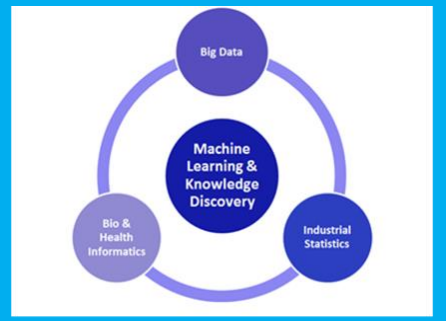
521156S
TOWARDS DATA MINING

MATKALLA TIEDONLOUHINTAAN

UNIVERSITY OF OULU

BISG Oulu



Data Analysis and Inference Group

How to **prepare** your data to make sure that your data mining process will be successful…

Tuomo Alasalmi
tuomo.alasalmi@oulu.fi

TOPICS OF THE LECTURES

1. Introduction to data preprocessing
2. Data ethics, data security, privacy and open data
3. Data management and databases
4. Data gathering
5. **Missing data**
6. Noise, outliers, signal saturation
7. Normalization, transformation, dependence, distributions
8. Feature selection and ensuring the generality of the results

Data on uusi vesi

Datan mahdollistamat uudet palvelut ja liiketoiminnot ovat innoittaneet analyytikoita ja it-taloja vertaamaan dataa öljyyn. Vertauksissa data mahdollistaa samanlaisen murroksen kun öljyn hyödyntäminen 1900-luvun taitteessa.

Näissä analogioissa on aina ongelmansa. Nykyisin on mahdollista tehdä esimerkiksi analytiikkaa, tiedolla johtamista ja tekoälysovelluksia ennennäkemättömällä tavalla. Dataa on saatavilla ylen määrin joka puolella. Tässä kohtaa vertaus öljyyn ontuu.

ÖLJY ON rajallinen resurssi, jonka viimeisten pisaroiden hyödyntäminen on koko ajan vaikeampaa. Dataa puolestaan on yllin kyllin. Hyödyntämisessä ongelmana on enemmänkin se, että läheskään kaikki saatavilla oleva data ei ole yrityksille hyödyllistä. Virheellinen tieto on jopa haitallista.

Yhtä pätevä vertaus voisikin olla vesi. Vettä on maapallolla todella paljon, mutta vain pieni osa siitä on sellaisenaan ihmiselle juomakelpoista. Datastakaan ei voi hyödyntää kuin pienen osan ilman puhdistamista.

JOS PUHDISTETTUUN dataan sekoittaa likaista dataa, menee lopputulos helposti käyttökelvottomaksi. Jos pohjalla olevat bitit eivät ole kohdallaan, on asiakkaan kokema palvelu vähintäänkin kulmakarvoja kohottavaa.

Näin äskettäin esimerkin rengashotellista. Asiakkaan mobiilisovellukseen tuli raportti säilytyksessä olevien renkaiden kulumisesta. Raportissa oli yksi selvästi kulunut rengas, johon piti kiinnittää huomiota. Palvelu on sinällään upea, mutta valitettavasti asiakkaan hotelliin toimittamat renkaat olivat iskemättömät.

Erityisen vaativaa on luoda luotettavaa pohjaa erilaisista big data -lähteistä. Tiedon rakenne on aina puutteellista ja sen ajantasaisuutta ei pysty varmistamaan.

MYÖNNETÄÄN, ei data ole myöskään uusi vesi. Vertailukohtia on mukava hakea kaikille tutuista elementeistä. Data on raaka-aineena kiitollinen, sen määrä kasvaa eksponentiaalisesti. Hyvääkin dataa riittää aika monelle, kunhan vain pidämme huolen sen puhdistamisesta.

