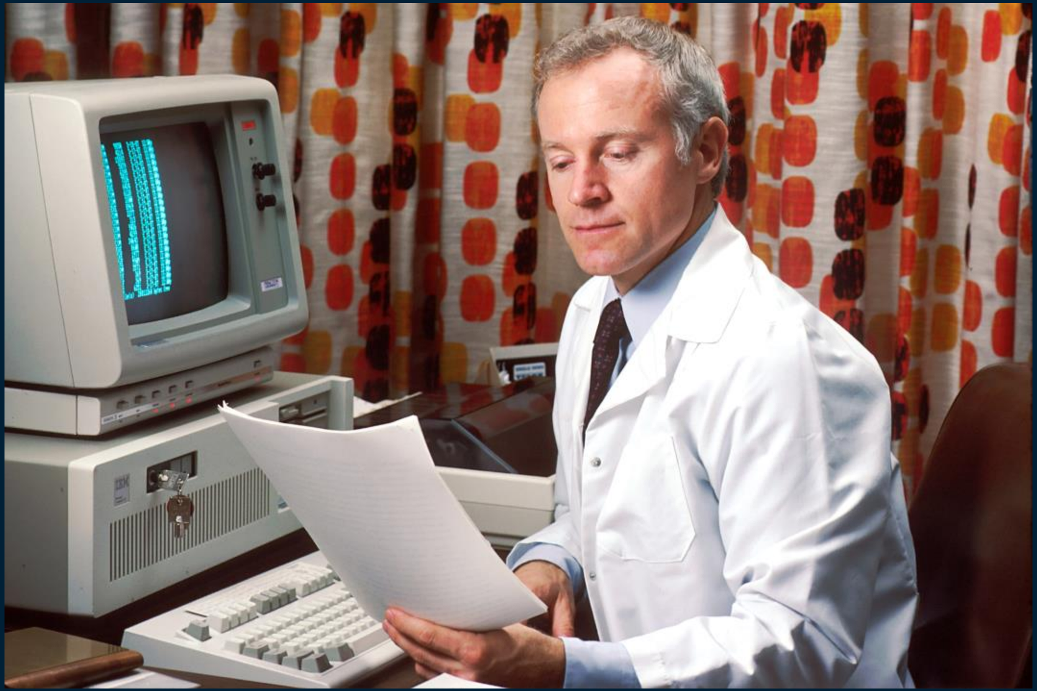
Culture as a part of the modelling toolbox

Copyright © SAS Institute Inc. All rights reserved.

How do we handle the pitfalls?

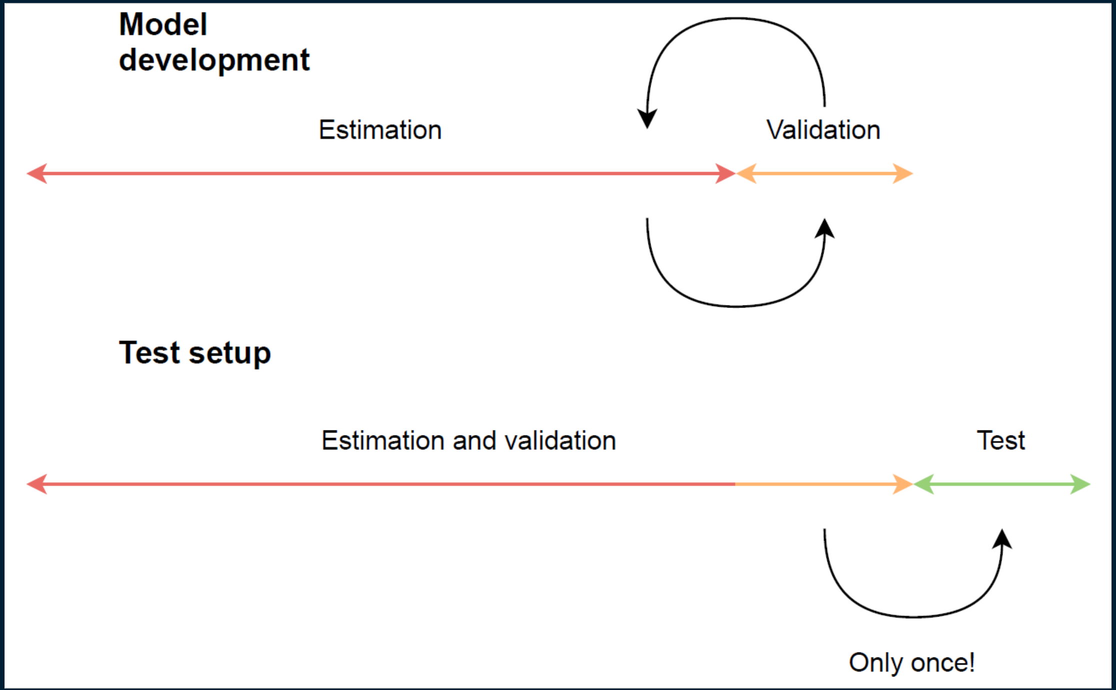
Copyright © SAS Institute Inc. All rights reserved.

How do we handle the pitfalls?



Copyright © SAS Institute Inc. All rights reserved.

How do we handle the pitfalls?



Copyright © SAS Institute Inc. All rights reserved.

How do we handle the pitfalls?

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

Copyright © SAS Institute Inc. All rights reserved.

Which opportunities does it create?

Copyright © SAS Institute Inc. All rights reserved.

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)

Copyright © SAS Institute Inc. All rights reserved.

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

Copyright © SAS Institute Inc. All rights reserved.

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

Copyright © SAS Institute Inc. All rights reserved.

Rune Hjorth Nielsen
Providing insights within data science and AI
for SAS customer advisory

Copyright © SAS Institute Inc. All rights reserved.