

Data Wrangling and Data Analysis

Missing Data and Imputation

Daniel Oberski

Department of Methodology & Statistics

Utrecht University

This week

- Monday: Missing data
- **Tuesday: What to do about missing data**
- Wednesday: Clustering #1 (**video!**)

Strategies to deal with missing data



Prevention

(the golden road?)

Imputation


- Ad-hoc methods
- Single imputation
- Multiple imputation

Adjustment

- Weighting methods
- Likelihood methods
- EM-algorithm
- ...

Imputation

Replacing missing values with *guessed* values

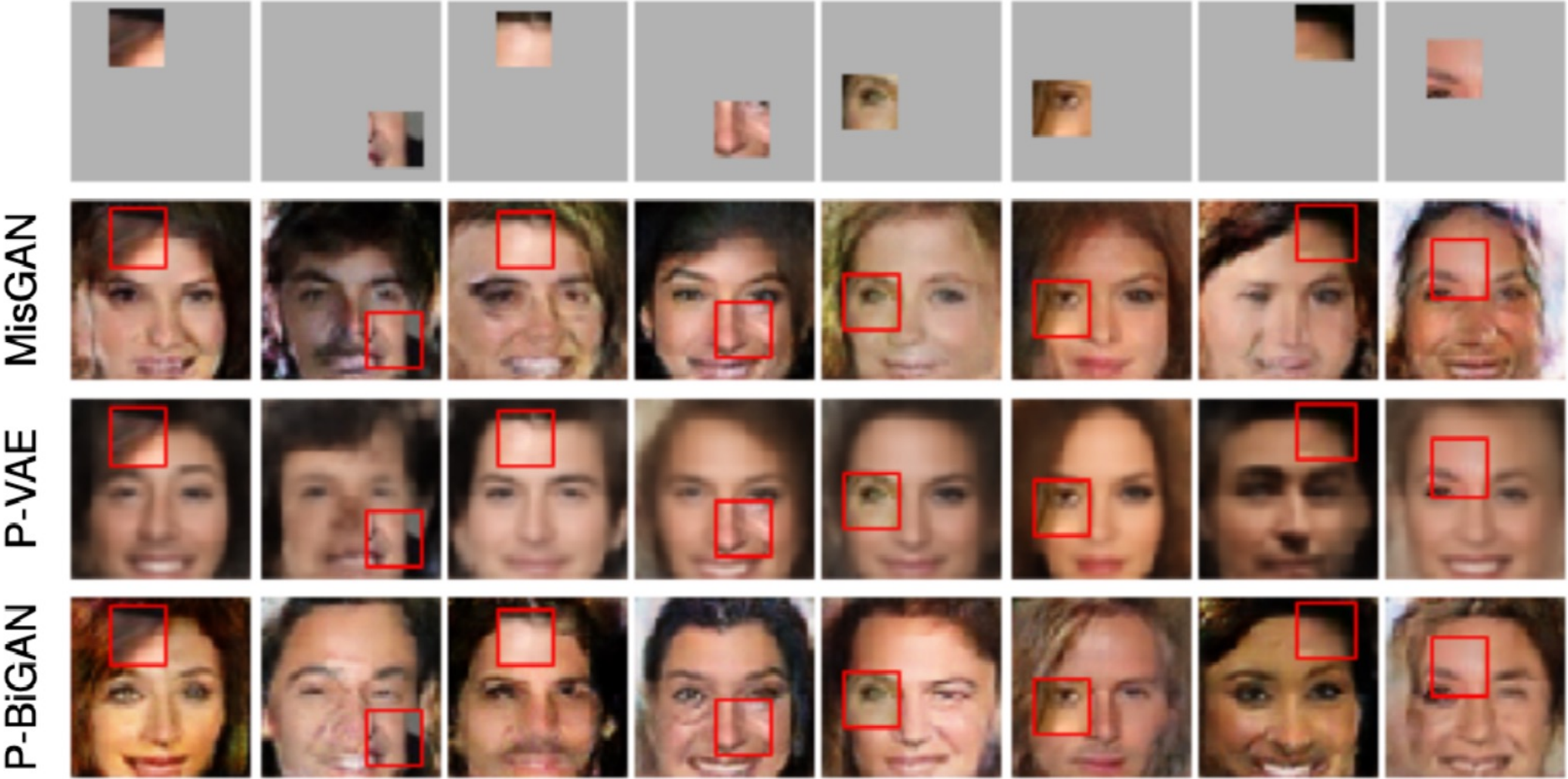
##	age	weight		##	age	weight		
##	1	13	42	##	1	13	42	
##	4	14	NA		##	4	14	47
##	6	18	61	##	6	18	61	
##	5	23	70	##	5	23	70	
##	3	24	73	##	3	24	73	
##	7	25	68	##	7	25	68	
##	2	40	80	##	2	40	80	

Imputation

- Yields “complete” data
- Statistics are now defined
- Hooray?



Imputation can look quite OK with the right model



Cheng-Xian Li & Marlin (2020), ICML

Deductive imputation

- If we know height and weight, we can calculate BMI
- A male child is probably not pregnant

Converse also holds: If an *observed* value “must” be wrong, we can correct it 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 (perhaps it should have been 41?). We can set this value missing and let our algorithms do the rest.

Missing data procedures and your (implicit) assumptions

- Missing data important **because they can bias the analysis**
- Bias depends on NDD/SDD/UDD
- **Goal: prevent the bias**
- **Logic of all missing data procedures:**
 - IF NDD/SDD is causing the bias,
 - THEN procedure *xyz* will remove that bias
- (On very rare occasions, specific forms of UDD can be dealt with)
- In other words, each procedure has an **implicit assumption**

Methods and their assumptions

Assumption needed to get unbiased estimate

Mean	Regression coef.	Correlation	Standard Error
------	------------------	-------------	----------------

Listwise deletion

Mean imputation

Regression imputation

Stochastic imputation

Listwise deletion

Also known as:

- “Complete Case Analysis”
- “we deleted missing data”
- “huh? Whaddayamean missing data?”

