Data Wrangling and Data Analysis
Missing Data and Imputation

Daniel Oberski

Ayoub Bagheri

Erik-Jan van Kesteren

Anastasia Giachanou

Department of Methodology & Statistics

Utrecht University

This week

- What is missing data
- Sources of missingness
- Missing data mechanisms
- Missingness patterns

- Ad-hoc solutions to missing data in databases
- Multiple imputation and Sensitivity

Assignments this week

- Monday: Exercise on missing data in python/R, understanding and visualising missingness.
- Tuesday: Correcting for missingness in databases in R
- Wednesday: Multiple choice test on missingness mechanisms and solutions
- Thursday: either (a) resit for the test, or (b) assignment on multiple imputation in R

Strategies to deal with missing data in the data wrangling process

Data collection → Data wrangling → Data Analysis

Prevention

(of course, this is what we want)

Imputation

- Ad-hoc methods
- Multiple imputation

Adjustment

- Weighting methods
- Likelihood methods
- EM-algorithm

Imputation

Replacing missing values with *guessed* values

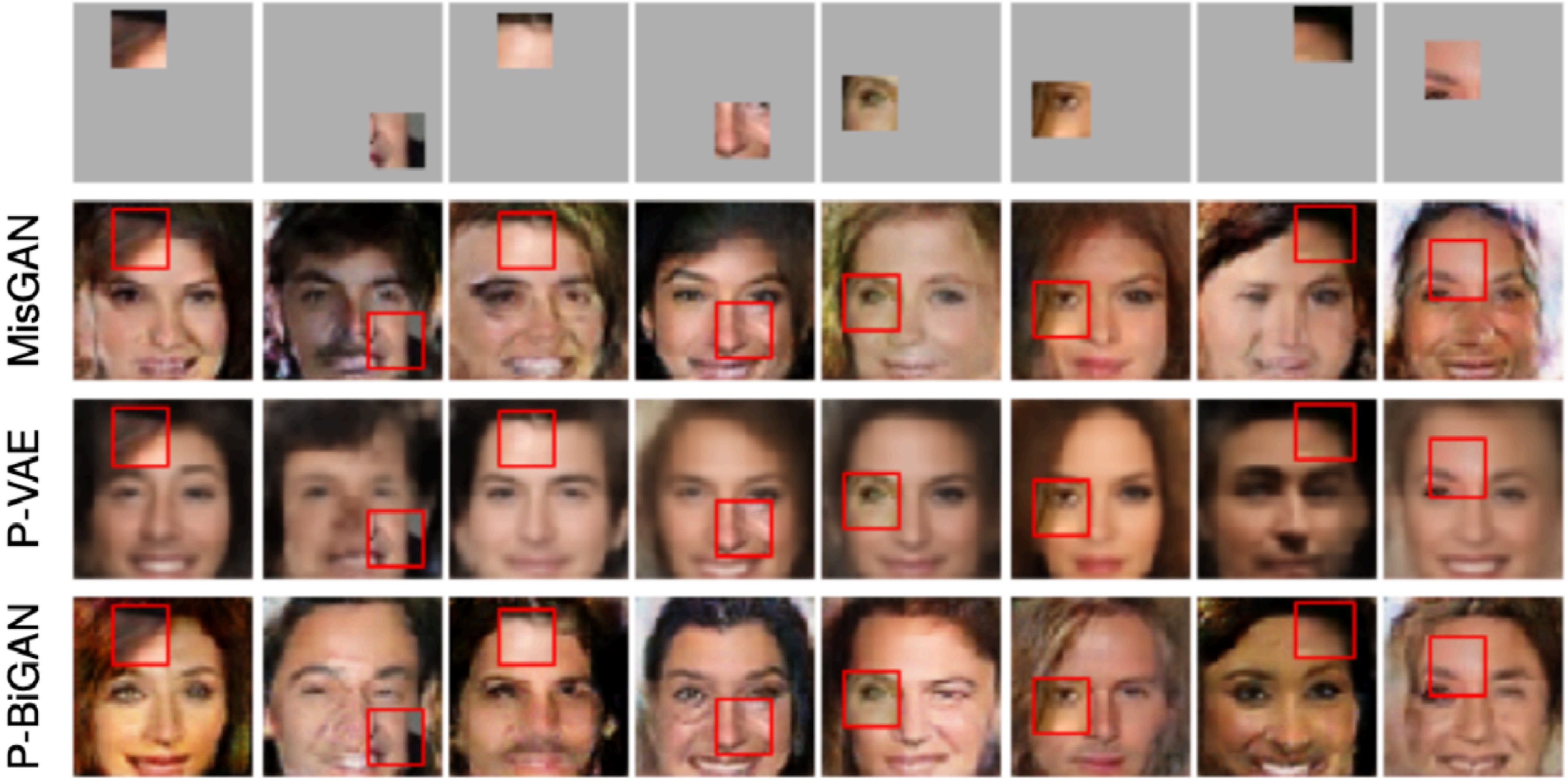
```
##    age weight
## 1 13   42
## 4 14   NA
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

```
##    age weight
## 1 13   42
## 4 14   47
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

Imputation

- Yields “complete” data
- Statistics are now defined
- Convenient!
- But… we just made up some data!
- How close does this get us to the target…?

Imputation can look quite OK with the right model



Cheng-Xian Li & Marlin (2020), ICML

Deductive imputation – always a good idea

- If we know height and weight, we can calculate BMI
- If someone is unemployed, we know that person has zero income out of labour

Inverse also holds: If we can prove that an observed value must be wrong, we must correct it (if we can) or make it missing

Example: database may contain a 14 yr old female with 3 kids, married for 20 years and working as a manager at a public school

- This may be a data entry error: her age could be 41

Listwise deletion

- Also known as Complete Case Analysis (CCA)

Listwise deletion

```
##   age weight
## 1 13  42
## 4 14  NA
## 6 18  61
## 5 23  70
## 3 24  73
## 7 25  68
## 2 40  80
```

→

```
##   age weight
## 1 13  42
## 6 18  61
## 5 23  70
## 3 24  73
## 7 25  68
## 2 40  80
```

Listwise deletion

- Advantages
    - Simple (default in most software)
    - Unbiased under MCAR
- Disadvantages
    - Wasteful
    - Large standard errors
    - Biased under MAR, even for simple statistics like the mean
    - Inconsistencies in reporting

Mean imputation

- Replace the missing values by the mean of the observed data
- Advantages
  - Simple
  - Unbiased for the mean, under MCAR

