Missing Data

Soc 212C

Ian Lundberg

Learning goals for today

By the end of class, you will be able to

- make assumptions about missing data
  - missing completely at random
  - missing at random
- connect these assumptions to causal inference
- use two strategies for missing data
  - listwise deletion
  - multiple imputation

Abraham Wald

- b. 1902, Austria-Hungary
- Jewish, persecuted in WWII
- Fled to U.S. in 1938
- Namesake of the Wald test
- Statistical consultant for U.S. Navy in WWII

[1]

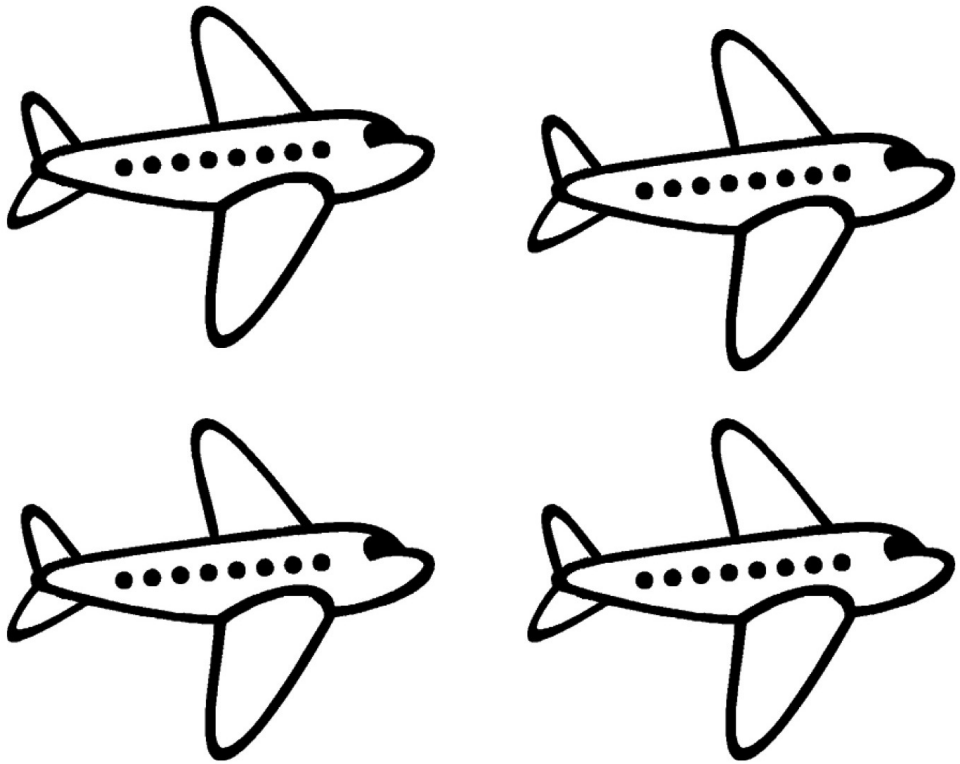
[1] PC: Wikimedia commons

Question: Where should armor be added to protect planes?

Data: Suppose we saw the following planes.[2]

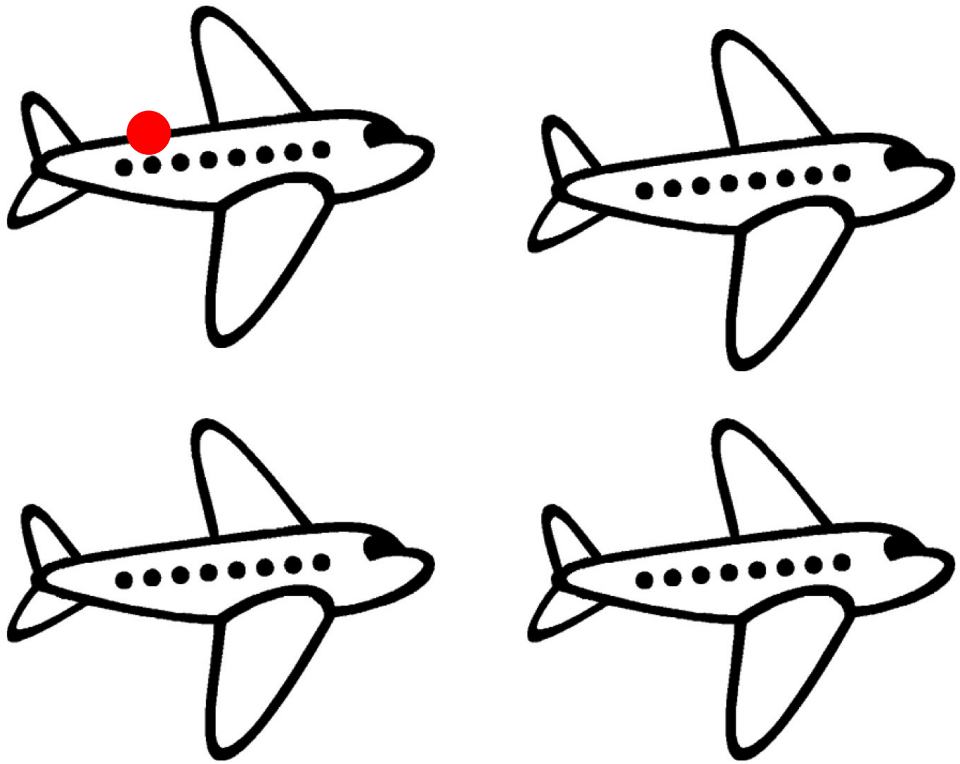
[2] Story told by Mangel and Samaniego 1984 [link].
Presentation style inspired by Joe Blitzstein. See the original here [link]

Observed data: Planes that came home with damage



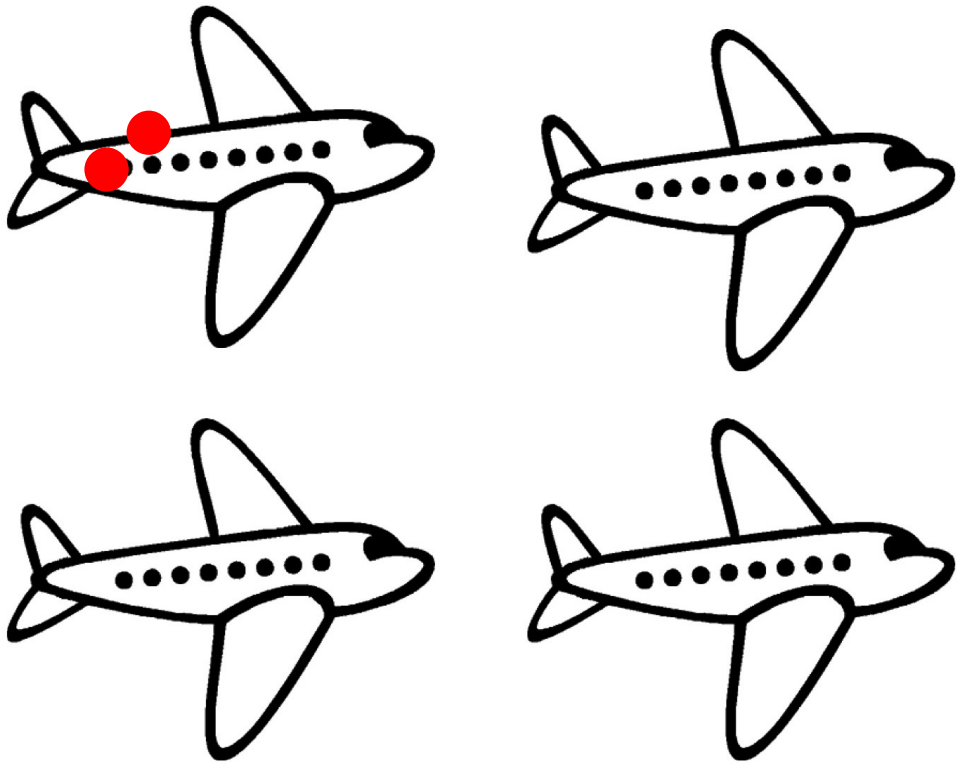
Q: Where would you add armor?