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 NDD

Disadvantages

- Large loss of information, hopeless with many features
- Biased under SDD and UDD, even for simple stuff like mean

Mean imputation

Replace the missing values by the mean of the observed data

Advantages

- Simple
- Unbiased for the mean, under NDD

Disadvantages

- Doesn't work
- (okay, except in some very specific circumstances)

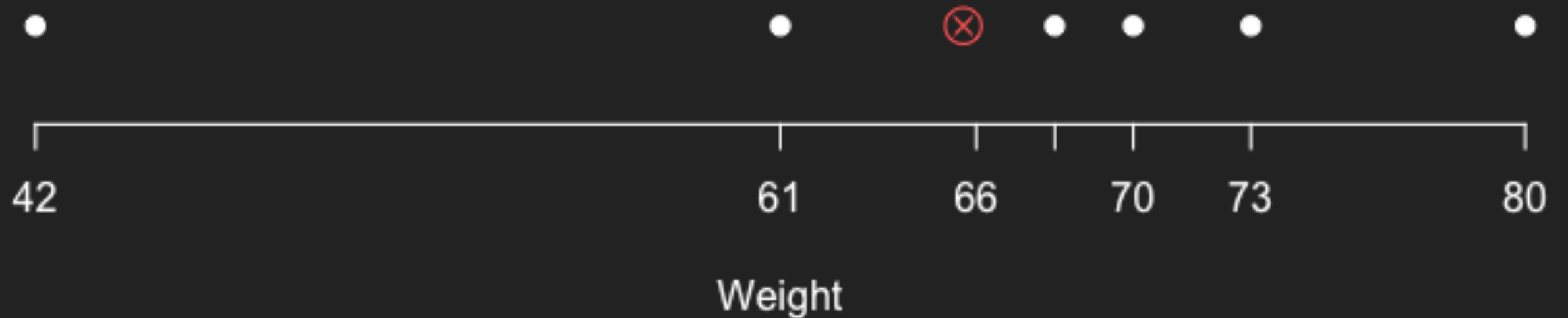
Mean imputation

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

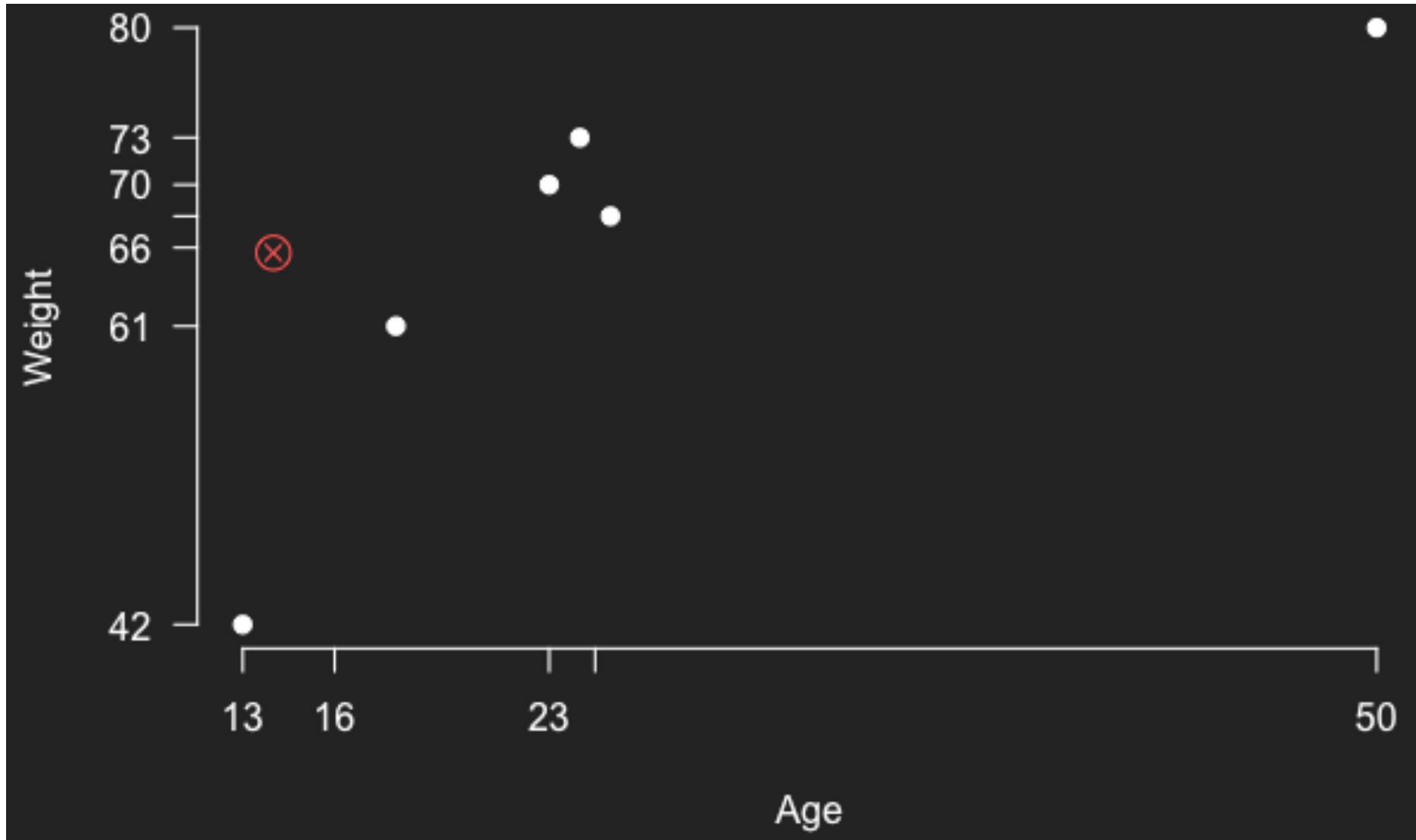


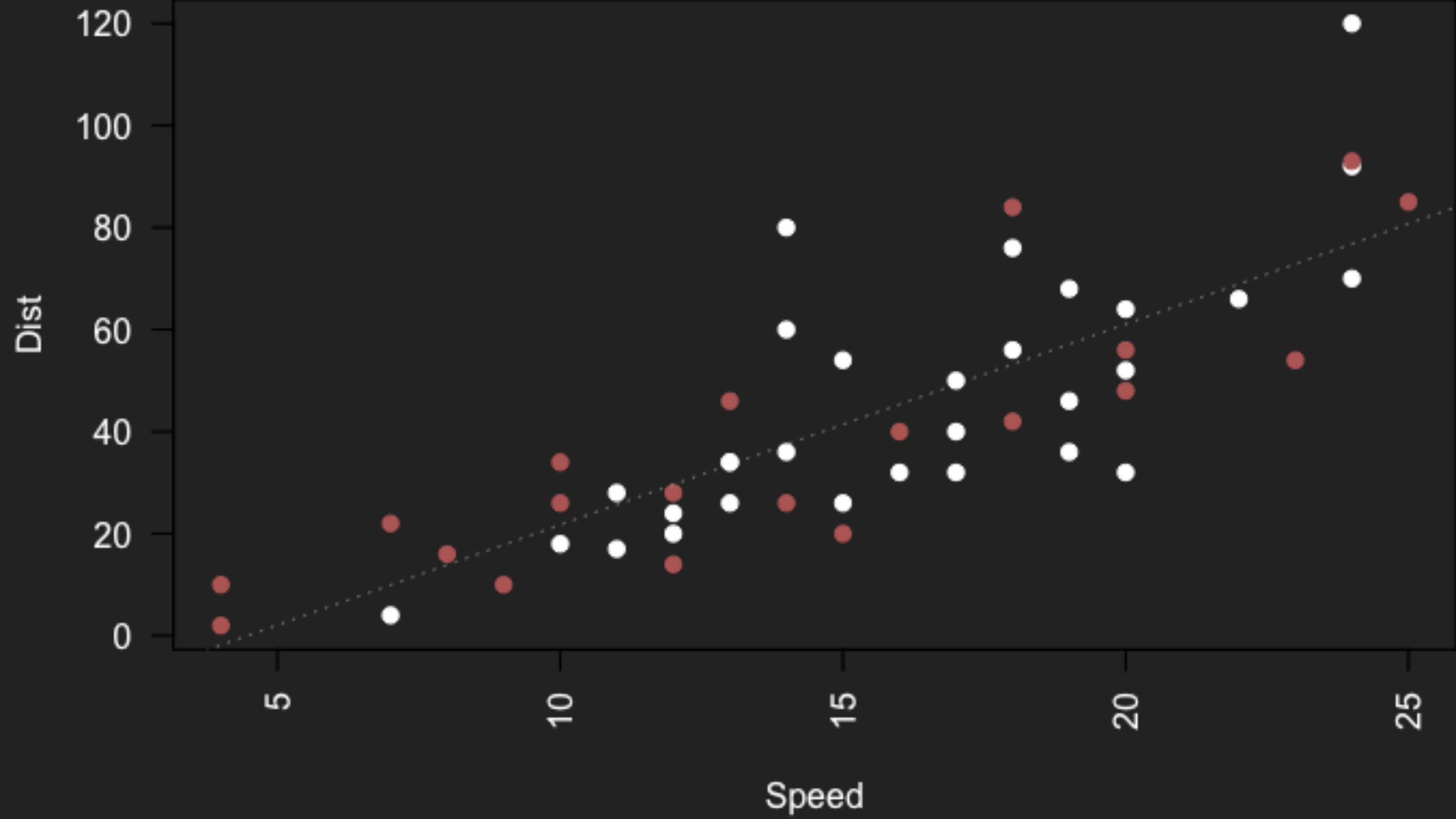
##		age	weight
##	1	13	42
##	4	14	65.667
##	6	18	61
##	5	23	70
##	3	24	73
##	7	25	68
##	2	40	80