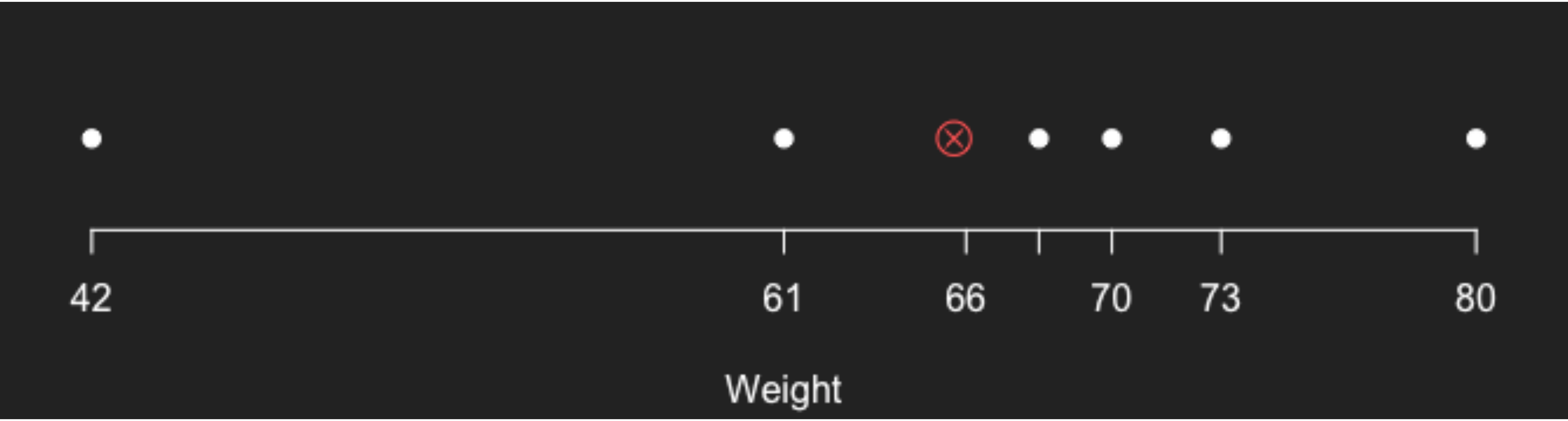
Mean imputation

```
##    age weight
## 1 13   42
## 4 14   NA
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

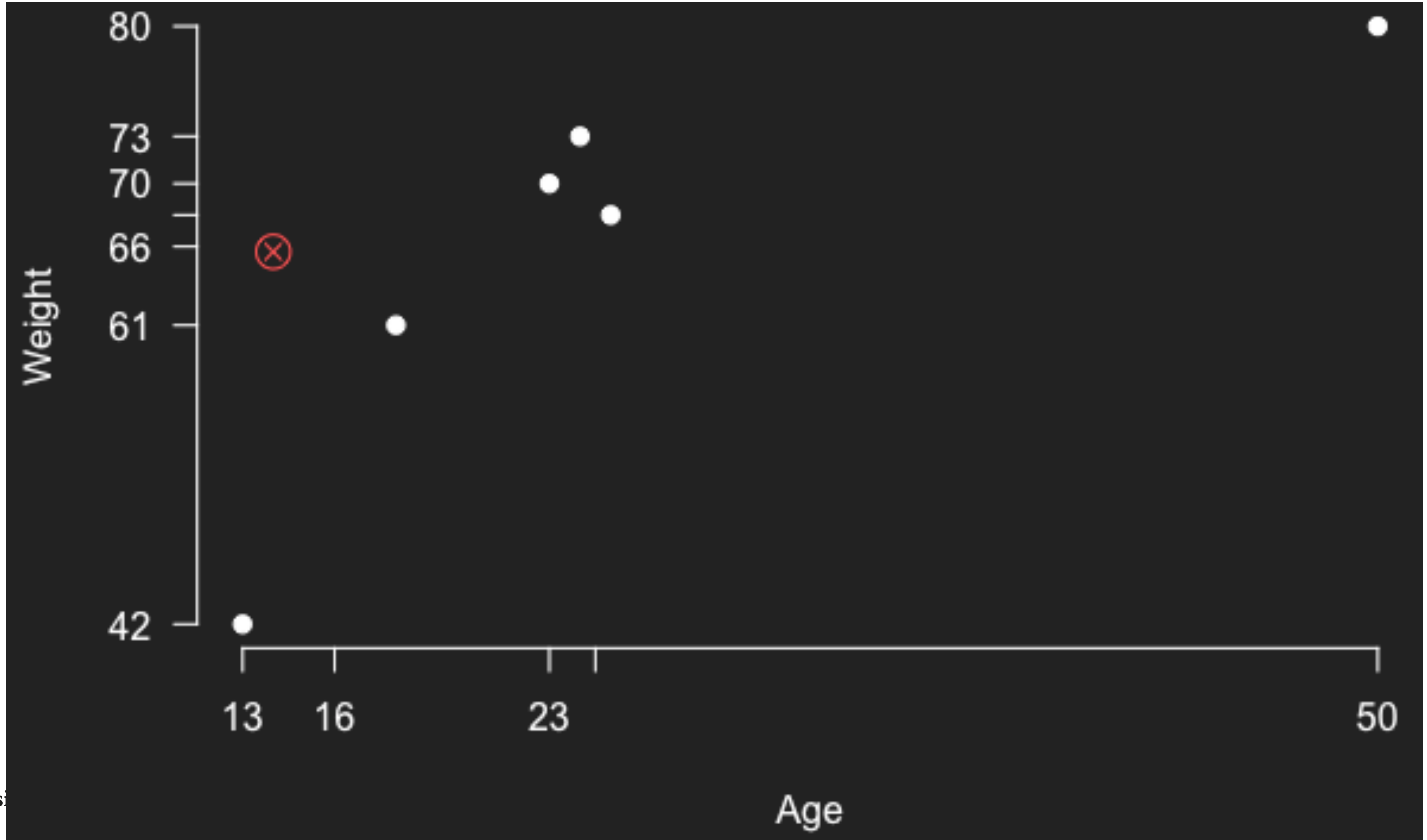
→

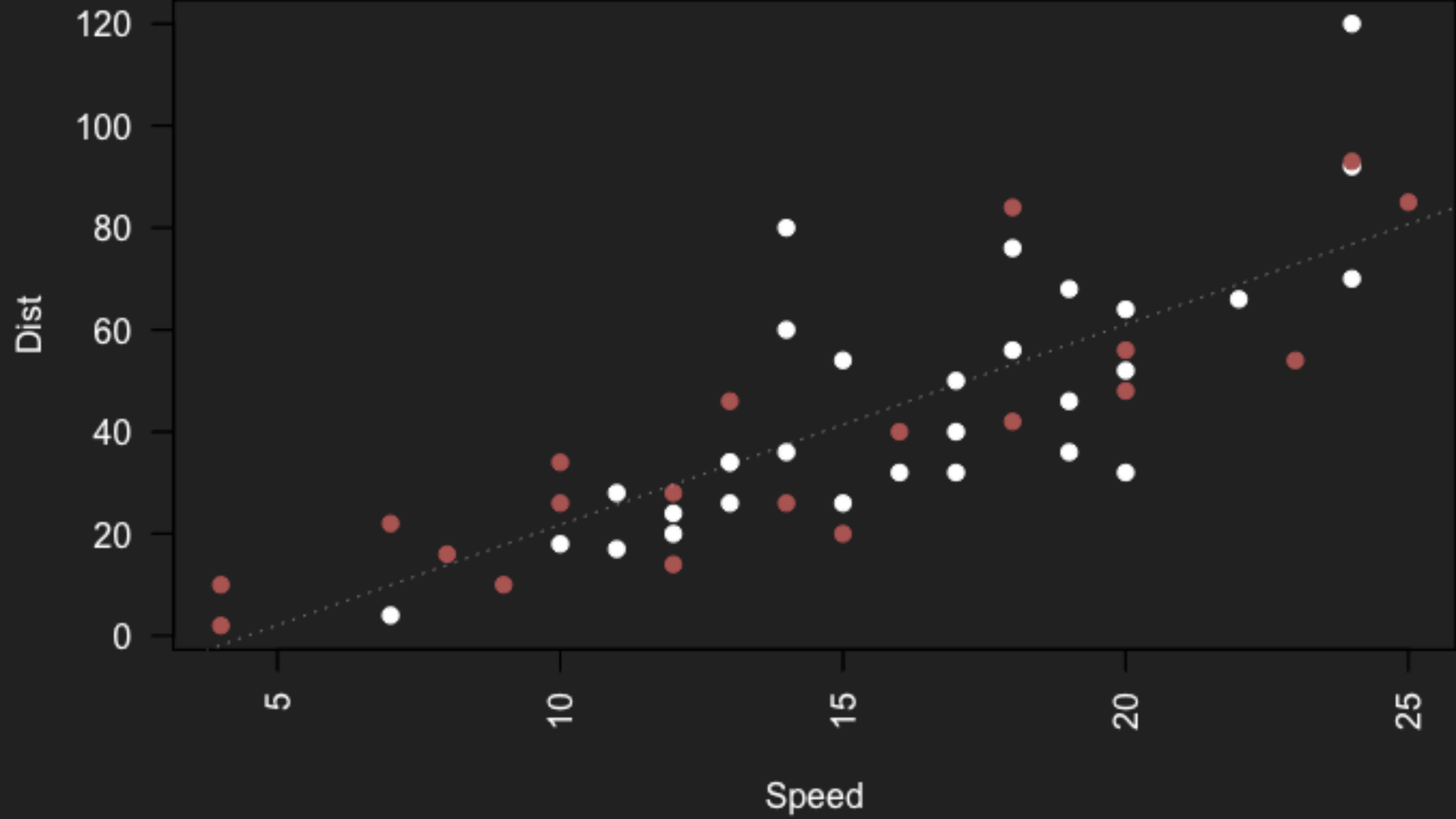
```
##    age weight
## 1 13   42
## 4 14   65.667
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