Observed data: Planes that came home with damage



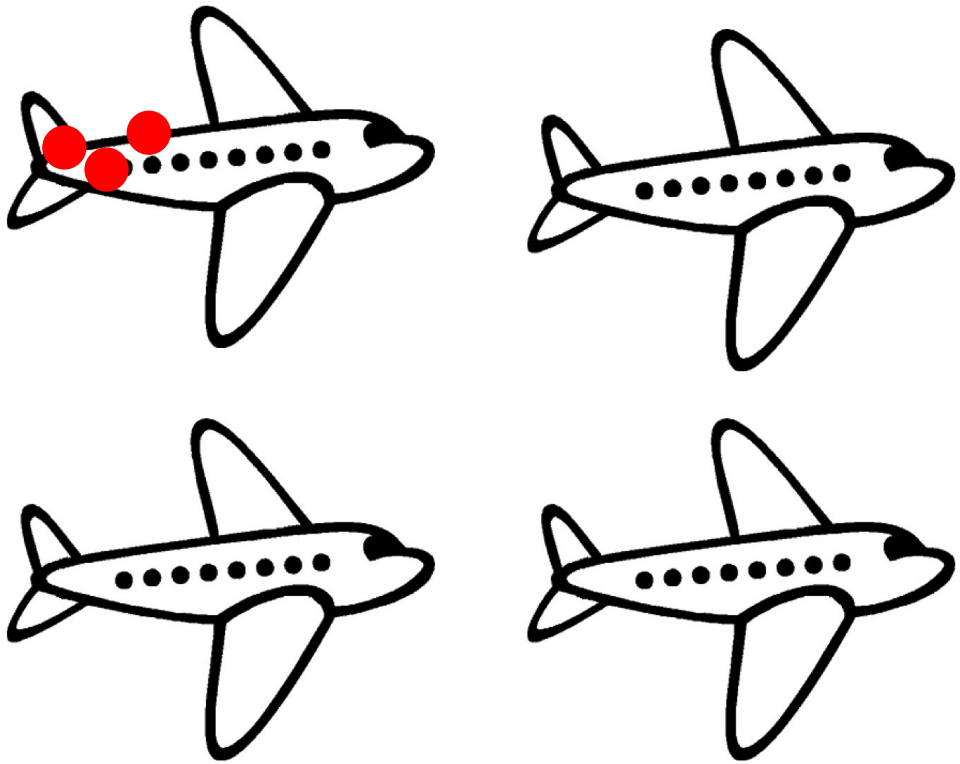
Q: Where would you add armor?

Observed data: Planes that came home with damage



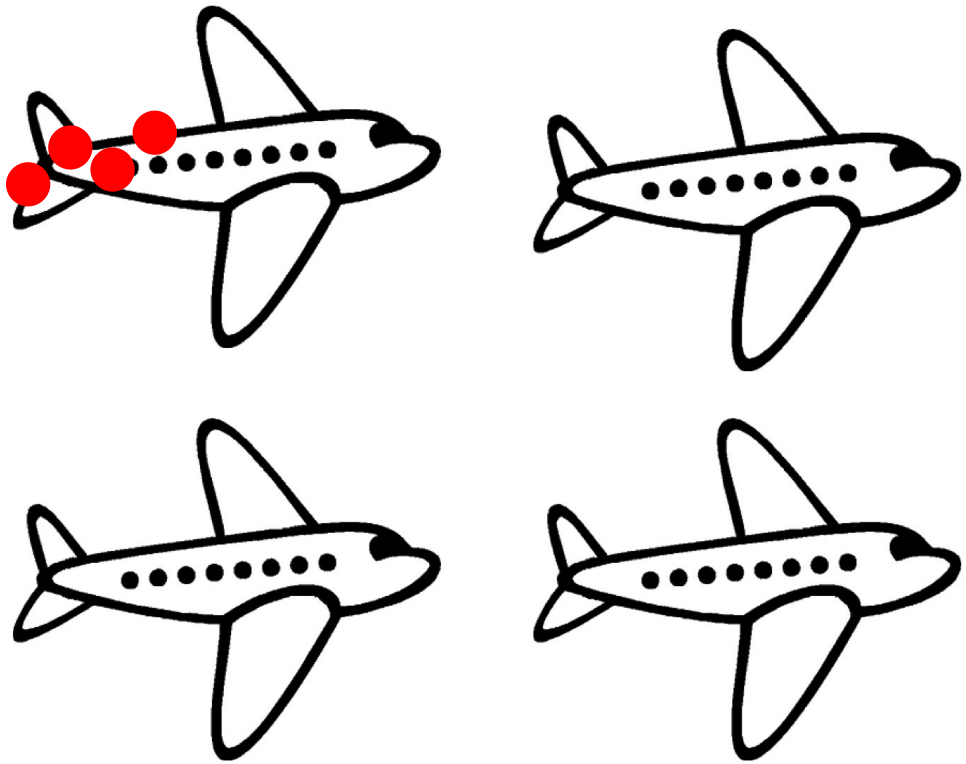
Q: Where would you add armor?

Observed data: Planes that came home with damage



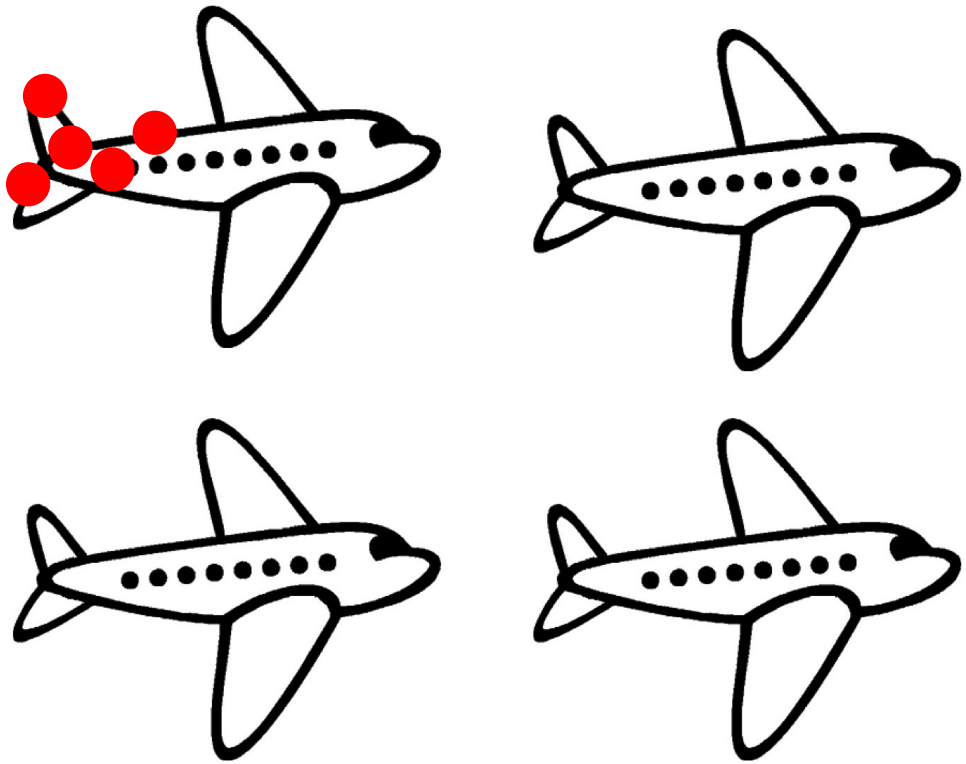
Q: Where would you add armor?

Observed data: Planes that came home with damage



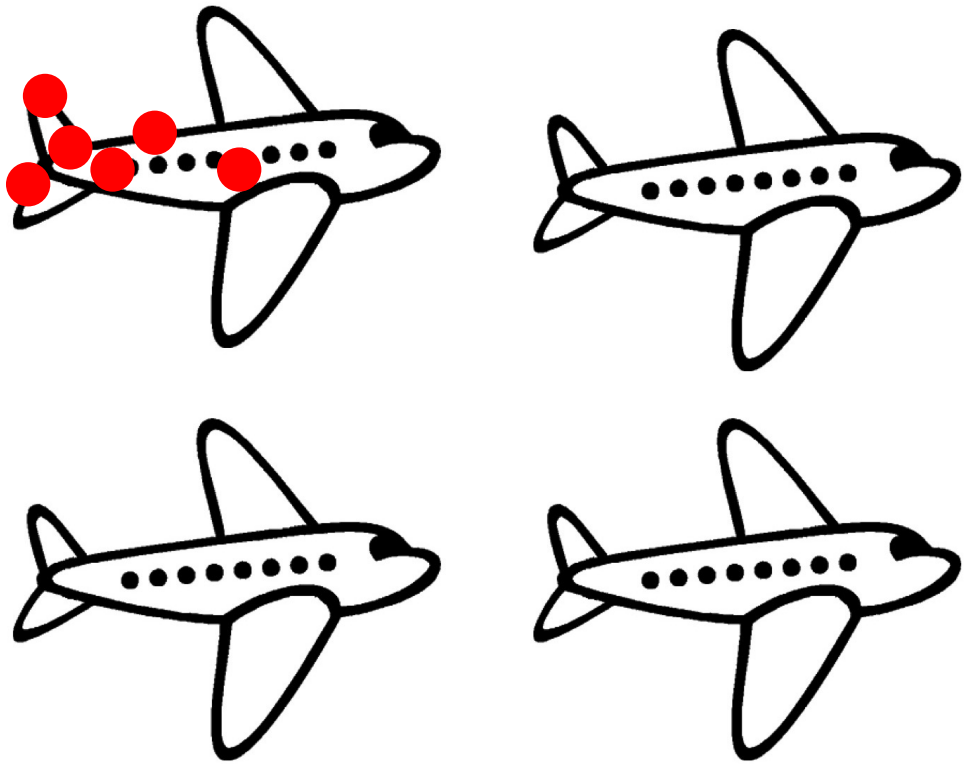
Q: Where would you add armor?