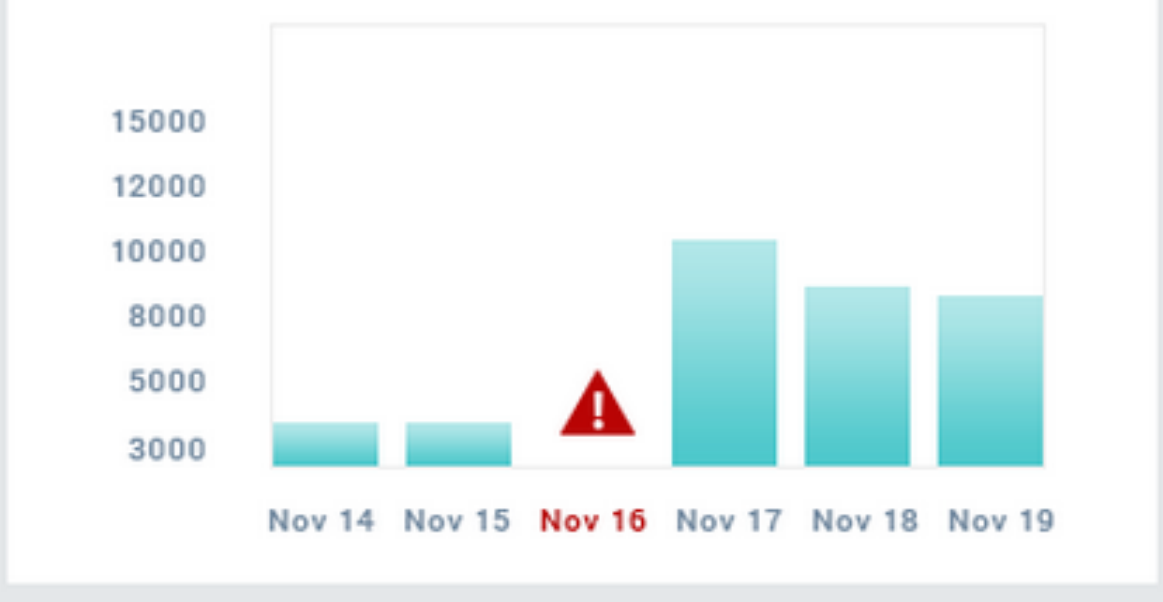
MOTIVATION

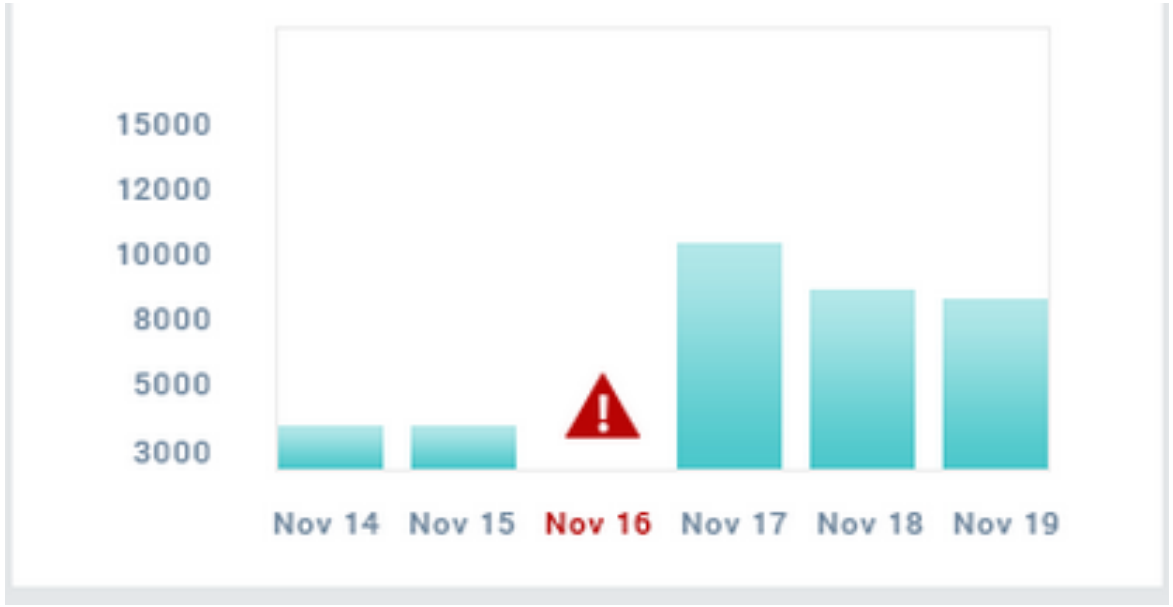
- **Data is everywhere and everyone wants to make use of it**
- **Raw, uncleaned data is not always very usefull**
- **Even the mainstream is starting to wake up to this (it's about time!)**
  - On the left is an editorial of an IT magazine TIVI from wk 40/2017 making the point that only some of data is usefull without cleaning
- **Even if we (well I don't) like to talk about artificial intelligence nowadays, there really isn't any real intelligence in these systems (at least yet)**
- **Without intelligence what we'll get instead is garbage in, garbage out**
- **We need to be the real intelligence in the process to make sure garbage is not what our models output**

MISSING DATA

- Missing values (MV) in data sets are very common in the real-world

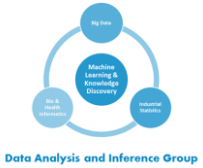
- Why?
- What kind of problems can missing values cause?





MISSING DATA

- **Missing values (MV) in data sets are very common in the real-world**
- **Why?**
  - Equipment errors
  - Incorrect measurements
  - Mistakes in manual data entry
  - Non-response in surveys
  - Using data that was not originally designed for this use or combining data from different sources
  - Etc.
- **Modelling methods often assume that data is complete**
- **Inappropriate handling of MVs can**
  - Introduce bias
  - Lead to misleading conclusions
  - Limit the generalizability of the research findings



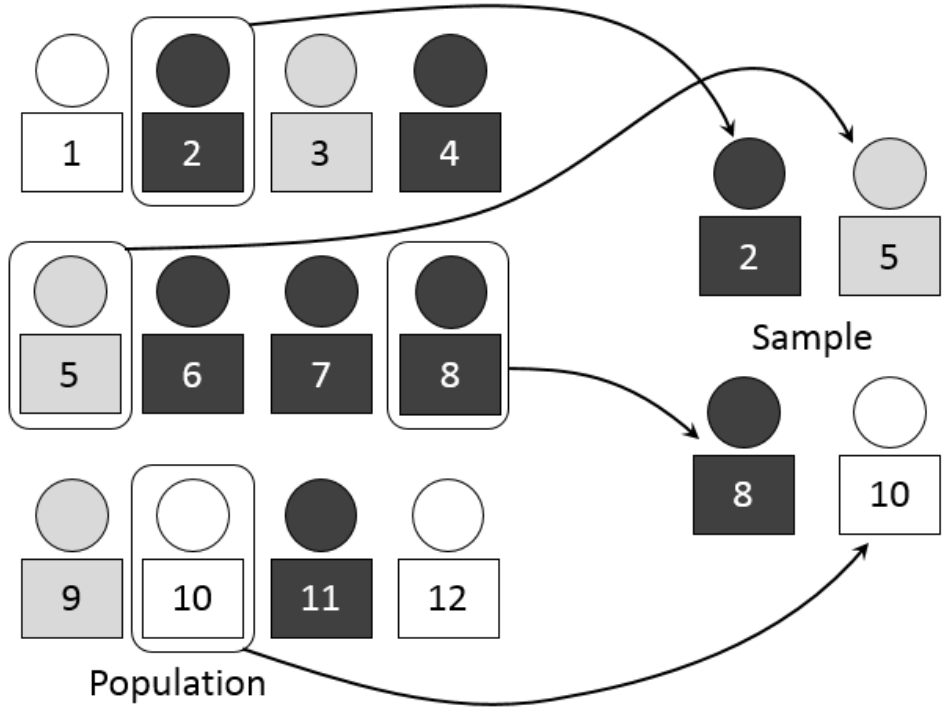
39. Oletko koskaan tupakoinut elämäsi aikana?