Mean imputation: **uni**variate perspective

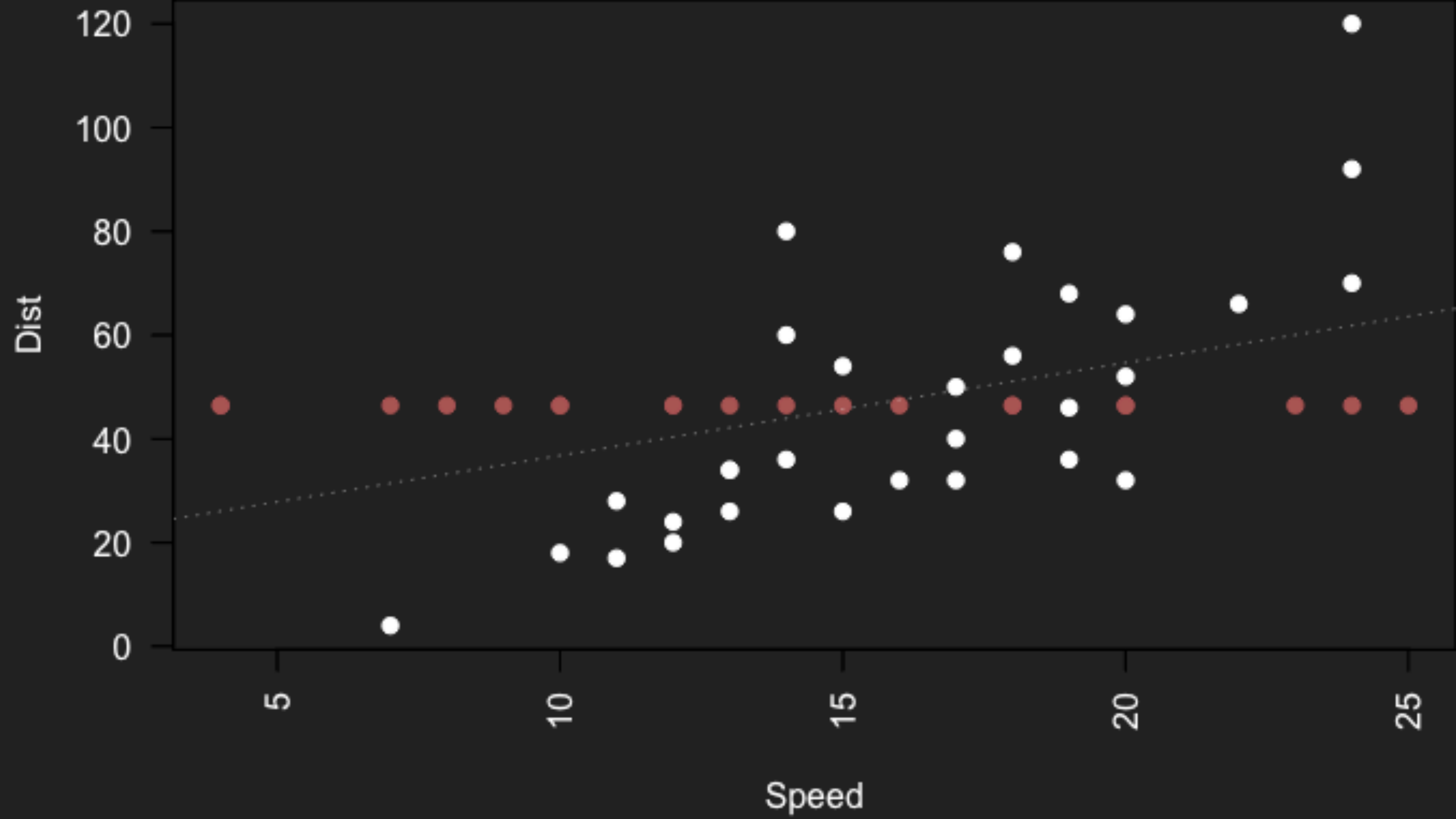


Mean imputation: **bi**variate perspective





Dist
Speed
0
20
40
60
80
100
120
5
10
15
20
25



120
100
80
60
40
20
0
Dist
5
10
15
20
25
Speed

Mean imputation

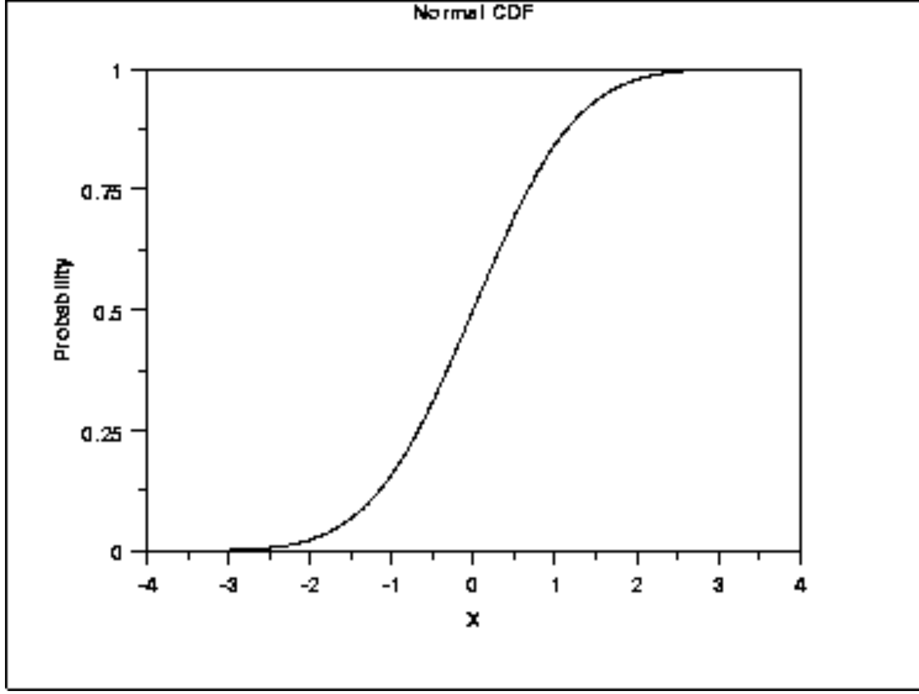
- Disadvantages
  - Disturbs the distribution
  - Underestimates the variance
  - Biases correlations to zero
  - Biased under MAR

- AVOID (unless you know what you are doing)

"Uncertainty"

- A number we're trying to estimate has a "standard error"
- This is the hypothetical standard deviation I would get for my number (say, a mean or regression coefficient) if I repeated the random sampling procedure lots of times and calculated the number each time
- From the "standard error", you can estimate a "confidence interval", often by just going **2×standard error** above and below the number you got. For example if you got mean=66 and se=2, then CI: 66±4
- You can also get a "p-value", often by taking the number and **dividing it by the standard error.** Then running that result through a specific function, namely the standard normal CDF

"Uncertainty"



So:

- **Standard error** se comes from variance and sample size
- p-value comes from estimate/se
- CI comes from mean $\pm$ 2$\times$se

- There's also a different way of quantifying uncertainty (Bayesian)

Regression imputation

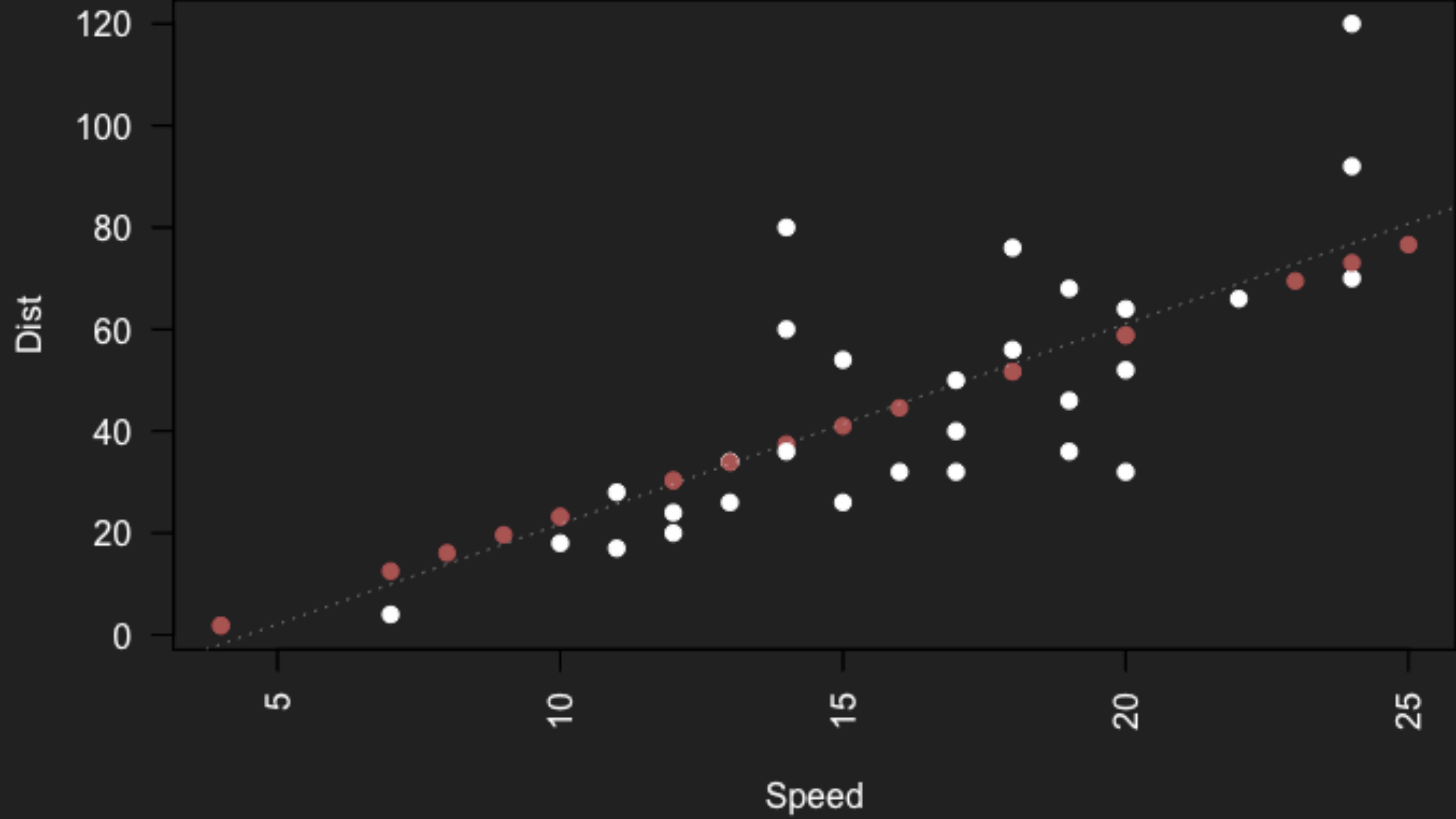
- Also known as prediction
- Fit model for weight under listwise deletion: the **imputation model**
- Predict weight for records with missing weight
- Replace missing values by prediction