Observed data: Planes that came home with damage



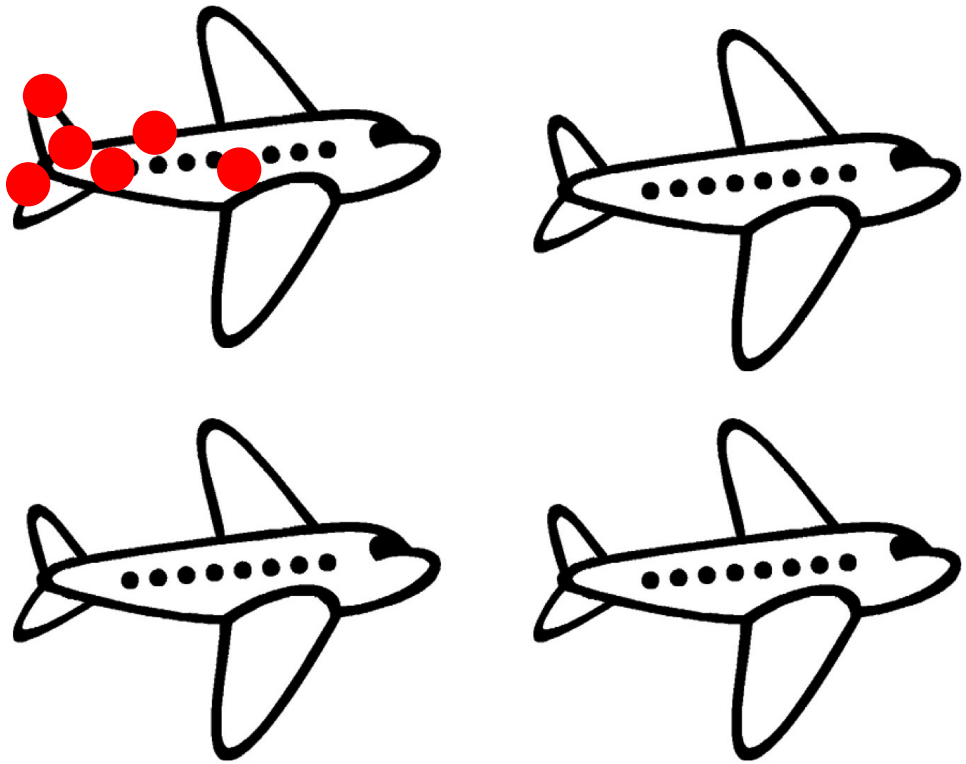
Q: Where would you add armor?

Observed data: Planes that came home with damage



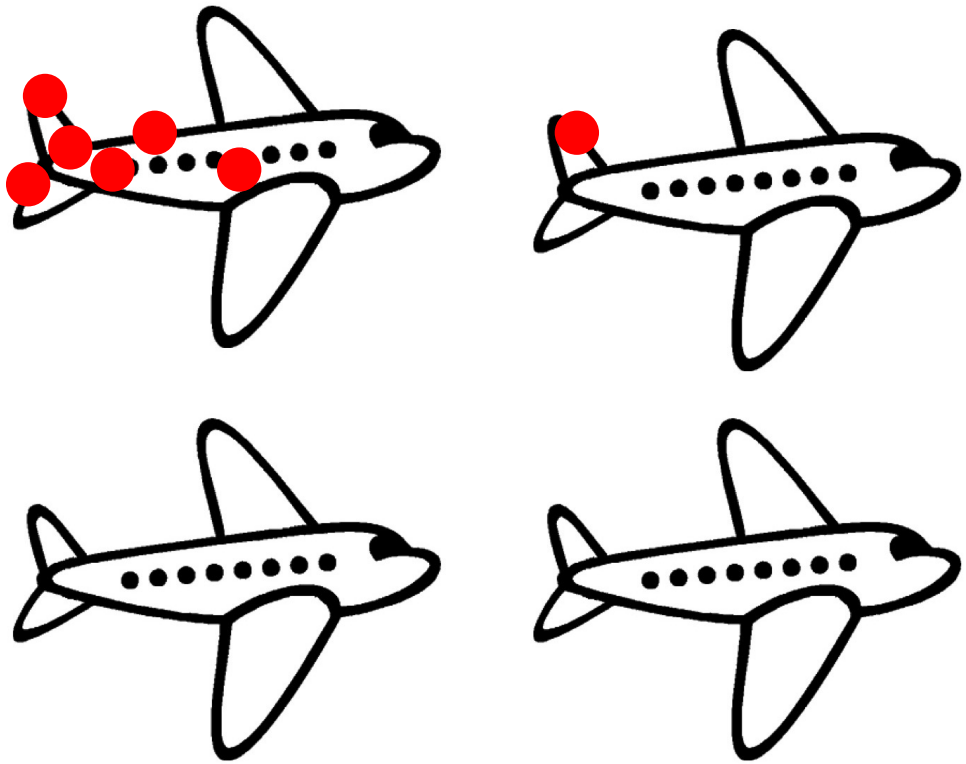
Q: Where would you add armor?

Observed data: Planes that came home with damage



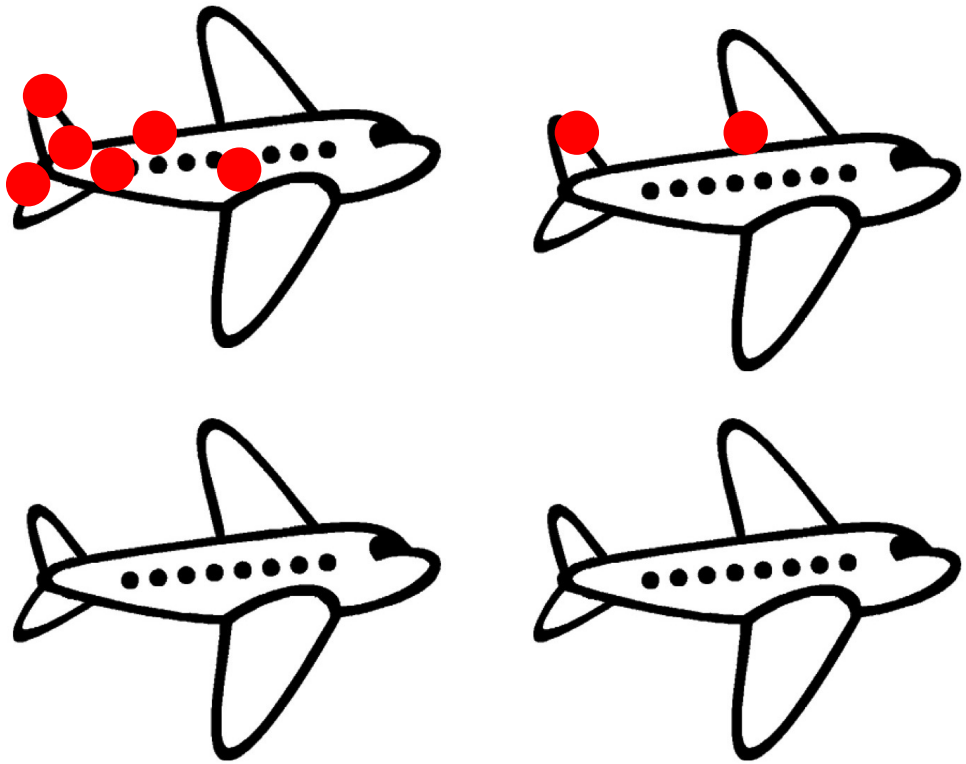
Q: Where would you add armor?