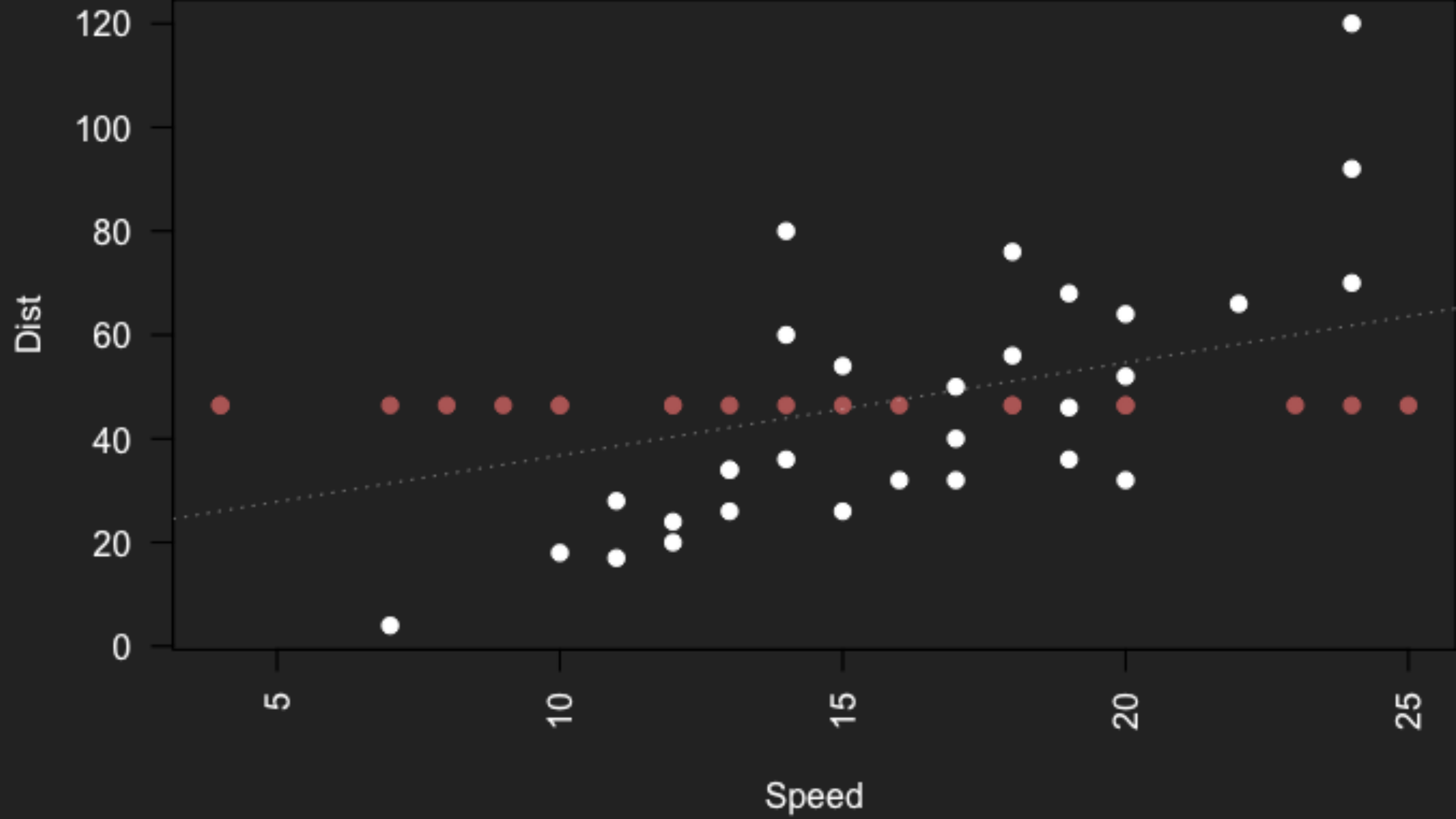
Mean imputation: univariate perspective



Mean imputation: bivariate perspective







Mean imputation

Disadvantages

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

AVOID

(unless you know what you are doing (and probably even then))

**Intermezzo: tiny crash course/reminder
of classical statistics**



Sample

Population/Data-generating process (DGP)

“Uncertainty”

- Suppose there is a target “number of interest” in the DGP
- We only have **ONE** sample that will let us estimate this target number
- How close can we expect this estimate to be to the true number, given the sample size?



“Uncertainty”

- **Think about:**
 1. all the possible spoonfuls I could have taken
 2. the estimates of our target number I would have gotten
 3. the standard deviation of estimates over spoons
- Here is the **Big Trick**: we can estimate this **standard deviation** from our ONE sample!
- This estimate is called the “**standard error**”
- You can think of the **standard error** as an estimate of the **uncertainty** about your estimate



Using the standard error to deal with uncertainty

- Every estimate you care about should preferably come with a standard error
- Or some other measure of uncertainty (e.g. Bayesian)



You can also use the standard error to:

- Calculate “confidence intervals”

Intervals with the following handy interpretation: “in 95% of the spoonfuls you will obtain a confidence interval that contains the true target value”

- Calculate “p-values”

A decision rule that can tell you that when you lead your life as though some hypothesis were true, then you would be wrong 5% of hypothetical spoonfuls.