1 en *(→ siirry kysymykseen 45)*

2 kyllä, aloitin I____I____I vuotiaana

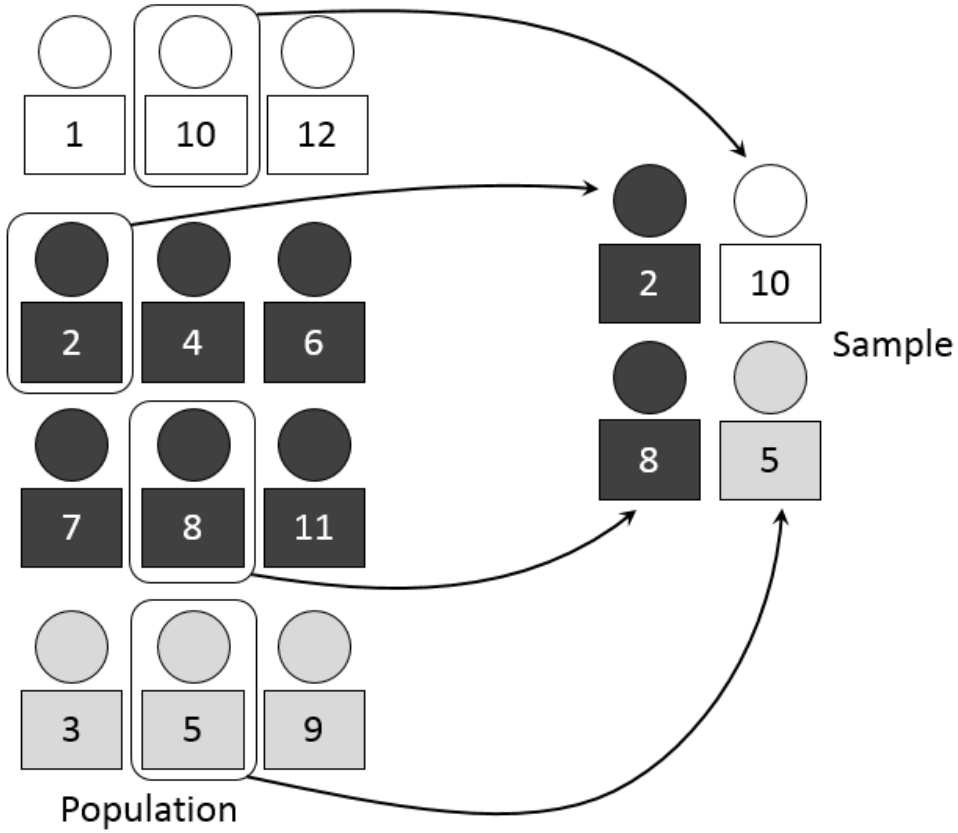
MISSING DATA

- **We need to know the mechanism that caused the missing values**
- **Is it clear from the data why some values are missing?**
  - Conditional questions in a survey
  - Some data relevant only in some cases (e.g. refraction information of glasses if one does not have glasses)
- **These things need to be taken into account in the analysis**



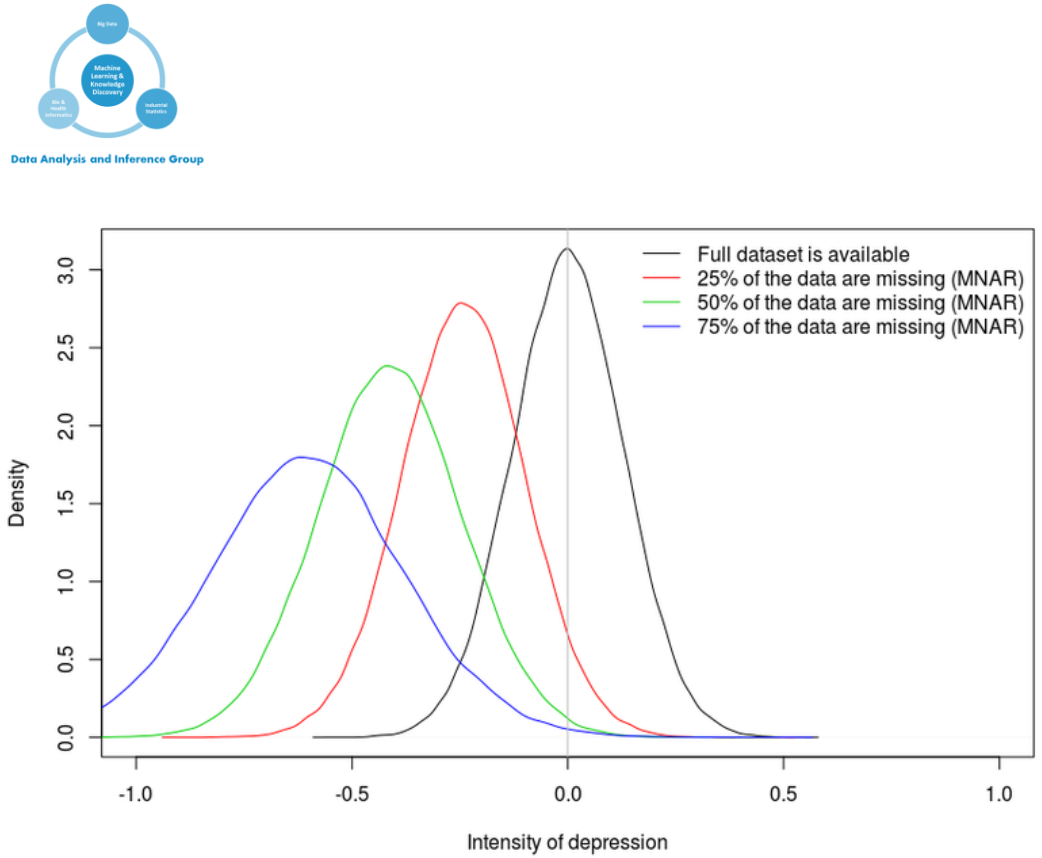
MISSING DATA

- Missing completely at random (MCAR)
- The probability that a value is missing does not depend on the missing value itself or the observed data
- Ignoring the MVs does not bias the results although some information is lost
- E.g. malfunctioning equipment or random sampling of a population (figure, data about people that are not in the sample are missing)
- Not often a realistic assumption