Observed data: Planes that came home with damage



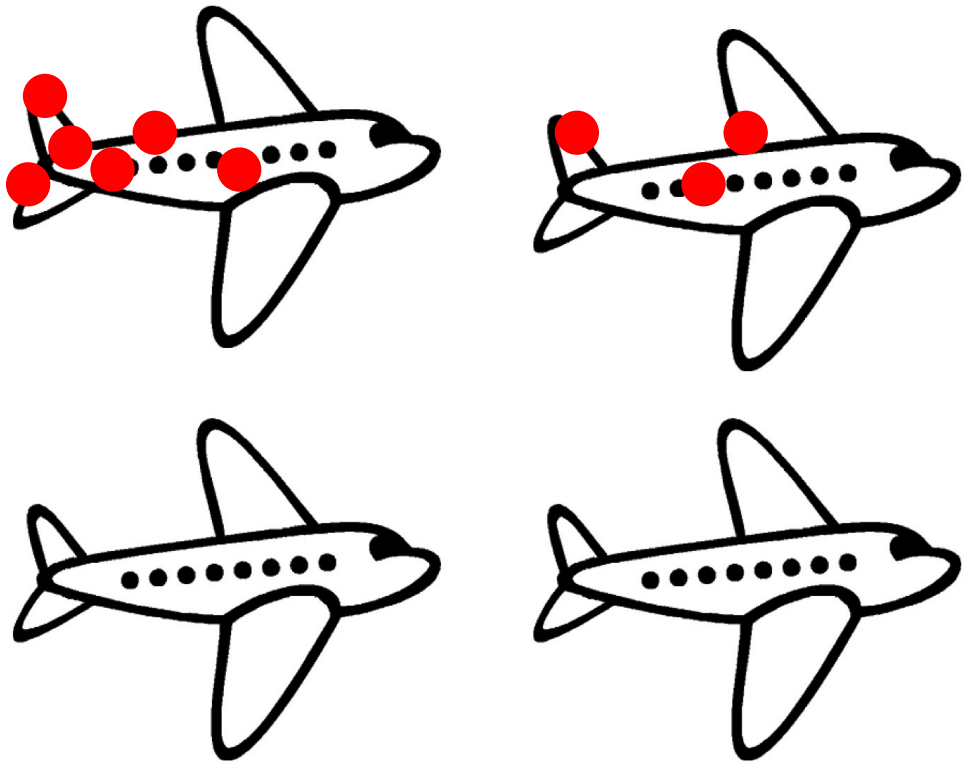
Q: Where would you add armor?

Observed data: Planes that came home with damage



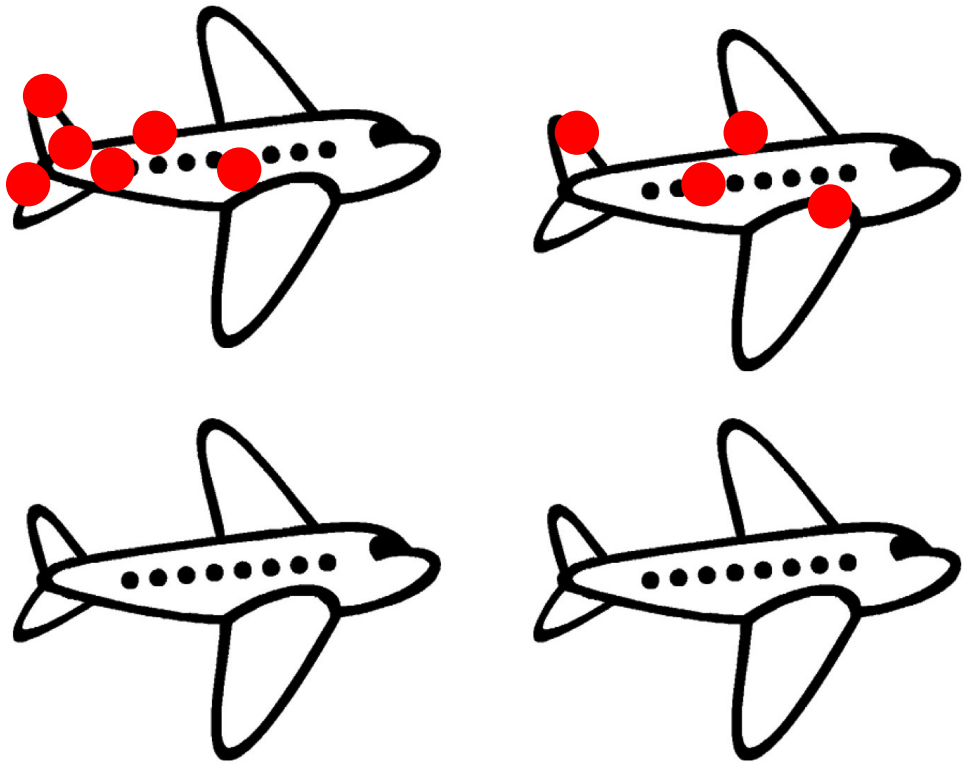
Q: Where would you add armor?

Observed data: Planes that came home with damage



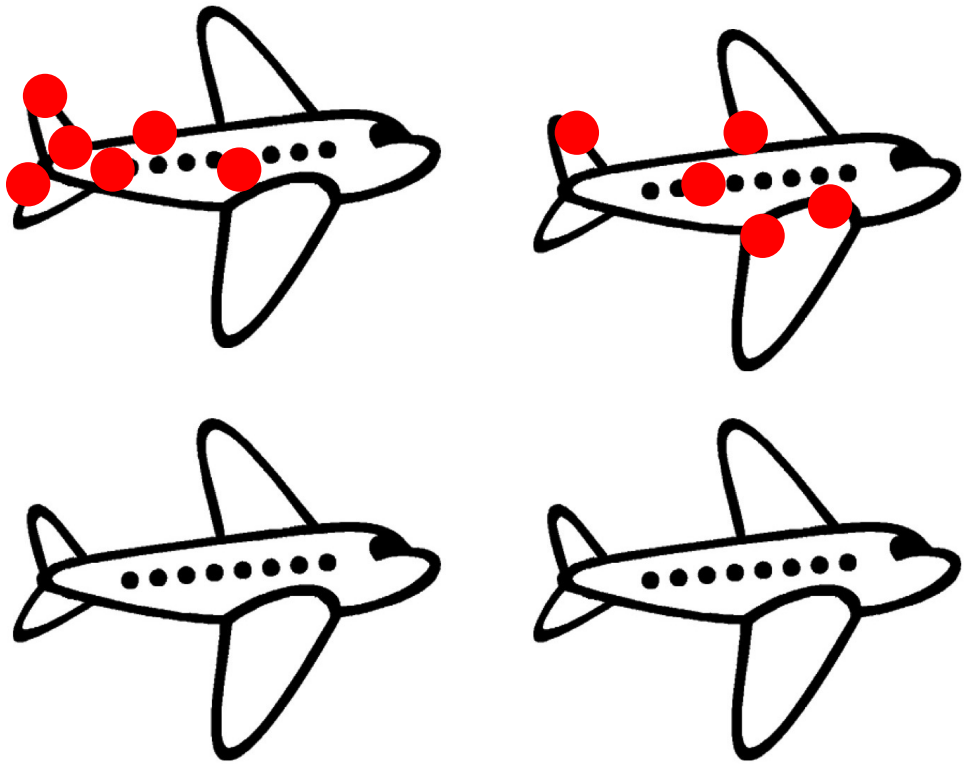
Q: Where would you add armor?

Observed data: Planes that came home with damage



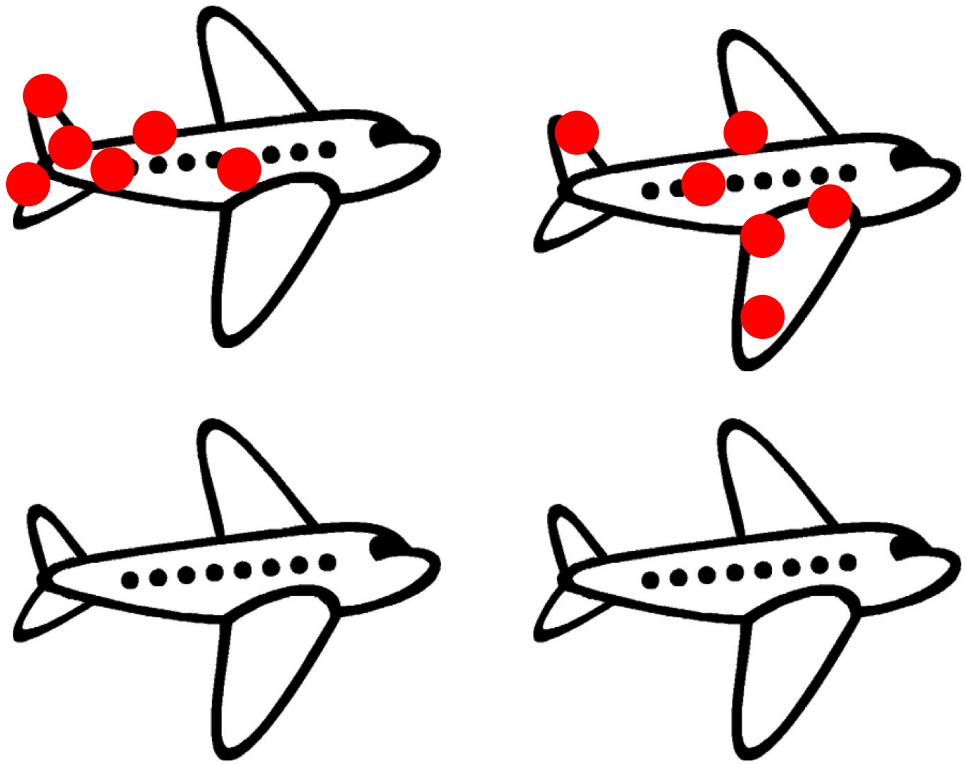
Q: Where would you add armor?