“Regression” imputation (in the statistics sense)

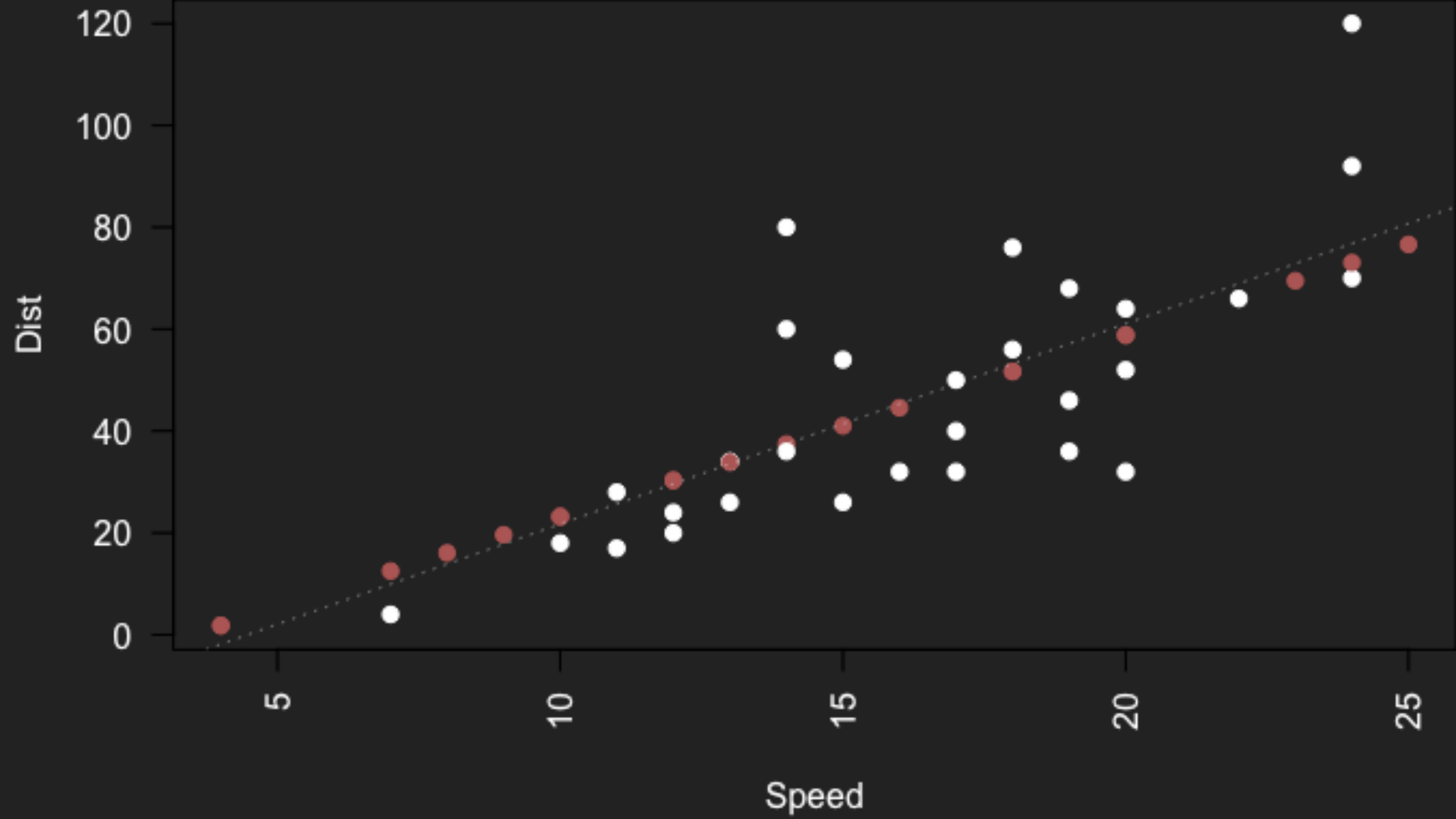
- Just *predict* the missing value
- Fit model for weight under listwise deletion: the **imputation model**
- Predict body weight for records with missing weight
- Replace missing values by prediction