Regression imputation

```
##    age weight
## 1 13   42
## 4 14   NA
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

→

```
##    age weight
## 1 13   42
## 4 14   53.45
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

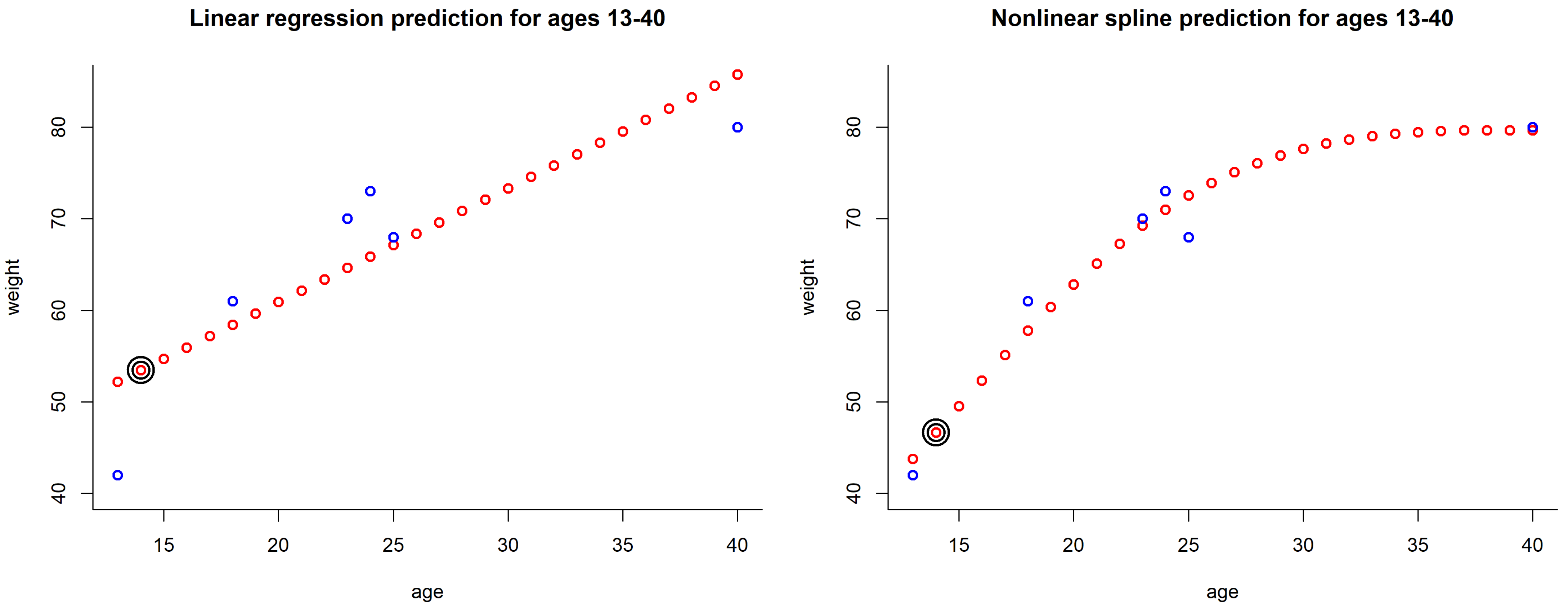


0
20
40
60
80
100
120
Dist
5
10
15
20
25
Speed

Regression imputation

- Advantages
    - Unbiased estimates of regression coefficients (under MAR)
    - Good approximation to the (unknown) true data if explained variance is high
    - The better your prediction, the better your approximation

The better your prediction, the better your approximation



Utrecht University

Regression imputation: spline

```
##    age weight
## 1 13   42
## 4 14   NA
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

→

```
##    age weight
## 1 13   42
## 4 14   46.65
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

Regression imputation

- Disadvantages:
  - Artificially increases correlations
  - Systematically underestimates the uncertainty
  - p-values too optimistic, confidence intervals too narrow