Observed data: Planes that came home with damage



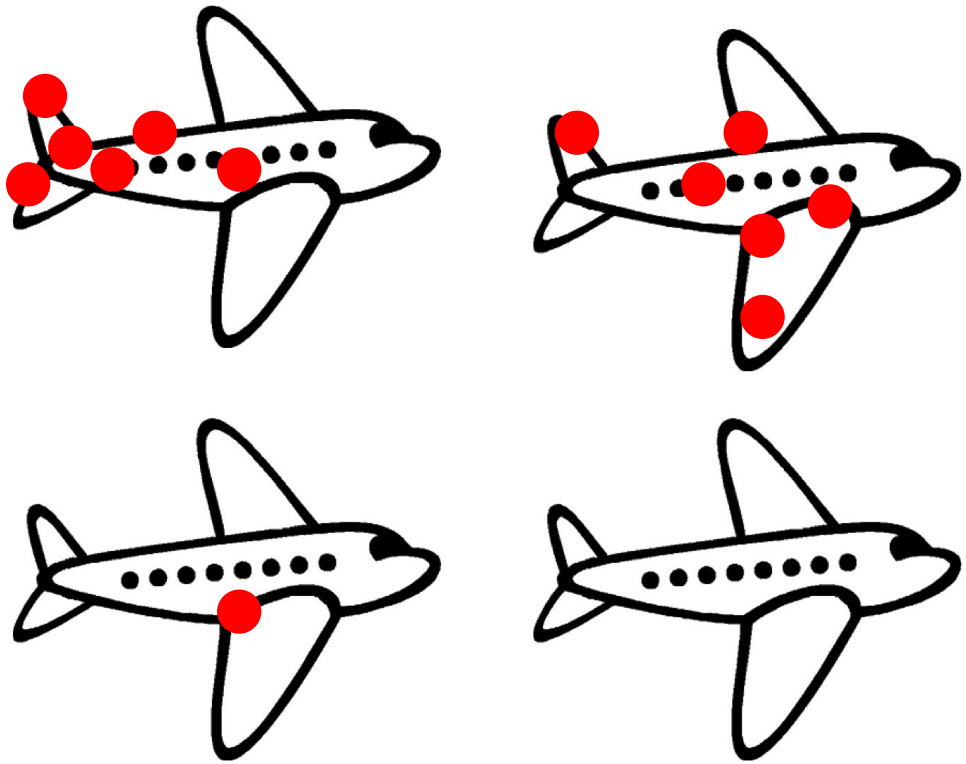
Q: Where would you add armor?

Observed data: Planes that came home with damage



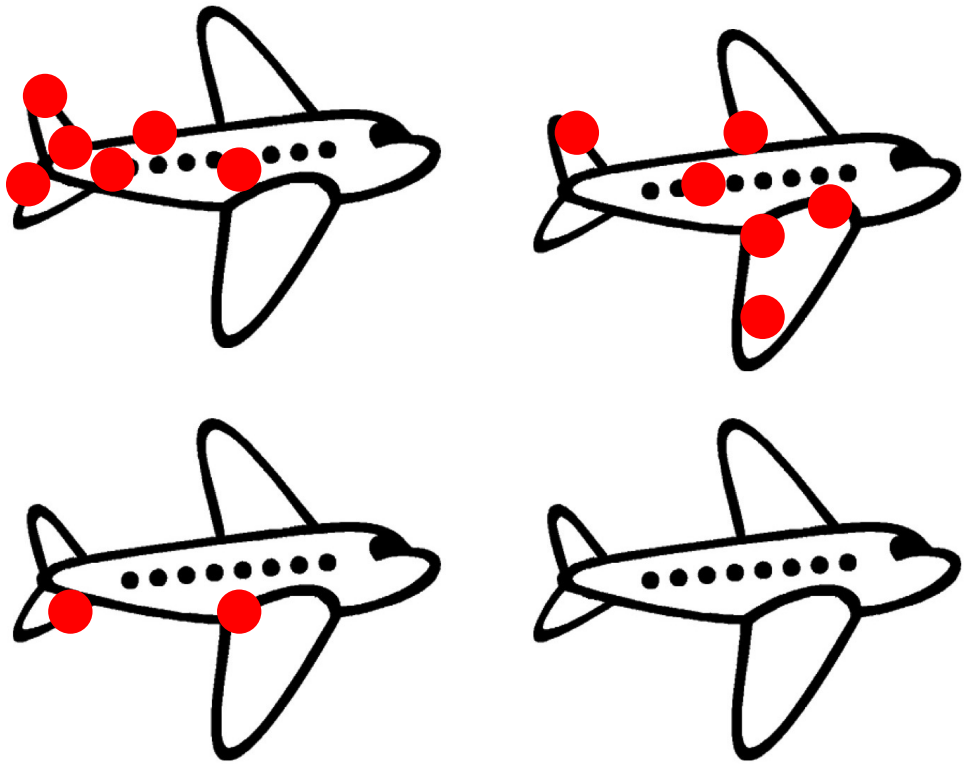
Q: Where would you add armor?

Observed data: Planes that came home with damage



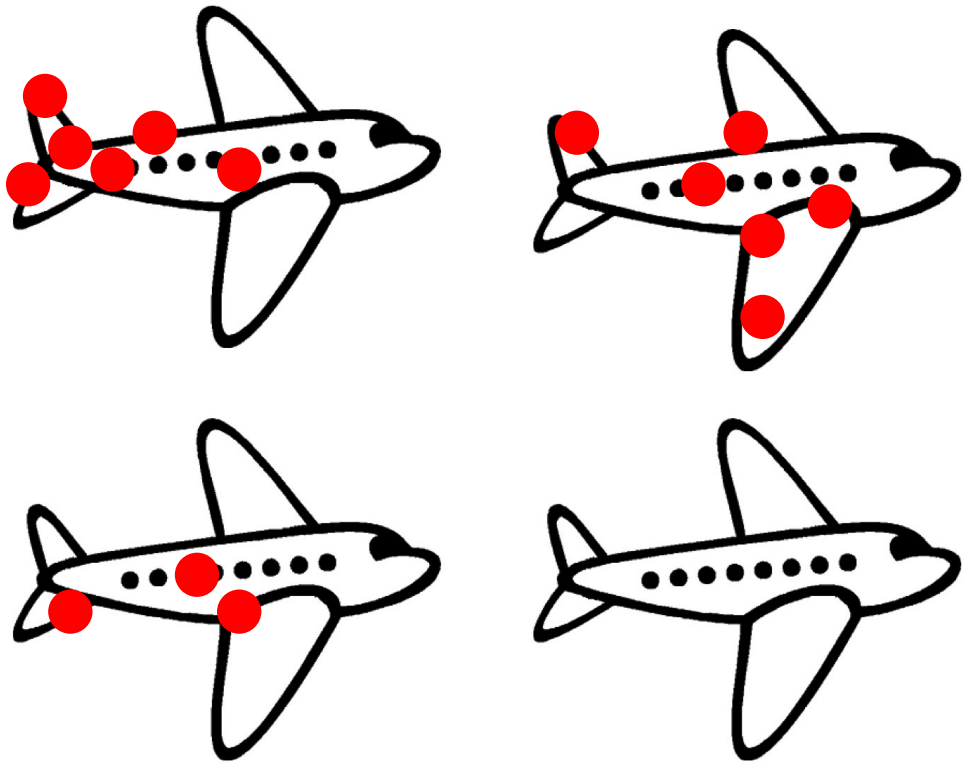
Q: Where would you add armor?

Observed data: Planes that came home with damage



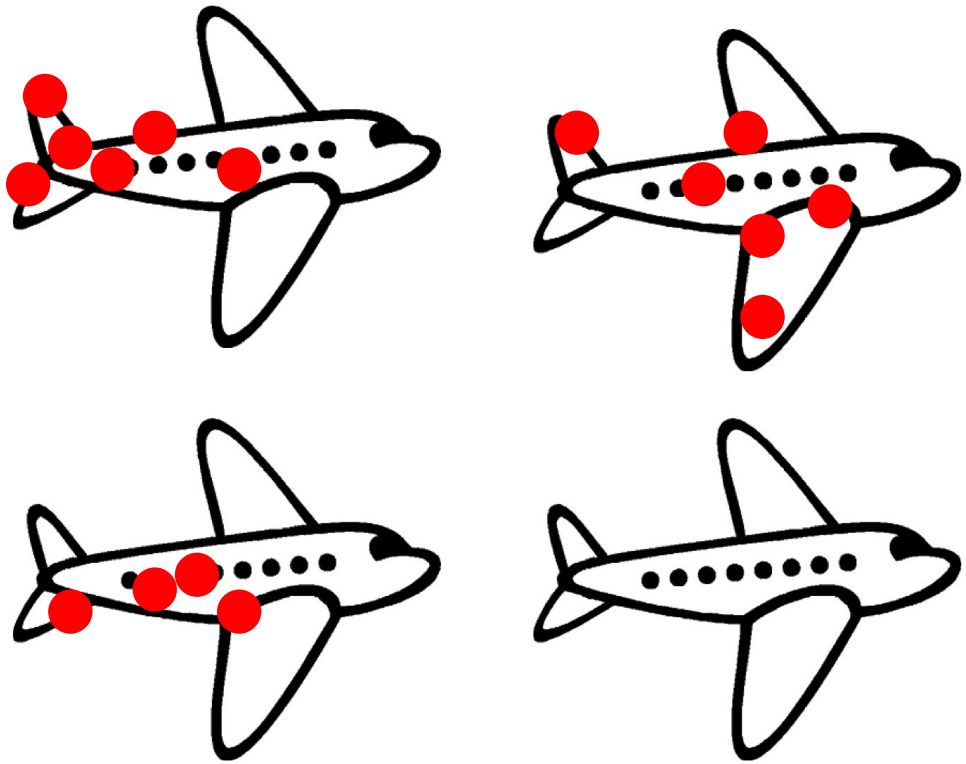
Q: Where would you add armor?