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



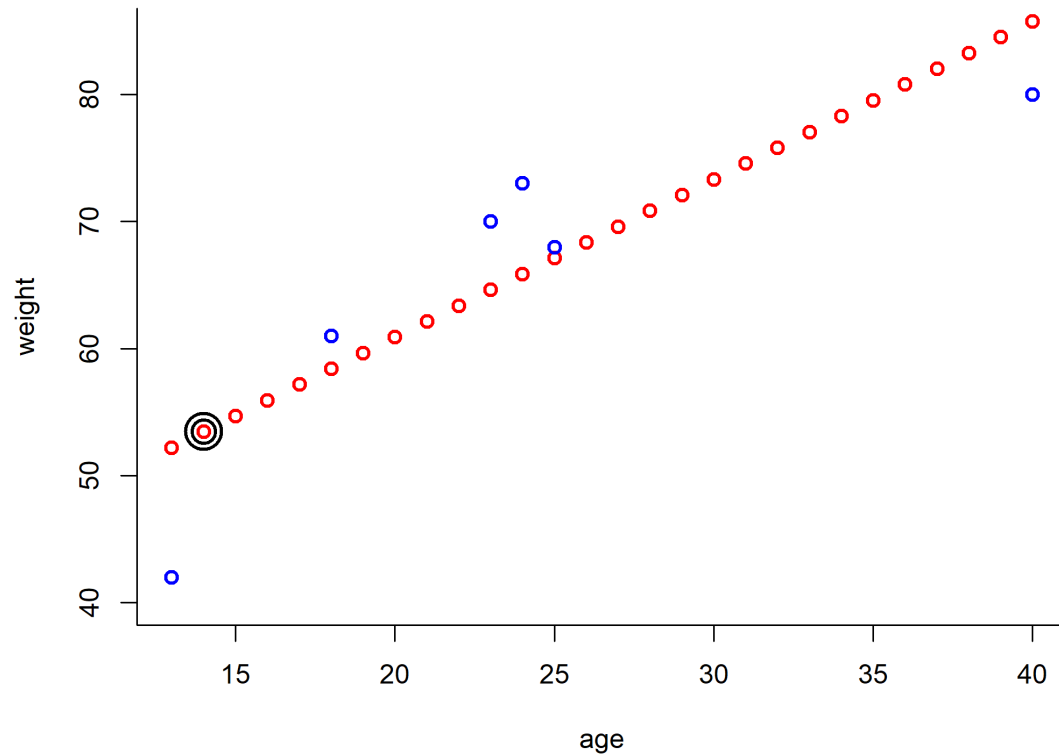
Regression imputation

Advantages

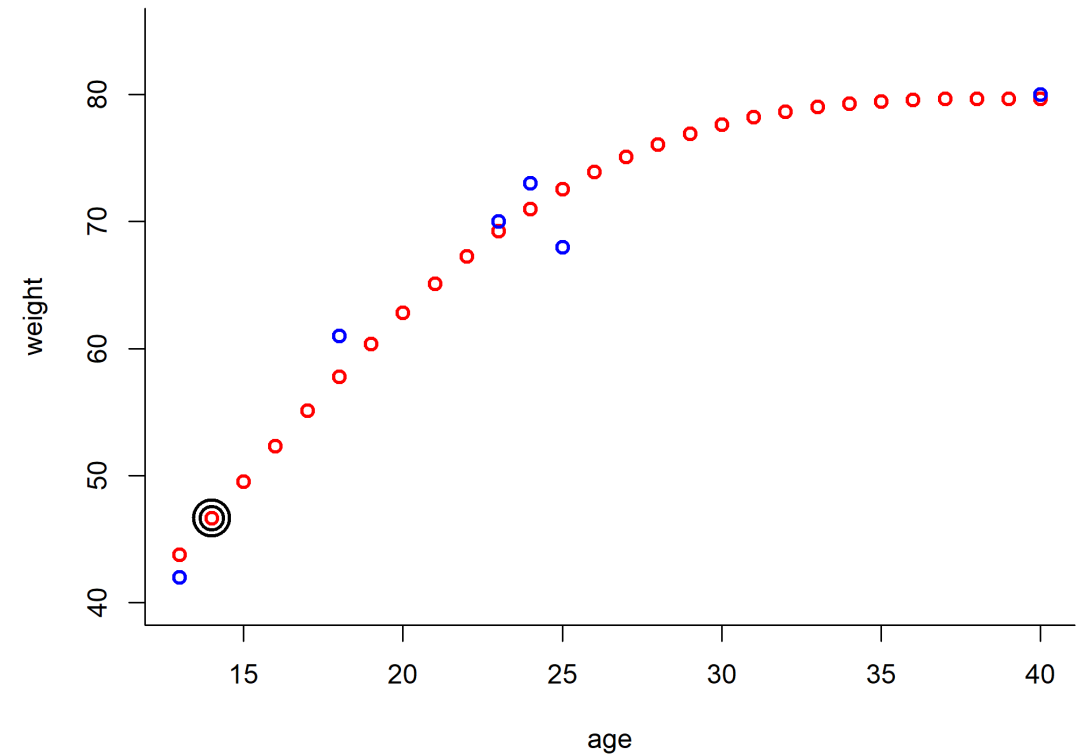
- Unbiased estimates of regression coefficients (under SDD)
- Good approximation to the (unknown) true data if explained variance is high
- The better your prediction, the better your approximation

The better your prediction, the better your approximation

Linear regression prediction for ages 13-40



Nonlinear spline prediction for ages 13-40



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

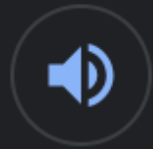
Regression imputation

Disadvantages:

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

Stochastic regression imputation

- Like regression imputation, but adds 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

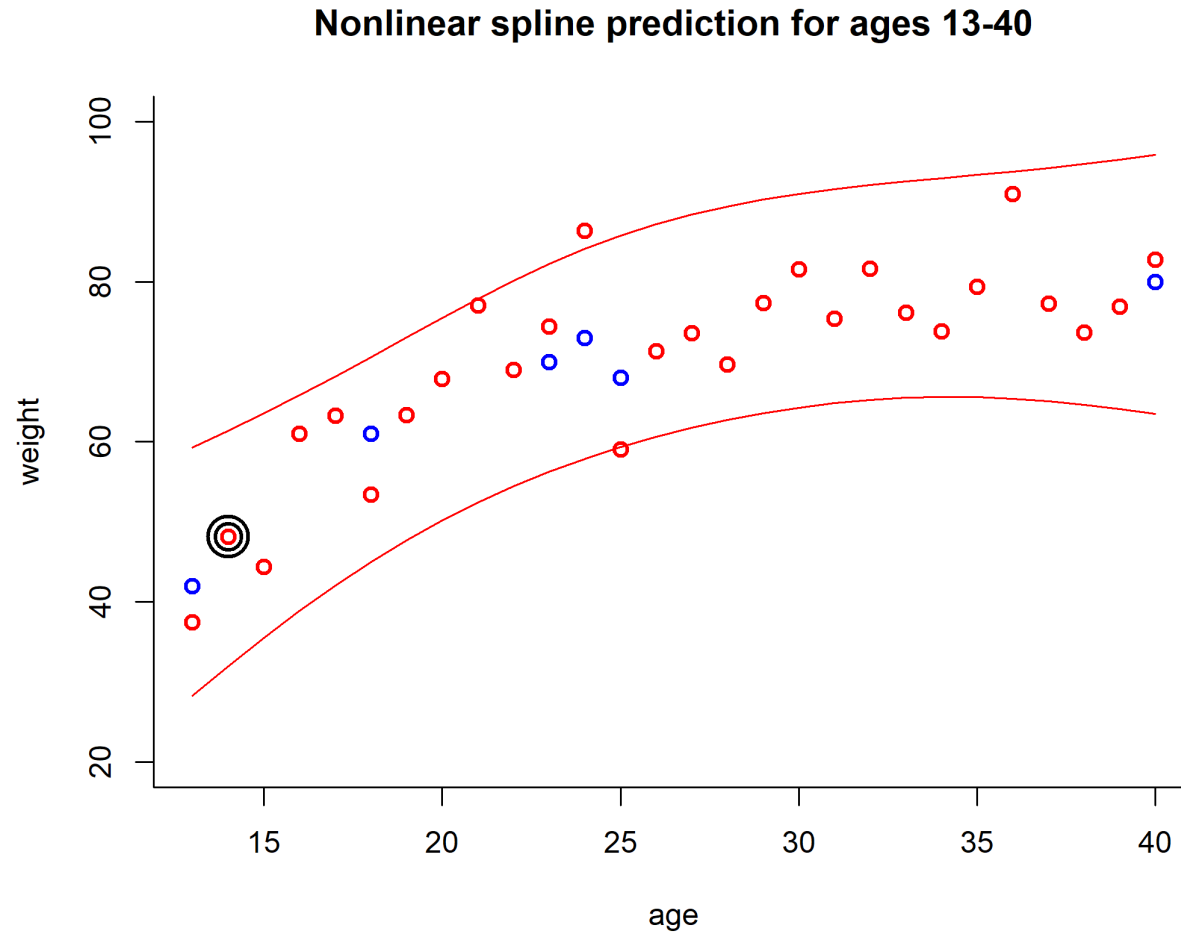
/stə'kastɪk/

adjective

TECHNICAL

having a random probability distribution or pattern that may be analysed statistically but may not be predicted precisely.