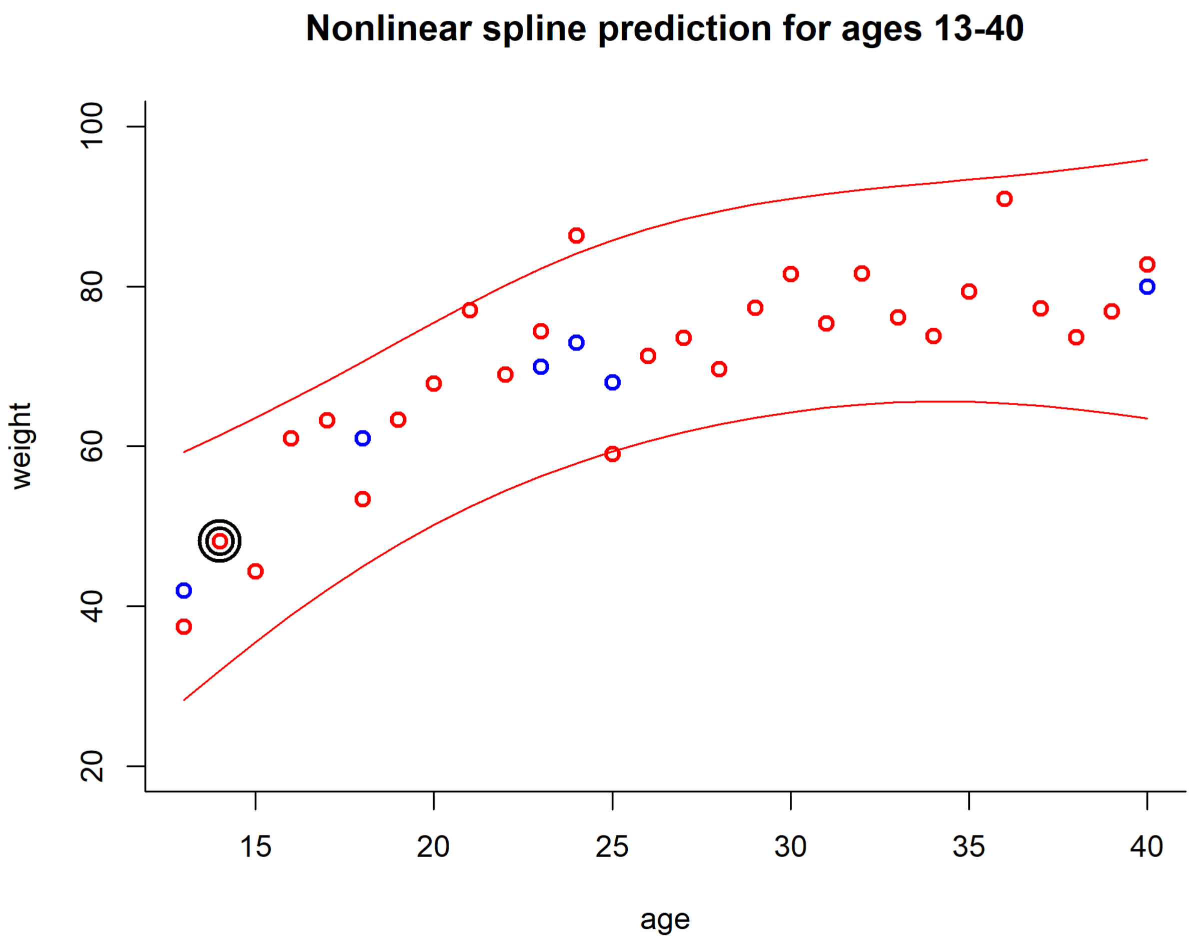
- Harmful to statistical inference

Stochastic regression imputation

- Like regression imputation, but adds appropriate noise to the predictions to reflect uncertainty
- Uncertainty in the form of:
  - Parameter uncertainty in the prediction model
  - Uncertainty due to unexplained variance in the target feature

- Related to *prediction intervals* (ISLR sections 3.2.2-four and exercises on pages 111-112)

Stochastic regression imputation



Stochastic regression imputation

```
##    age weight
## 1 13   42
## 4 14   NA
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

→