Observed data: Planes that came home with damage



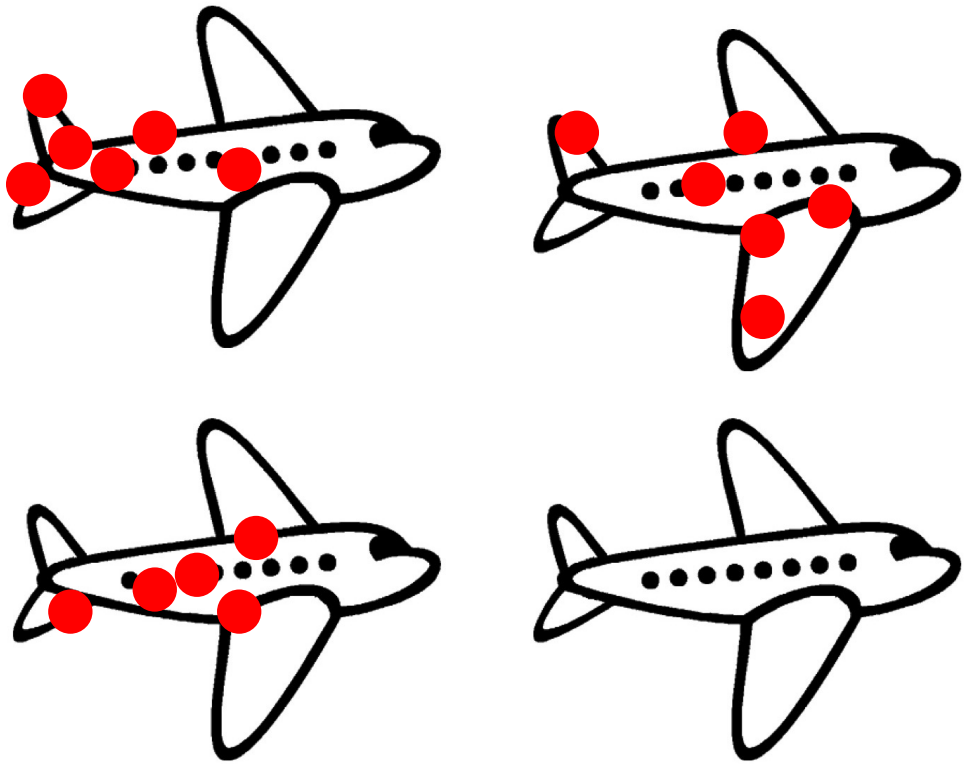
Q: Where would you add armor?

Observed data: Planes that came home with damage



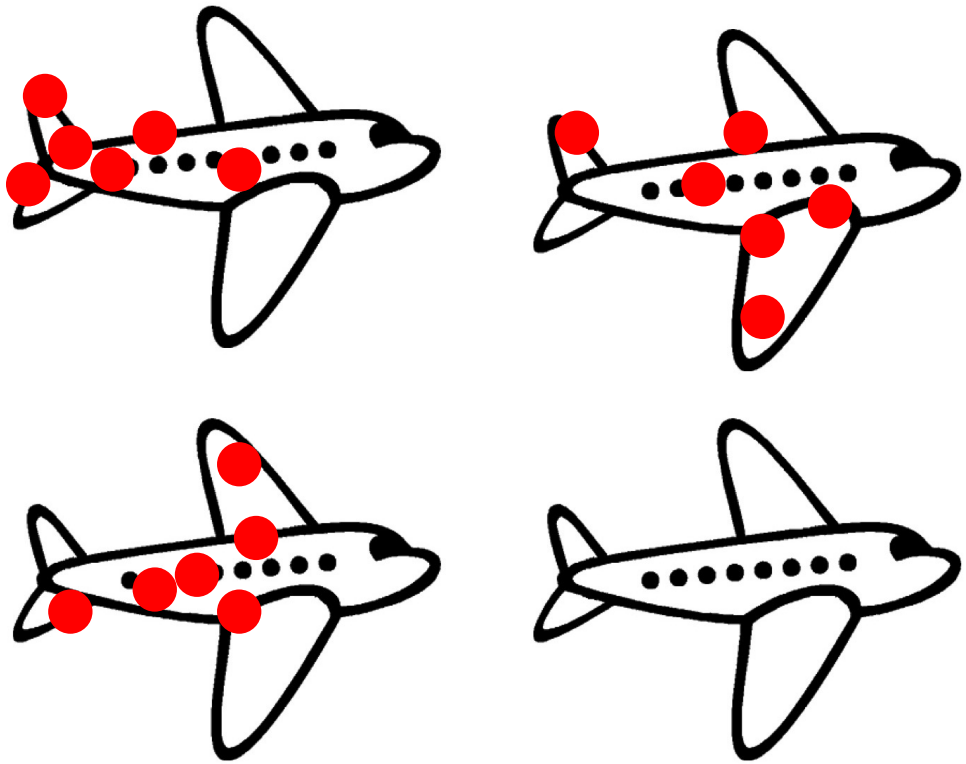
Q: Where would you add armor?

Observed data: Planes that came home with damage



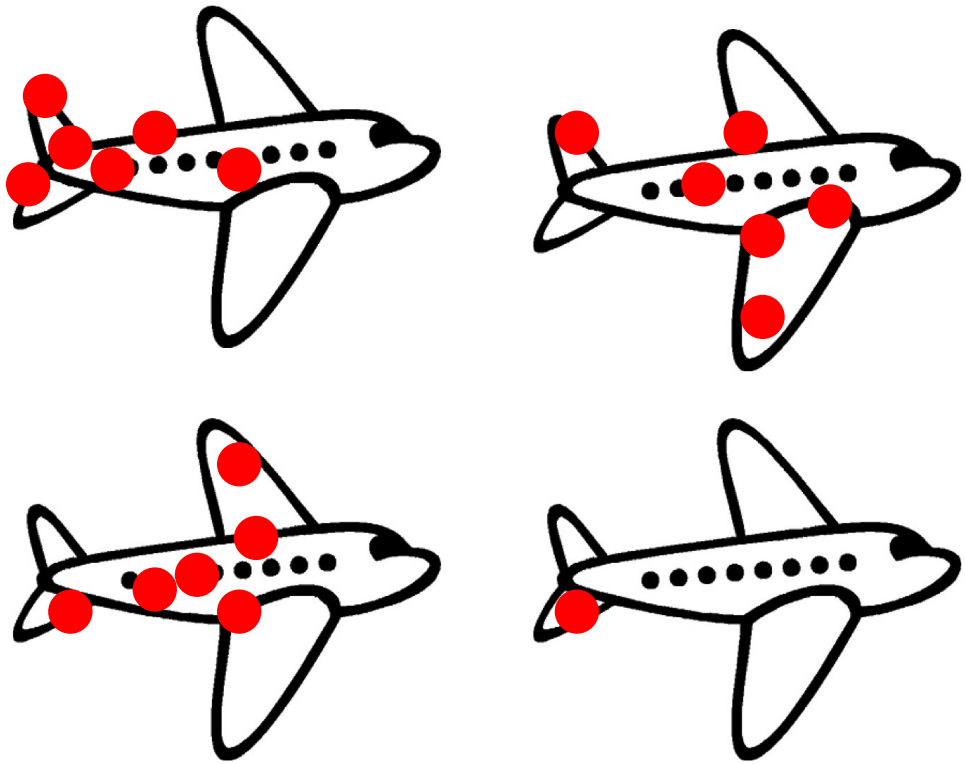
Q: Where would you add armor?