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 body weight
- Preserves the correlation between age and weight in the imputed data

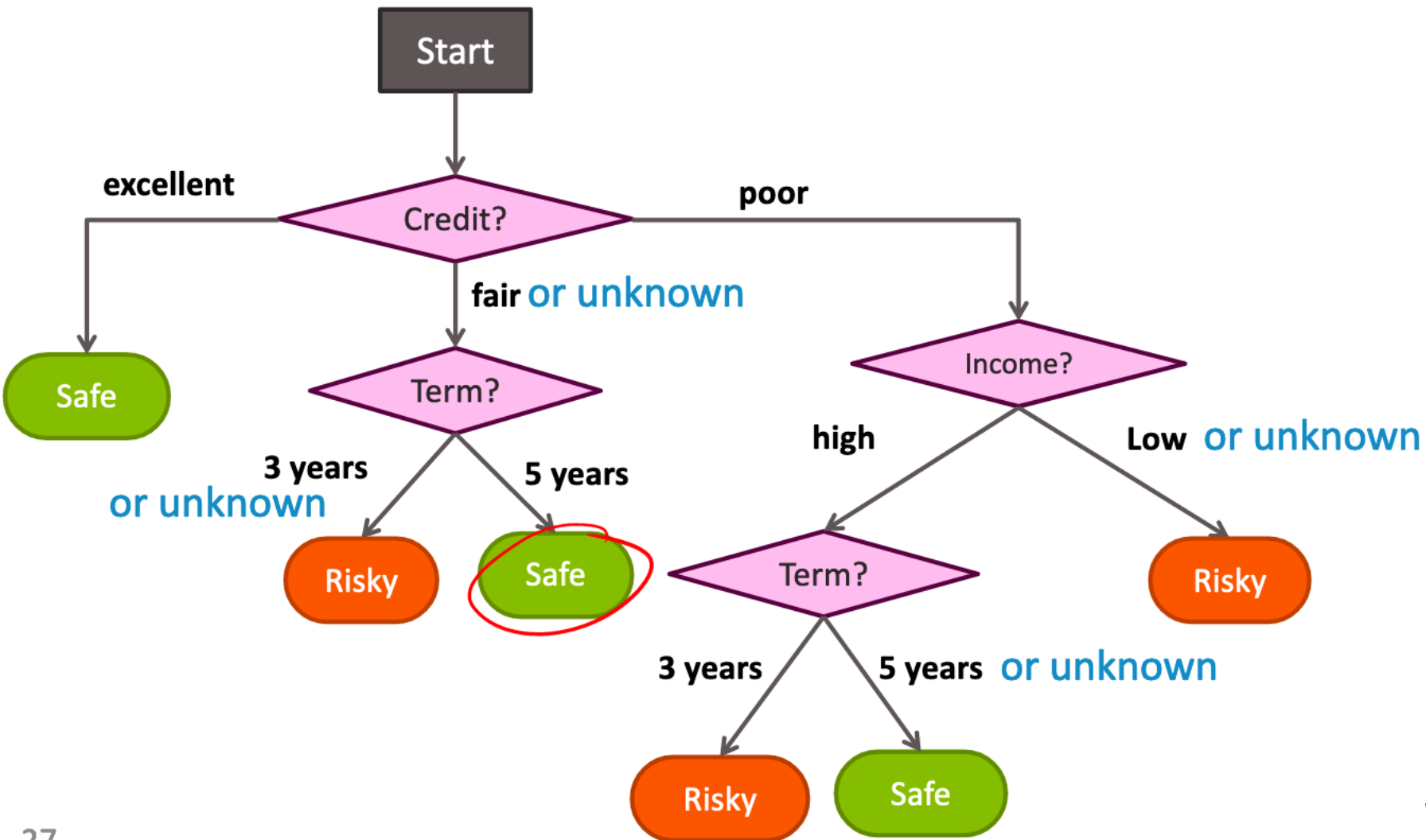
Disadvantages:

- Fiddly
- Single imputation does not take uncertainty imputed data into account, and incorrectly treats them as real

“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 SDD implicitly
- Example on next slide given with classification tree, but other models may have a different approach

$x_i = (\text{Credit} = ?, \text{Income} = \text{high}, \text{Term} = 5 \text{ years})$



Source: Fox (2018), <https://courses.cs.washingt>

Handy overview of methods

	Assumption needed to get unbiased estimates			
	Mean	Regression coef.	Correlation	Standard Error
Listwise deletion	NDD	NDD	NDD	Too large
Mean imputation	NDD	-	-	Too small
Regression imputation	SDD	SDD	-	Too small
Stochastic imputation	SDD	SDD	SDD	Too small