MISSING DATA

- Missing at random (MAR)
- The probability an observation is missing depends on the observed data but not on the missing data
- E.g. a certain sensor brand fails more often than another brand (brand is indicated in the observed data) or stratified sampling (figure)



MISSING DATA

- **Missing not at random (MNAR)**
  - Sometimes called not missing at random (NMAR)
- **The model for missingness is unknown (i.e. cannot be explained by the observed data)**
- **The probability of a value being missing could be explained by the missing value itself or some other data we do not have**
- **E.g. a sensor fails more often on certain data value range than in others or nonresponse in survey question is more frequent in people with a disease than without that disease (figure)**
- **MNAR analyses are difficult and do not necessarily perform better than methods assuming MAR**

"The starting point for any data preparation project is to locate the data. This is sometimes easier said than done!"

--Dorian Pyle. Data preparation for data mining.

MISSING DATA

- **We learned earlier that MVs can cause problems in our analysis**
  - Our modelling tools might not work on samples that contain MVs
  - We might get biased results and do wrong conclusions if we ignore the problem
- **What can we do to fix the problem?**

"The starting point for any data preparation project is to locate the data. This is sometimes easier said than done!"

--Dorian Pyle. Data preparation for data mining.

MISSING DATA

- **What can we do to fix the problem?**

- **Do not let it happen in the first place!**
- **Recollect the data**
  - Might be impossible or expensive

MISSING DATA

– **What can we do to fix the problem?**

– **Delete the missing values or incomplete cases (i.e. pretend there is no problem)**

- If the data is MCAR, this does not bias the results but does decrease the statistical power if lots of data need to be discarded from the analysis
- But data is not often MCAR
  - Can severely bias estimates of means, regression coefficients and correlations
- In predictive modelling it is often crucial to be able to handle all test cases, even those with MVs, so it might not be wise to just discard them

– **Deletion is, however, the most common method used!**

Complete cases:

MISSING	NUMBER OF FEATURES 5	10	20	40	80
1 %	95.1 %	90.4 %	81.8 %	66.9 %	44.8 %
2 %	90.4 %	81.7 %	66.8 %	44.6 %	19.9 %
5 %	77.4 %	59.9 %	35.8 %	12.9 %	1.7 %
10 %	59.0 %	34.9 %	12.2 %	1.5 %	
20 %	32.8 %	10.7 %	1.2 %		
40 %	7.8 %	0.6 %			
60 %	1.0 %				

MISSING DATA

– Complete-case analysis / list wise deletion

- Discard any rows containing MVs
- Some data mining algorithms do this by default so be careful and get to know your data first!

– Available-case analysis / pair wise deletion

- Use all the instances of a variable (or a group of variables) that are present
- E.g. mean or covariance of variables

	Var 1	Var 2		Var 1	Var 2
	0,5	2,12		0,5	2,12
	0,46	2,63			2,63
	0,43	2,57		0,43	
	0,52	2,03		0,52	2,03
	0,55	1,76			1,76
	0,57	1,97		0,57	1,97
	0,49	1,99		0,49	1,99
Mean	0,502857	2,152857		0,502	2,083333
Var	0,002049	0,090306		0,002056	0,071522
Covar	-0,01165			-0,0009	

E.g. Use this for var2 mean and var calculation but not for covar calculation

"The starting point for any data preparation project is to locate the data. This is sometimes easier said than done!"

--Dorian Pyle. Data preparation for data mining.

MISSING DATA

- **What can we do to fix the problem?**

- **Imputation**
  - Replace the missing values by some value
  - Analyse the now complete data

- **What to impute?**

MISSING DATA

– What to impute?

– Hot deck and cold deck imputation

- If missing values are rare, a replacement can be randomly drawn from a cluster of similar data
  - From the same data set (hot deck)
  - From another similar data set (cold deck)

– Mean imputation

- Widely used
- Replace missing values with mean (numerical) or mode (nominal)
- Global
  - Use all available instances as a reference
- Stratified
  - Use instances of the same class as reference
  - Test data in a predictive modelling task? Impossible