Observed data: Planes that came home with damage



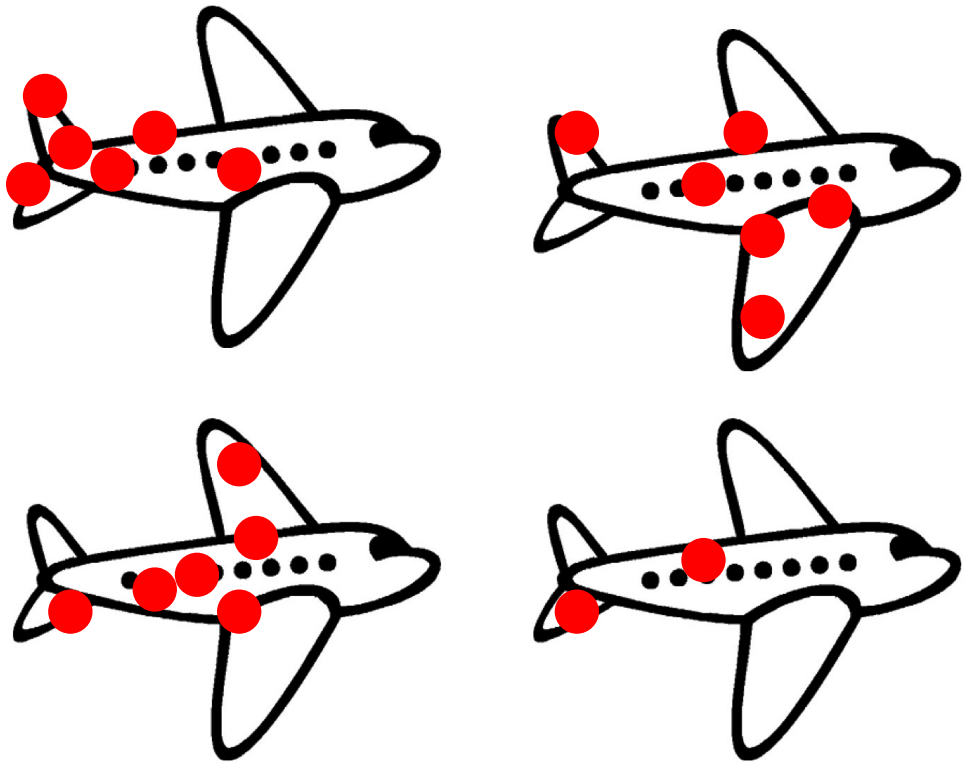
Q: Where would you add armor?

Observed data: Planes that came home with damage



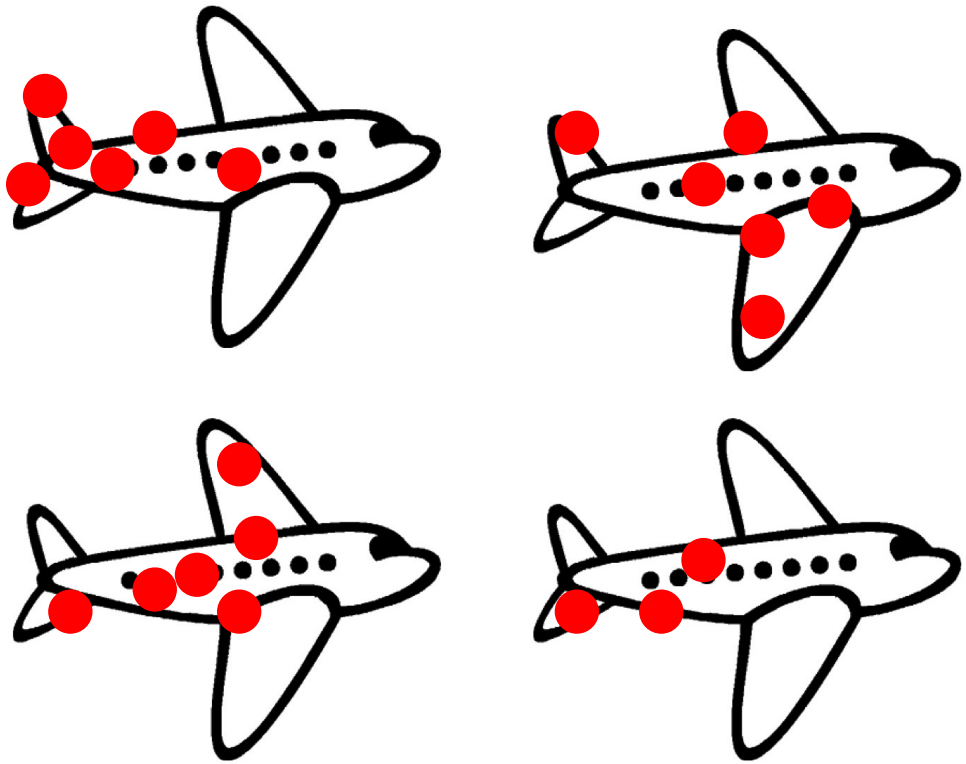
Q: Where would you add armor?

Observed data: Planes that came home with damage



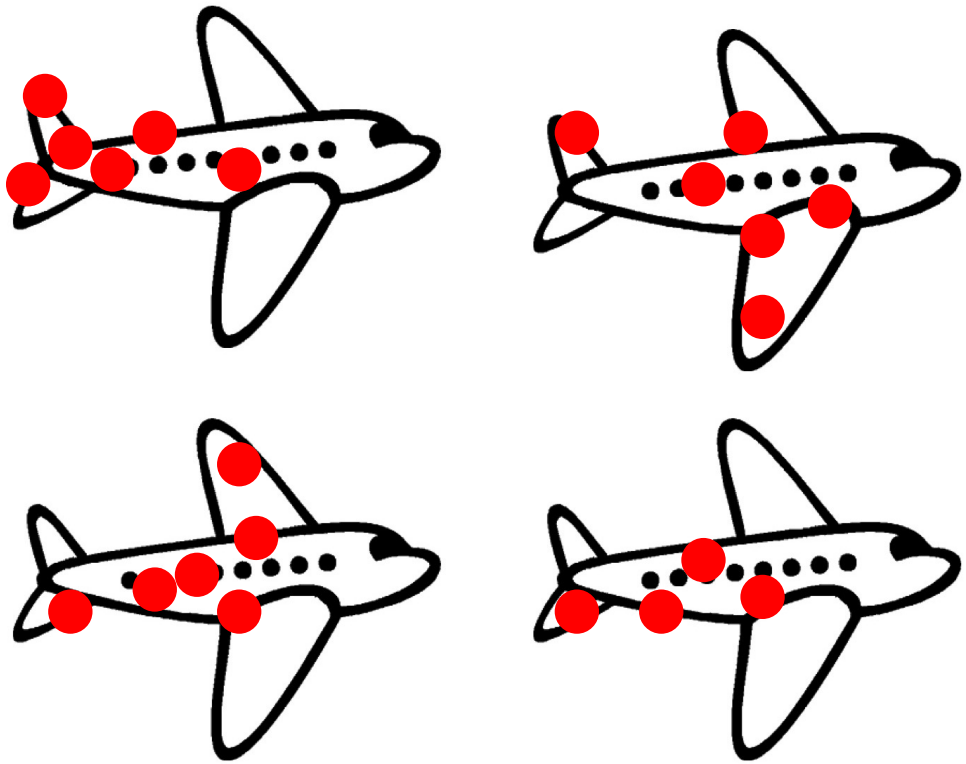
Q: Where would you add armor?

Observed data: Planes that came home with damage



Q: Where would you add armor?

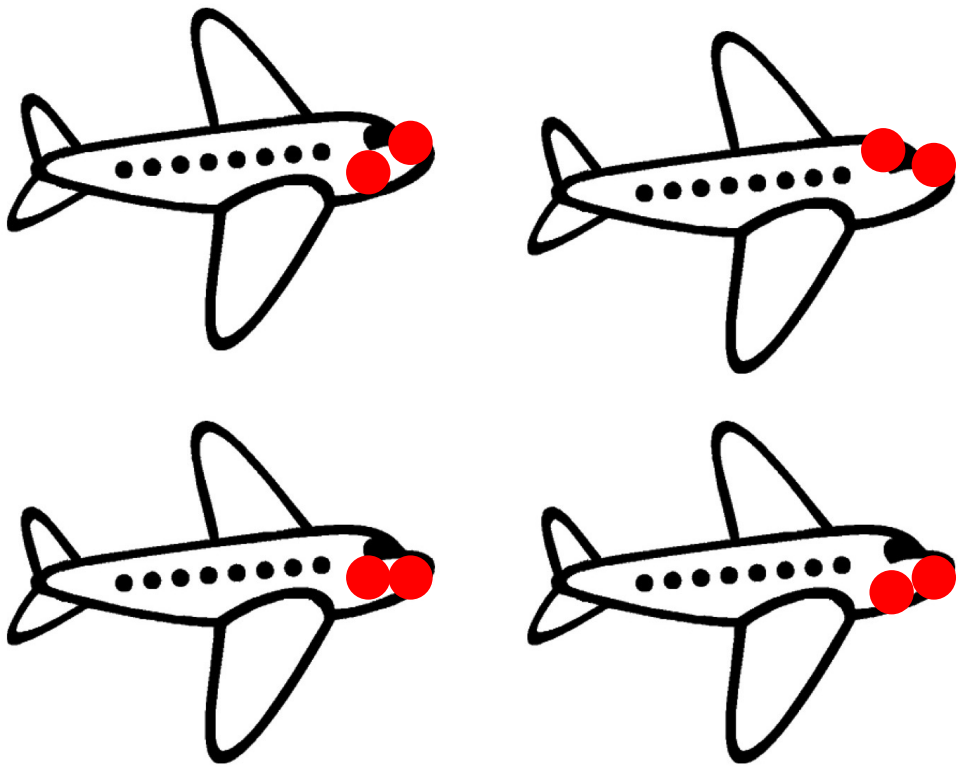
Observed data: Planes that came home with damage



Q: Where would you add armor?