```
##    age weight
## 1 13   42
## 4 14   48.13
## 6 18   61
## 5 23   70
## 3 24   73
## 7 25   68
## 2 40   80
```

Stochastic regression imputation

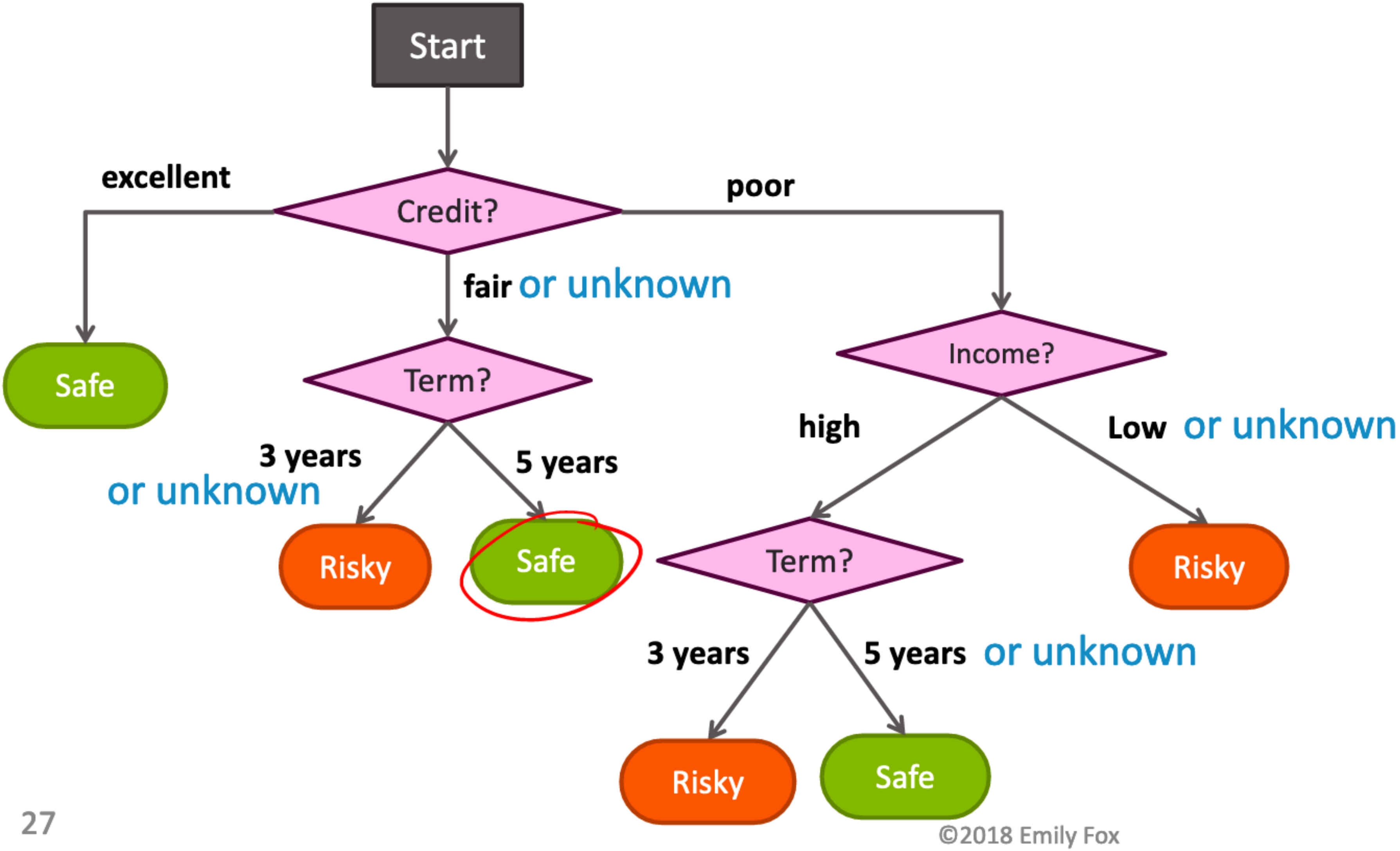
- Advantages:
  - Preserves the distribution of weight
  - Preserves the correlation between age and weight in the imputed data

- Disadvantages:
  - Symmetric and constant error restrictive
  - Single imputation does not take uncertainty imputed data into account, and incorrectly treats them as real
  - Not so simple anymore

“Embedded” methods (model-based)

- Don’t impute, deal with missing values somehow in the (prediction) model itself
- Depends on the model you are using
- Almost always assumes MAR implicitly
- Example on next slide given with classification tree, but other models may have a different approach

x_i = (Credit = ?, Income = high, Term = 5 years)



Source: Fox (2018), https://courses.cs.washing

Table: assumptions of the methods

	Assumption for unbiased statistics			
	Mean	Regression coef.	Correlation	Standard Error
Listwise deletion	MCAR	MCAR	MCAR	Too large
Mean imputation	MCAR	-	-	Too small
Regression imputation	MAR	MAR	-	Too small
Stochastic imputation	MAR	MAR	MAR	Too small

Interim conclusion

- Missing data problems are pervasive and important
- Ad hoc correction for missing values may work, but has assumptions