MISSING DATA

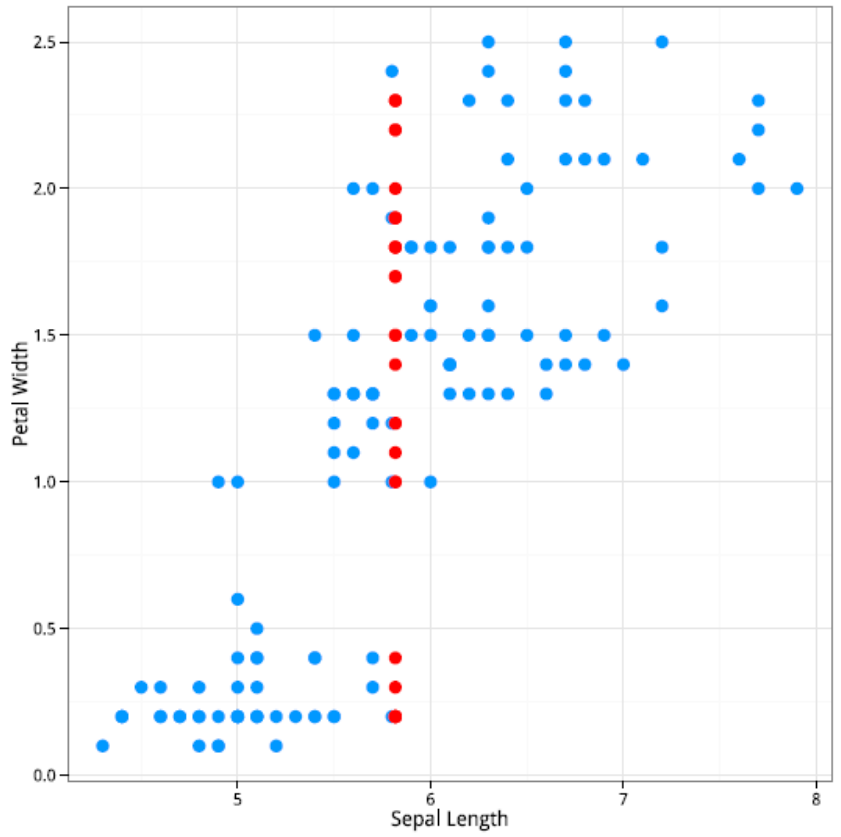
- What to impute?
- Mean imputation

- Makes it possible to analyse the data but:

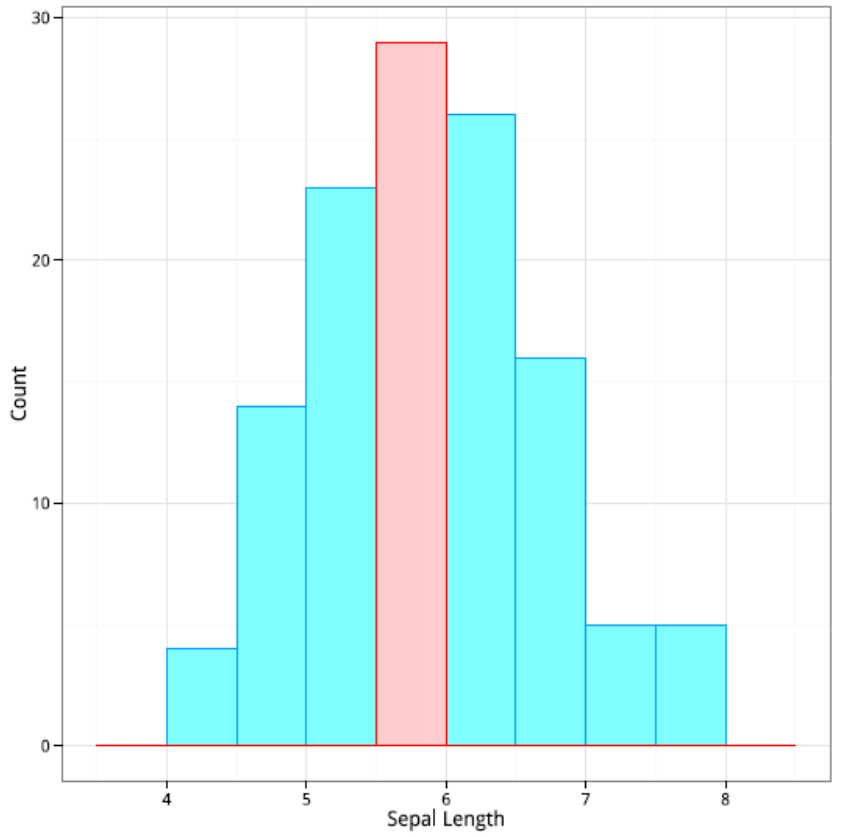
- Completely ignores the mechanism that yield the data values
- Imputed values pile up at the mean, variability is reduced
- Imputed values are constant and cannot therefore correlate with other variables
- Estimates are biased (except the mean itself under MCAR)

MISSING DATA

- What to impute?
- Mean imputation: example

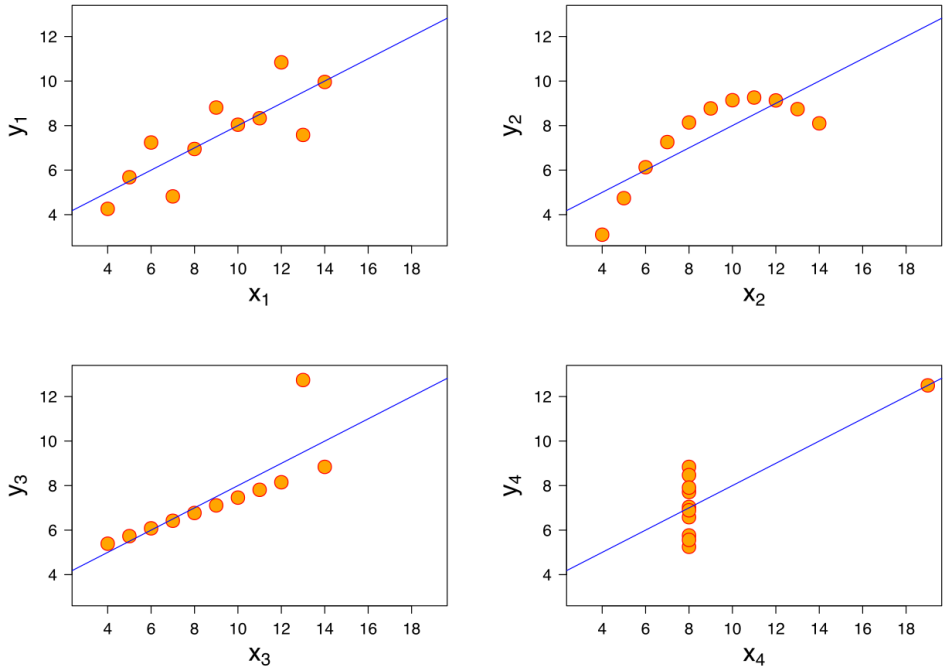


(a) Scatter plot of the features.