Missing data: Planes that never returned

Missing data: Planes that never returned



We should add armor to the nose!

The missing planes are unlike the observed planes.
Requires argument—no algorithm has the answer.

Missing data in social science

Any NA in your dataset is like a plane that never returned.

It requires careful thought, not a one-size-fits-all procedure.

A concrete example

What is the employment rate among these 6 people?

HS graduate? X	Employed? Y_{True}	Missing? M_Y	Y_{Observed}
1	0	0	0
1	1	0	1
1	0	1	NA
1	1	1	NA
0	0	0	0
0	0	0	0

Truth $\frac{2}{6} = \frac{1}{3}$

A concrete example

What is the employment rate among these 6 people?

HS graduate? X	Employed? Y_{True}	Missing? M_Y	Y_{Observed}
1	0	0	0
1	1	0	1
1	0	1	NA
1	1	1	NA
0	0	0	0
0	0	0	0

Truth $\frac{2}{6} = \frac{1}{3}$

Estimate dropping NA $\frac{1}{4}$

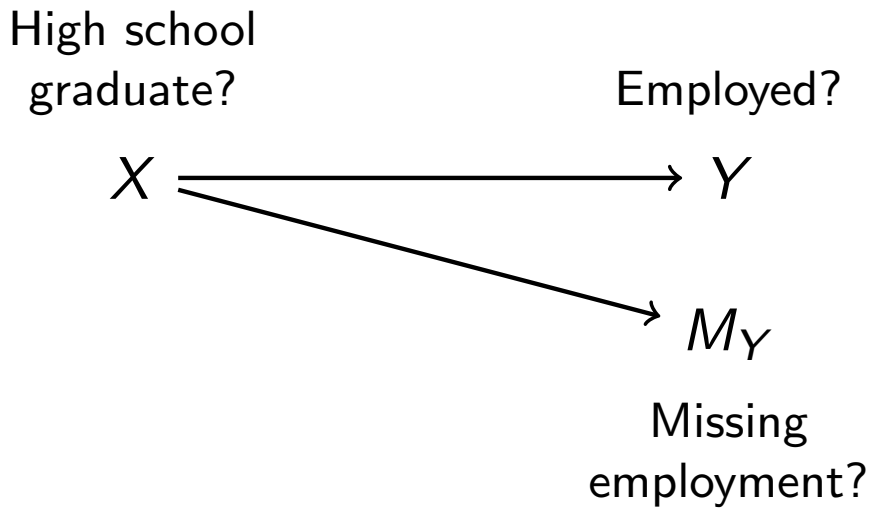
Missing completely at random and listwise deletion

If $\{M_X, M_Y\} \perp\!\!\!\perp \{X, Y\}$, then you can simply drop missing cases.

- This approach is called listwise deletion
- This assumption is called missing completely at random
- Assumption did not hold in this case

Missing at random: Listwise deletion within X

What if we assume $Y \perp\!\!\!\perp M_Y \mid X$?



Then we can estimate $E(Y)$ by adjusting for X.

Missing at random: Listwise deletion within X

$$
\begin{aligned}
E(Y) &= \mathrm{E}(Y \mid X = \text{HS graduate}) \times \mathrm{P}(X = \text{HS graduate}) \\
&\quad + \mathrm{E}(Y \mid X \neq \text{HS graduate}) \times \mathrm{P}(X \neq \text{HS graduate}) \\
&= \mathrm{E}(Y \mid X = \text{HS graduate}, M_Y = 0) \times \mathrm{P}(X = \text{HS graduate}) \\
&\quad + \mathrm{E}(Y \mid X \neq \text{HS graduate}, M_Y = 0) \times \mathrm{P}(X \neq \text{HS graduate}) \\
&= \frac{1}{2} \times \frac{4}{6} + 0 \times \frac{2}{6} \\
&= \frac{1}{3}
\end{aligned}
$$

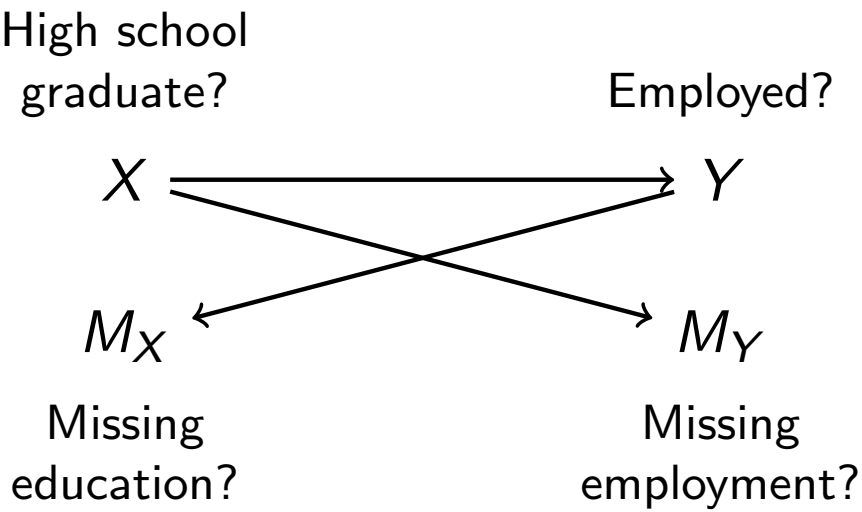
Missing at random: Imputation

HS graduate? X	Employed? Y_{True}	Missing? M_Y	Y_{Observed}	Y_{Imputed}
1	0	0	0	0
1	1	0	1	1
1	0	1	NA	0.5
1	1	1	NA	0.5
0	0	0	0	0
0	0	0	0	0

Exactly like causal inference:
when Y is not observed, impute its value under assumptions.

Generalizing to multivariate missingness

We hope for missing at random $M_X \perp\!\!\!\perp X \mid Y$ and $M_Y \perp\!\!\!\perp Y \mid X$



More generally, we hope $\vec{M} \perp\!\!\!\perp \text{Data}_{\text{Missing}} \mid \text{Data}_{\text{Observed}}$

Standard errors by the bootstrap

Use the estimator function $s()$

- Point estimate $\hat{\tau} = s(\text{data})$
- Bootstrap confidence interval by quantiles of $s(\text{data}^*)$

Within $s()$

- impute missing values
- carry out analysis
- return estimate

Standard errors by math

Approach is called Rubin's Rules

Create m imputed datasets.

Variance estimator involves variance

- **within** each imputation and
- **between** imputations.

Standard errors by math

Approach is called Rubin's Rules

Create m imputed datasets.

Variance estimator involves variance

- **within** each imputation and
- **between** imputations.

$$\hat{\mathrm{V}}(\hat{\theta}) = \underbrace{\frac{1}{m}\sum_{i=1}^{m}\hat{\mathrm{V}}(\hat{\theta}_i)}_{\substack{\text{Mean} \\ \textbf{within-imputation} \text{ variance}}} + \underbrace{\left(1+\frac{1}{m}\right)}_{\substack{\text{Inflation} \\ \text{factor}}} \underbrace{\left(\frac{1}{m-1}\sum_{i=1}^{m}\left[\hat{\theta}_i - \bar{\hat{\theta}}\right]^2\right)}_{\textbf{Between-imputation} \text{ variance}}$$

What happens in practice

- Many variables in $\vec{X}$
- Swiss cheese missingness
- Very hard to make assumptions credible
- Reviewers prefer multiple imputation

My advice for dropping cases

Track dropped cases in a logical order

- drop outside target population (non-controversial)
- drop missing confounder values (mostly ok)
- drop missing treatment values (analogous to ATT)
- drop missing outcomes (very dubious)

My own workflow

No guarantees here; my workflow is uncommon.

I use the `Amelia` package because it is fast.

- Write an `estimator()` function
  - goes from data to an estimate
  - internally calls `amelia()` to impute
  - set `boot.type = FALSE` within `amelia()` (bootstrap will be handled separately)
- Use `estimator(data)` to make a point estimate
- Bootstrap to make a confidence interval by quantiles of `estimator(data*)`

Here is an example from my code.

Learning goals for today

By the end of class, you will be able to

- make assumptions about missing data
  - missing completely at random
  - missing at random
- connect these assumptions to causal inference
- use two strategies for missing data
  - listwise deletion
  - multiple imputation