- Missing data imputation (data wrangling) and conclusions (data analysis) are intertwined

- Today: assignment ad-hoc methods for data imputation

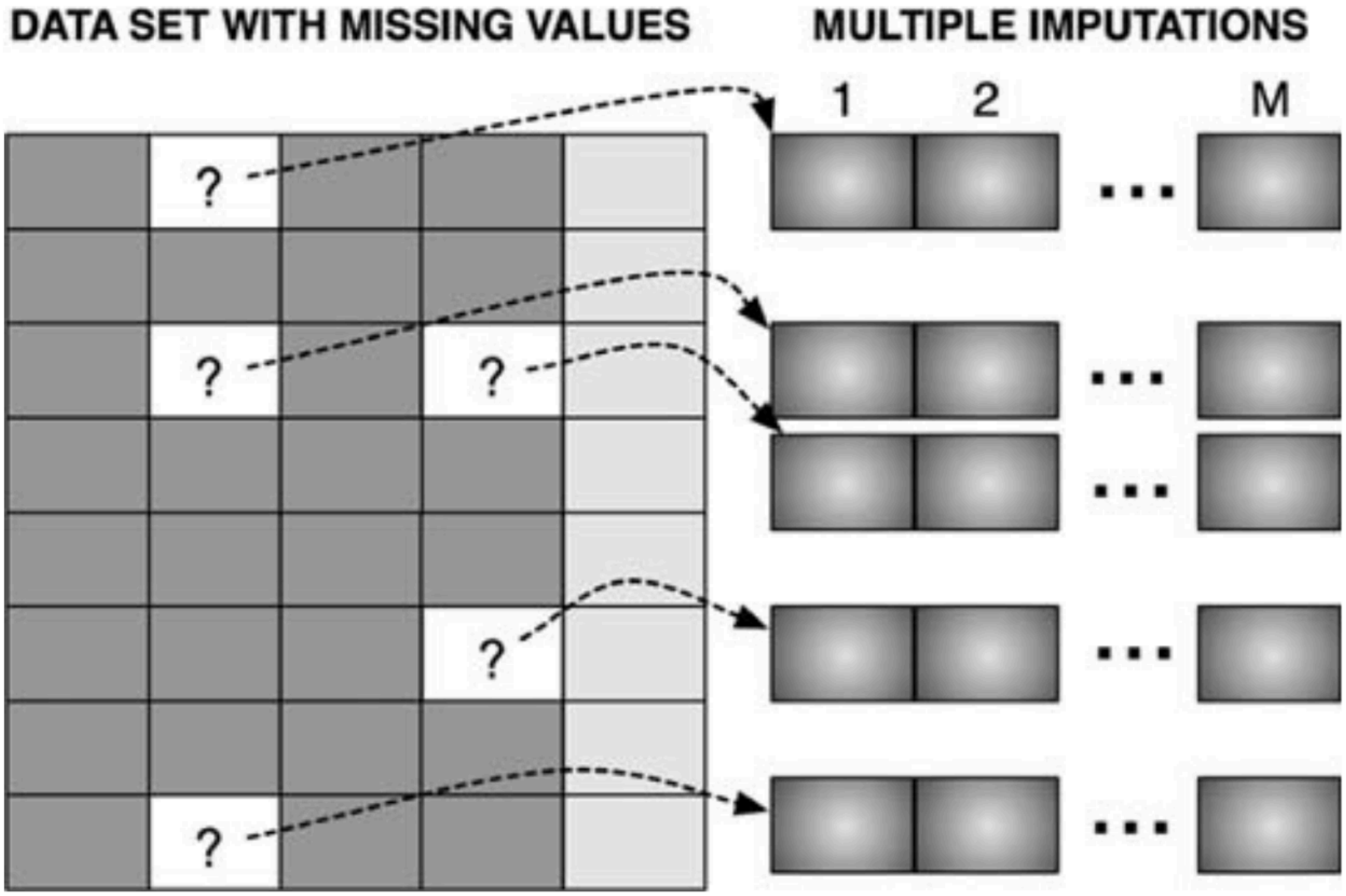
Imputing one value for a missing datum cannot be correct in general, because we don’t know what value to impute with certainty (if we did, it wouldn’t be missing).

Donald B. Rubin

Fixing the SE: Multiple Imputation

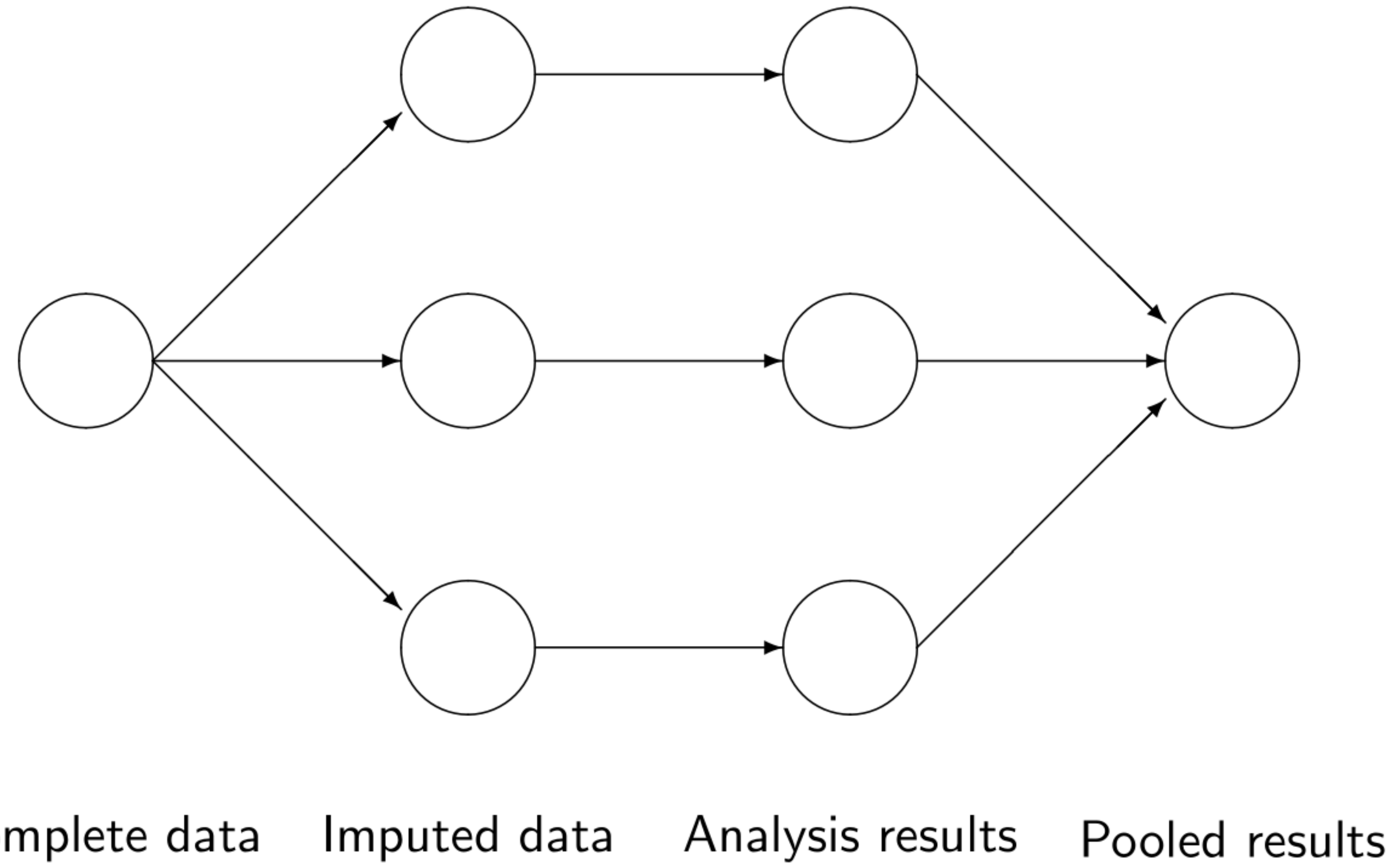
- Rubin (1987) "Multiple Imputation for Nonresponse in Surveys"
- Stochastic imputation, but create multiple datasets ($M = 5$ to 20)
- Each dataset is slightly different
- Appropriately consider the uncertainty around the imputed value

- Perform analysis multiple times
- Pool results



García-Laencina et al 2009

Multiple imputation



Difficult step: pooling

- How to pool the results of your data analysis?

- Estimand Q (e.g., mean length of population)
- Estimator $\hat{Q}_m$ (mean length in one imputed dataset m)
- Estimator $\overline{Q}$ (average of the M different $\hat{Q}_m$ estimates):

$$\overline{Q} = \frac{1}{M}\sum_{m=1}^{M}\hat{Q}_m$$

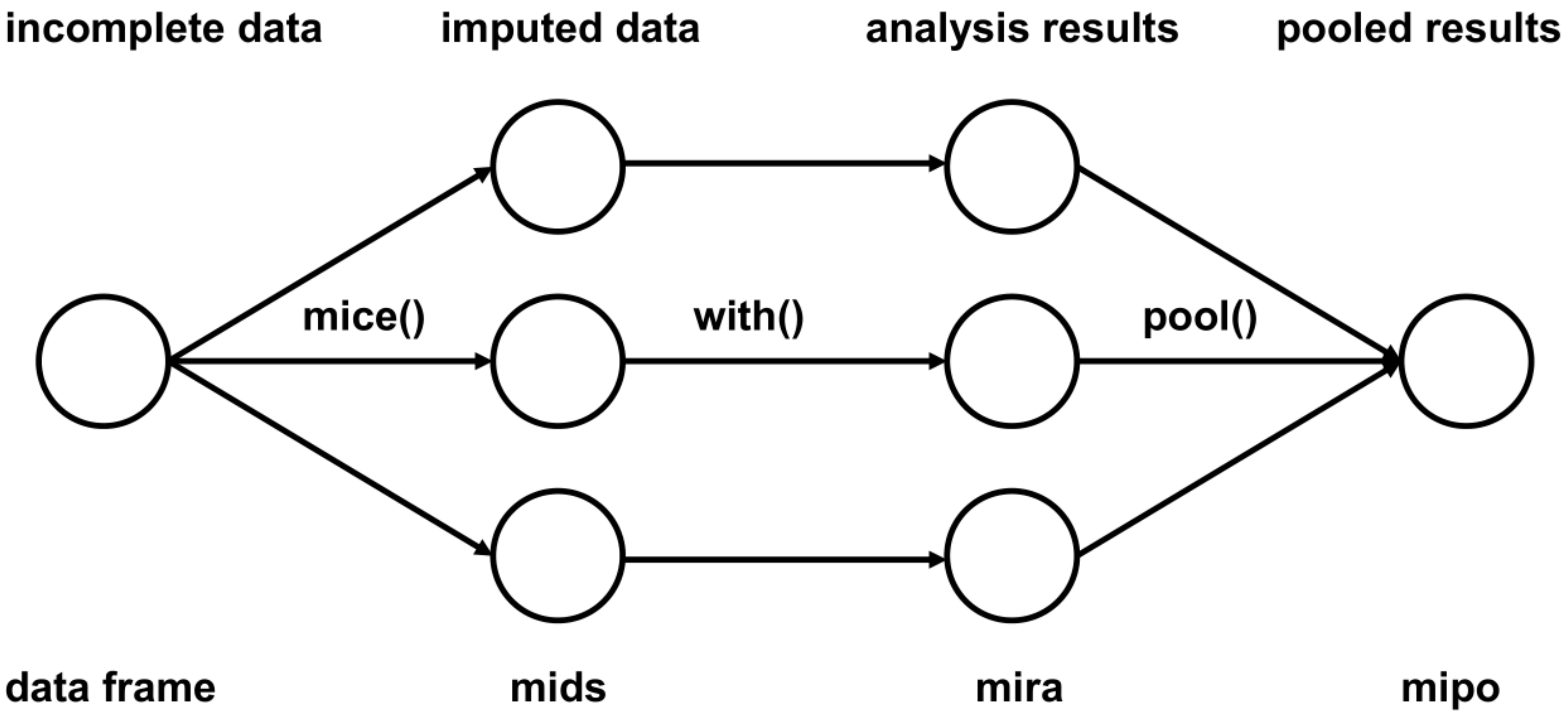
Difficult step: pooling

- Uncertainty / variance around estimator $\overline{Q}$ has three sources:
  - Within-dataset variance: the variance caused by the fact that we are taking a sample rather than the entire population. This is the conventional statistical measure of variability; the uncorrected standard error
  - Between-dataset variance: the extra variance caused by the fact that there are missing values in the sample;
  - Simulation error: the extra variance caused by the fact that $\overline{Q}$ itself is based on a finite amount of datasets M (this uncertainty decreases as M increases)

Solution: `mice`

- The R package mice performs multiple imputation and automatic pooling of results
- It has support for many analysis methods, such as anova(), lm(), glm(), and many more
- Thursday: assignment for data analysis with multiple imputation
- Read FIMD sections 2.1, 2.3, 2.4, 2.7, 2.8, 3.1, 3.4

Typical workflow for mice



Ad hoc alternative to pooling: sensitivity

- If conclusion of interest does not change across the m imputed datasets, we say the conclusion is not sensitive to the imputation
- For your topic this may be enough

Conclusion

- Single imputation does not account for all uncertainty
- One solution is multiple imputation
- Analysis needs to be pooled after being performed on multiply imputed datasets
- This is solved for many methods (mice), but might not be solved for all methods
- Sensitivity analysis can be an alternative to pooling
- Multiple imputation fixes inferences, but still has the MAR assumption!