(b) Histogram of Sepal Length.

Do not construct your regression imputation model blindly



MISSING DATA

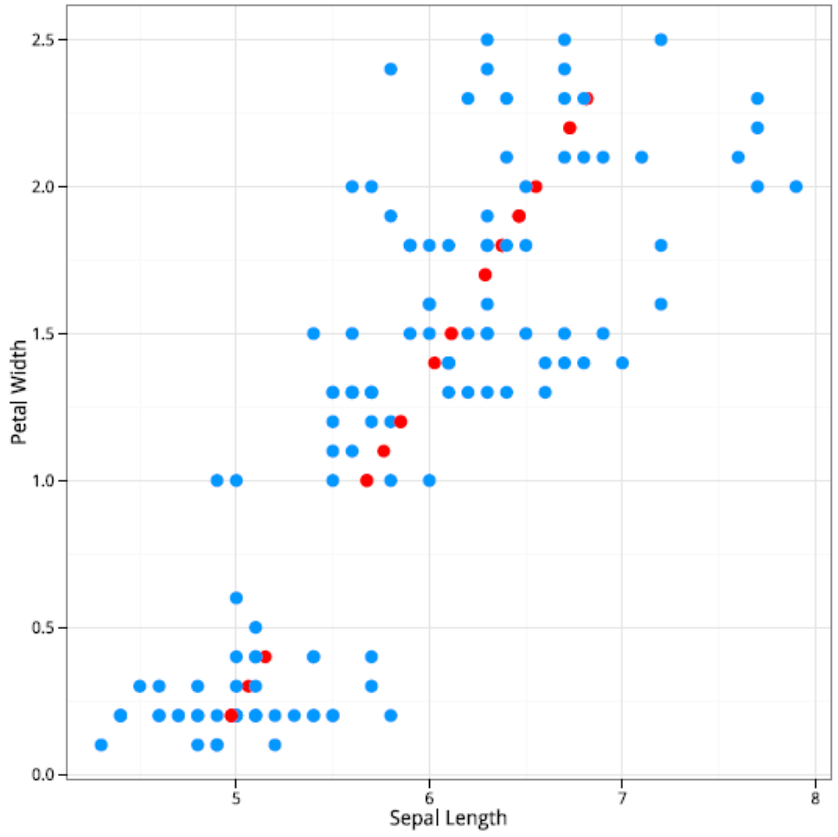
– What to impute?

– Regression imputation

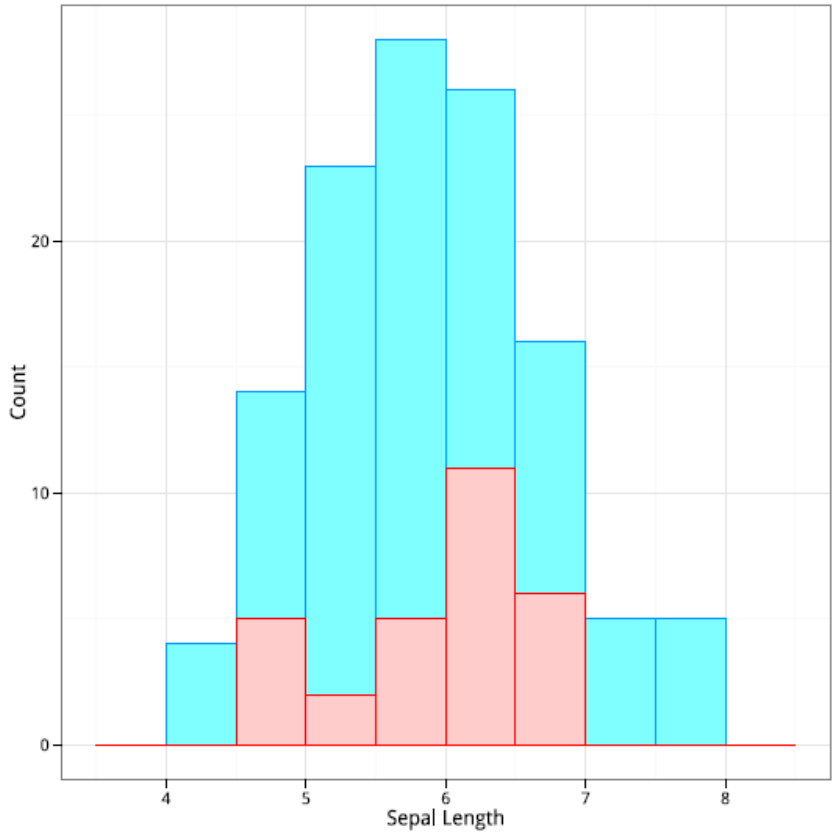
- Use a regression equation to predict the missing values
- Lack of variability because the imputed values fall directly on a regression surface
- Estimates of variation and association strength are biased
  - Variability of the imputed data is underestimated
  - $R^2 = 1$ for the imputed values
- Mean estimates are unbiased under MCAR
- Regression weights of the imputation model are unbiased if the explanatory variables are complete
- Regression weights are unbiased under MAR if the factors that influence the missingness are part of the regression model (i.e. our imputation model is correct)
- Note: if the imputation model yielded perfect prediction for the MVs, there was not really anything missing in the first place!

MISSING DATA

- What to impute?
- Regression



(a) Scatter plot of the features.



(b) Histogram of Sepal Length.