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 Rubin

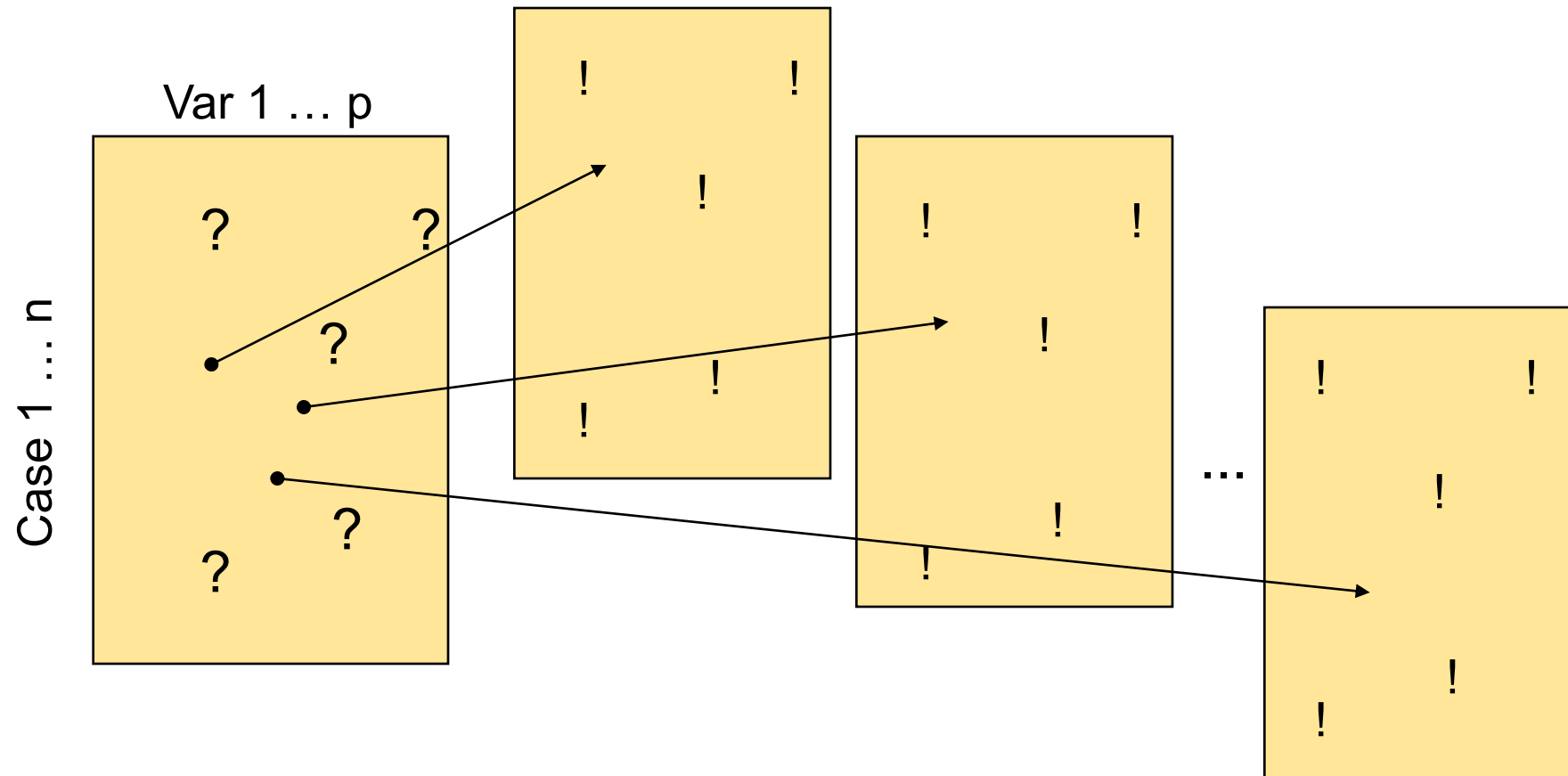
Fixing the SE: Multiple Imputation

- 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

Steps in Multiple Imputation

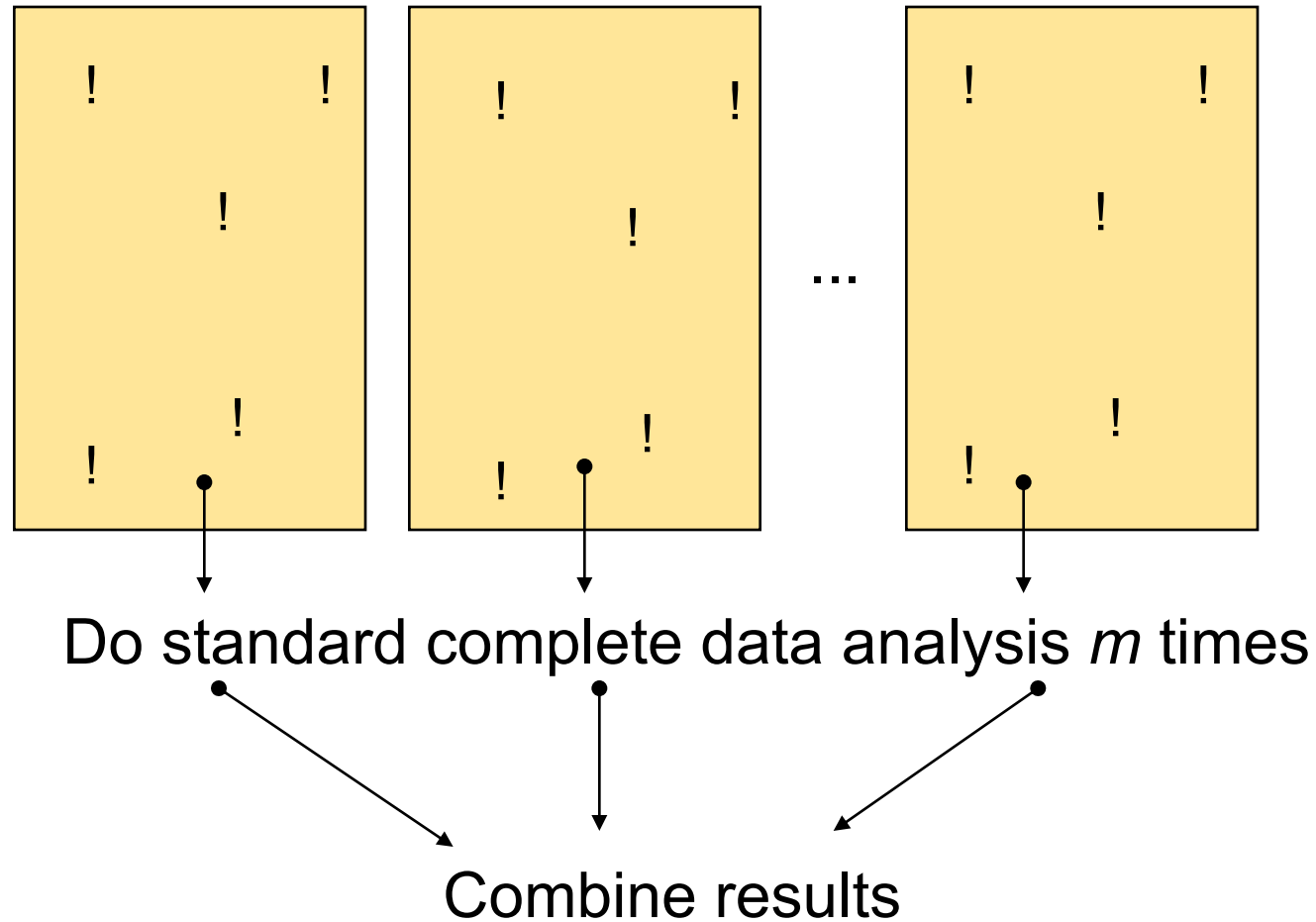
1. Create imputations
2. Analyze completed data sets
3. Combine the results

Multiple Imputation: Imputation Step

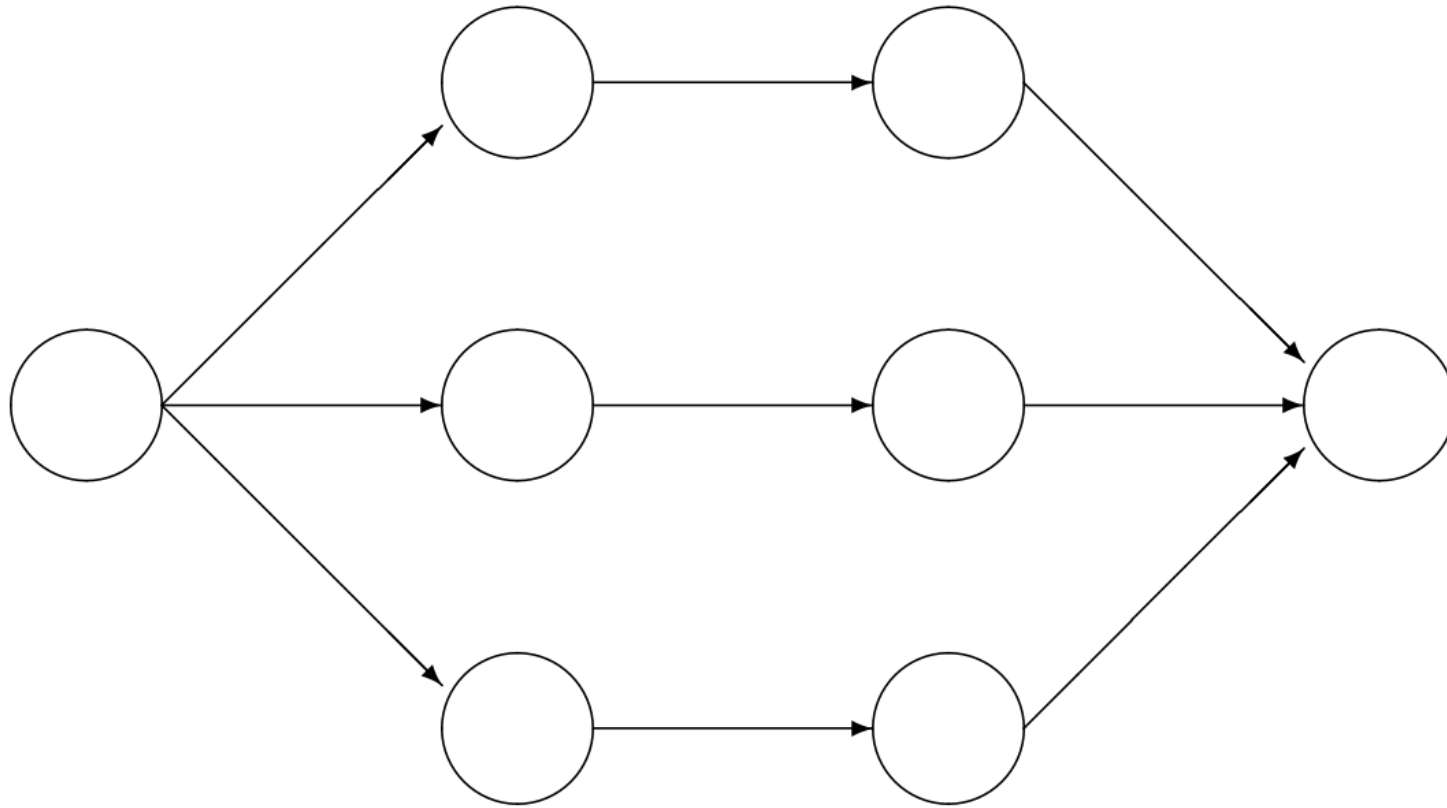


Create m different imputed data sets

Multiple Imputation: Analysis Step



Multiple imputation



Incomplete data

Imputed data

Analysis results

Pooled results

Combine the Results

Simply compute mean of m estimates

$$\bar{Q} = \frac{1}{m} \sum \hat{Q}_i$$

where \hat{Q}_i is the estimate in the i -th multiply imputed dataset.

“The overall estimate is the average of the estimates”

Difficult step: pooling

- Uncertainty / variance around estimator \bar{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 \bar{Q} itself is based on a finite amount of datasets M (this uncertainty decreases as M increases)

Combining Standard Errors

U : Within imputation variance = mean of m sampling variances (square of s.e.)

$$\bar{U} = \frac{1}{m} \sum U_i$$

B = Between imputation variance = variance of point estimates

T = Total error variance

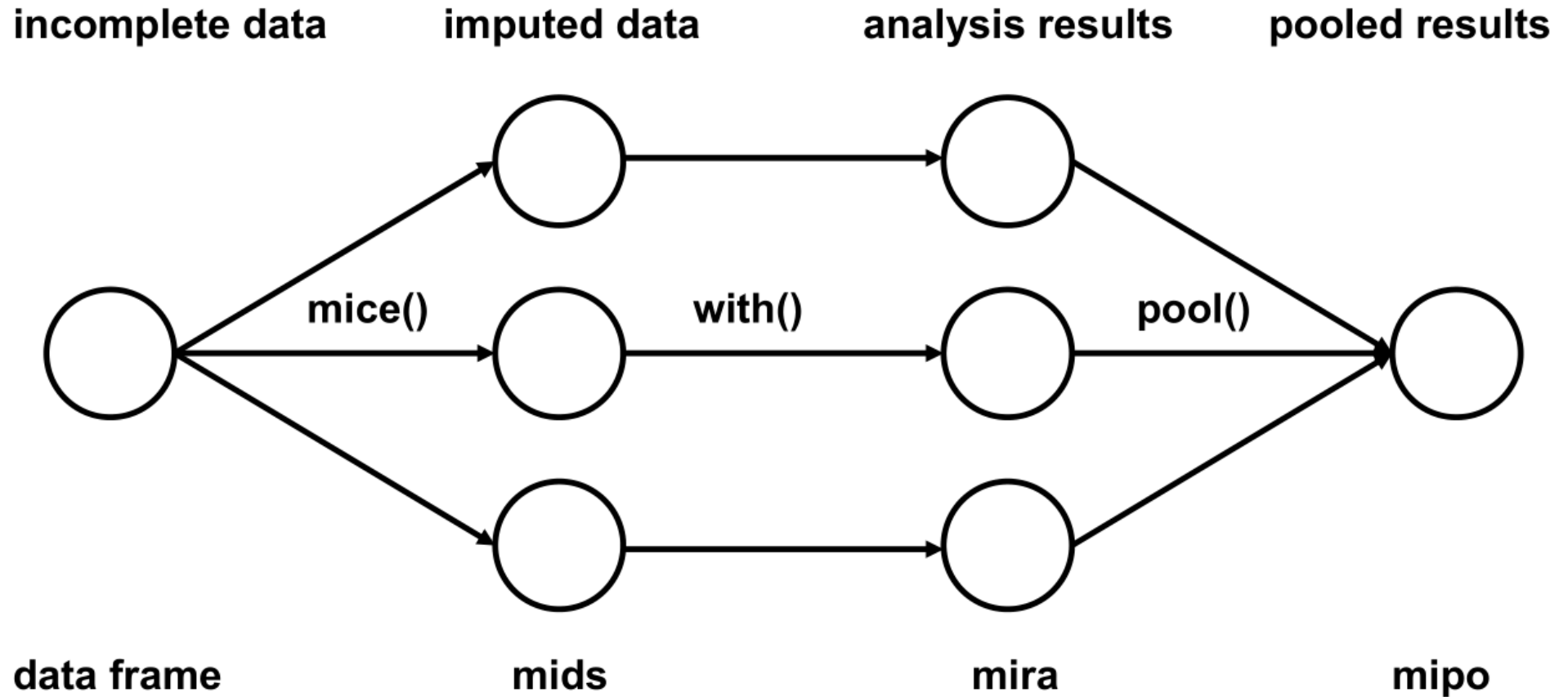
$$B = \frac{1}{m-1} \sum (\hat{Q}_i - \bar{Q})^2$$

$$T = \bar{U} + (1 + m^{-1})B$$

Software

- 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
- `sklearn` has an implementation of `mice` algorithm in `sklearn.impute.IterativeImputer`
- `sklearn` documentation: “It is still an open problem as to how useful single vs. multiple imputation is in the context of prediction and classification when the user is not interested in measuring uncertainty due to missing values.”

Typical workflow for mice



Ad hoc alternative to pooling: sensitivity

- If conclusion does not change across the m imputed datasets, conclusion 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
- Sensitivity analysis can be an alternative to pooling
- Multiple imputation fixes inferences, but still has the SDD assumption!