MISSING DATA

- **What to impute?**
- **Stochastic regression imputation**
  - Adding noise to regression imputation restores the lost variability in the imputed data points
  - A regression model is first fit to the complete data, residual variance is estimated, and finally the imputed values are drawn according to these parameters
  - Underestimates standard errors: why?
  - Preserves the relationship and correlation of the variables but it does not deal with the inherent uncertainty of the imputed values
    - These are NOT the real values that were lost even though they might look like they are
- **Regression imputation can lead to impossible values, e.g.:**
  - Male/female: pregnant yes/no?
  - Negative ozone concentration etc.

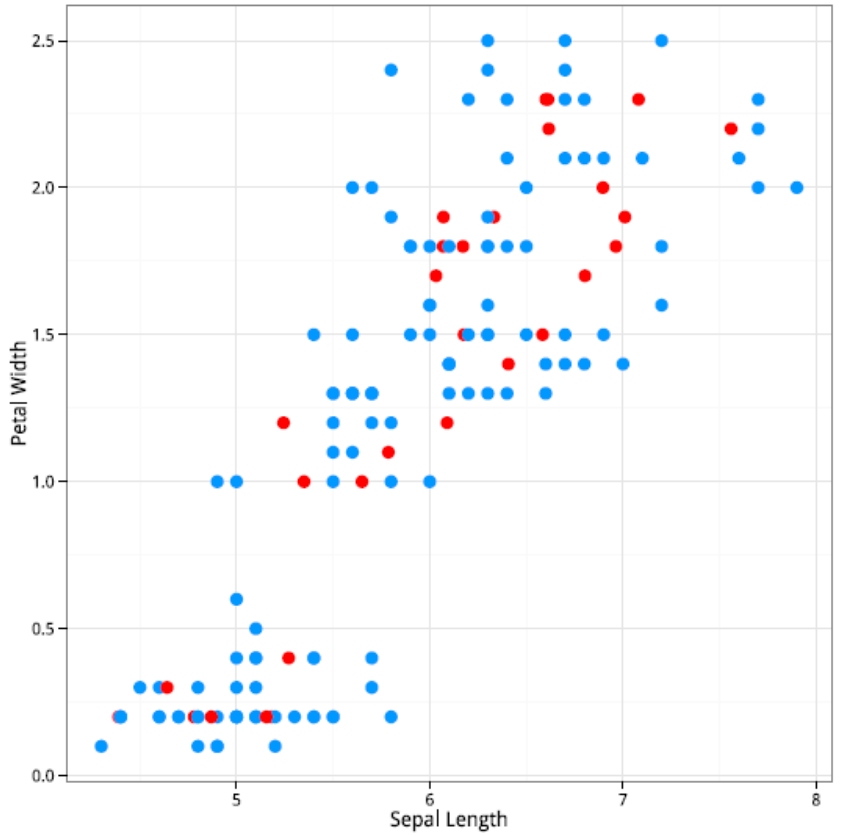
Länsnen media: Raskaana oleva mies kertoi puolisolleen viestillä odottavansa vauvaa – ”Petri tuli ihan äimistyneenä kotiin”

Suomessa on tiettävästi ensimmäistä kertaa tilanne, jossa mies on raskaana. Kuvituskuva. (KUVA: Colourbox)

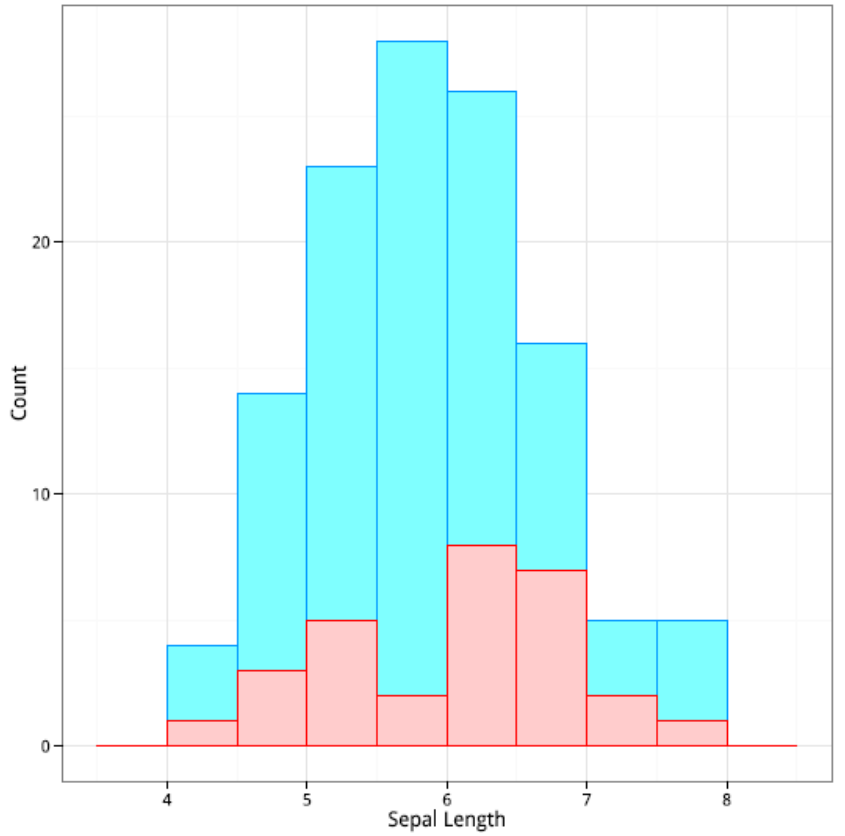
A news item about a (transgender) male being pregnant

MISSING DATA

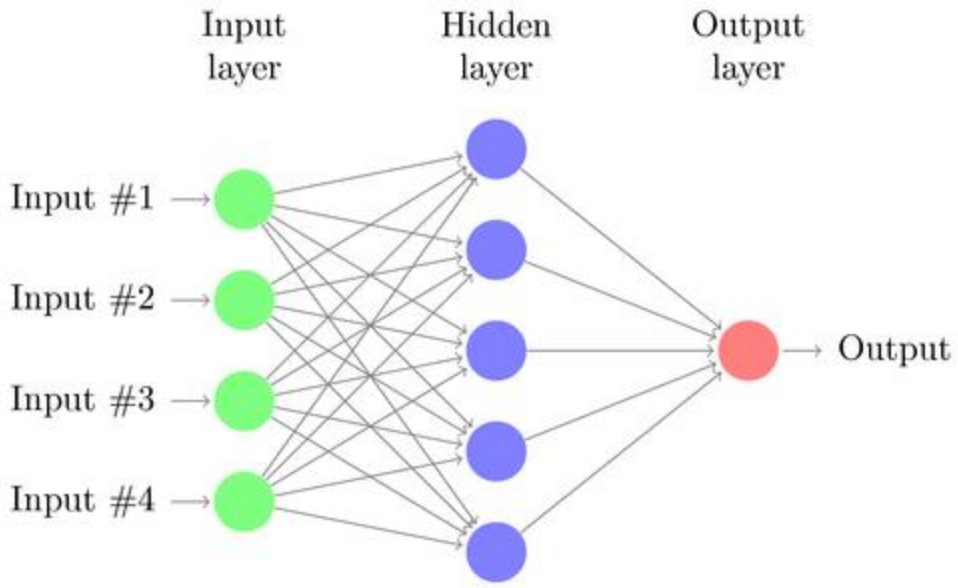
- What to impute?
- Stochastic regression



(a) Scatter plot of the features.



(b) Histogram of Sepal Length.



MISSING DATA

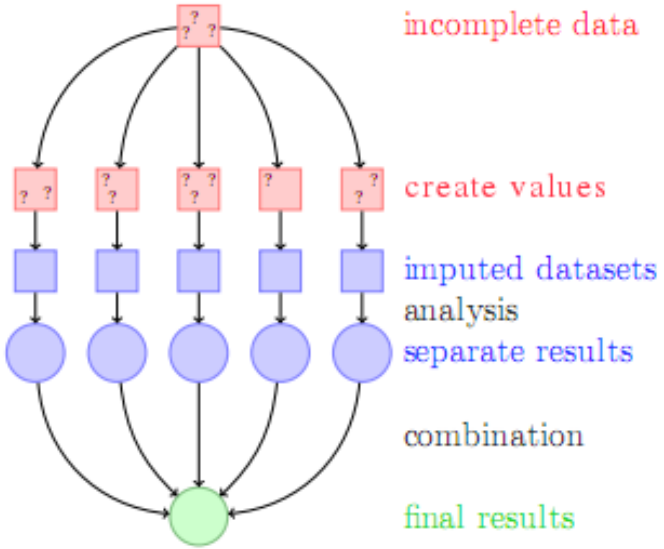
- **What to impute?**
- **Machine learning based methods**

- **No need to search for the underlying distribution of the data**
- **MAR assumption still holds**
- **Examples**

- k-Nearest Neighbour (kNN)
- K-means clustering
- Support Vector Machines
- Local Least Squares
- Neural networks

- **Some methods can handle MVs natively**

- C4.5, Naïve Bayes, etc.



MISSING DATA

- **What to impute?**
- **Multiple imputation**
- **M > 1 complete data sets drawn from a distribution specifically modelled for each missing entry**
  - The imputed datasets are identical for the observed data entries, but differ in the imputed values
  - The magnitude of these difference reflects the uncertainty about what value to impute
- **Each of the M complete data sets are analysed independently and combined into the final result**
  - Statistical analysis
  - Even classification is possible

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m}\hat{Q}_l$$

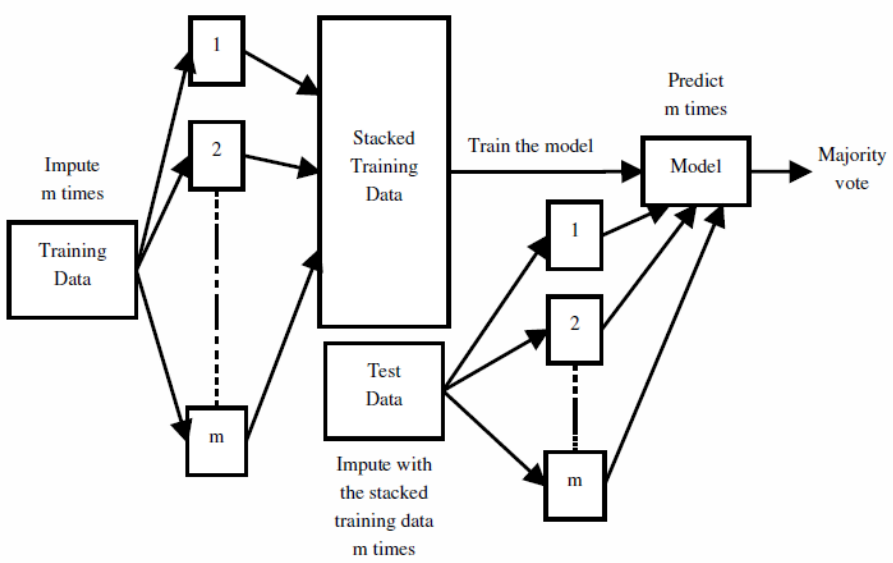
$$\bar{U} = \frac{1}{m}\sum_{l=1}^{m}\bar{U}_l$$

$$B = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{Q}_l - \bar{Q})(\hat{Q}_l - \bar{Q})'$$

$$T = \bar{U} + B + \frac{B}{m} = \bar{U} + \left(1 + \frac{1}{m}\right)B$$

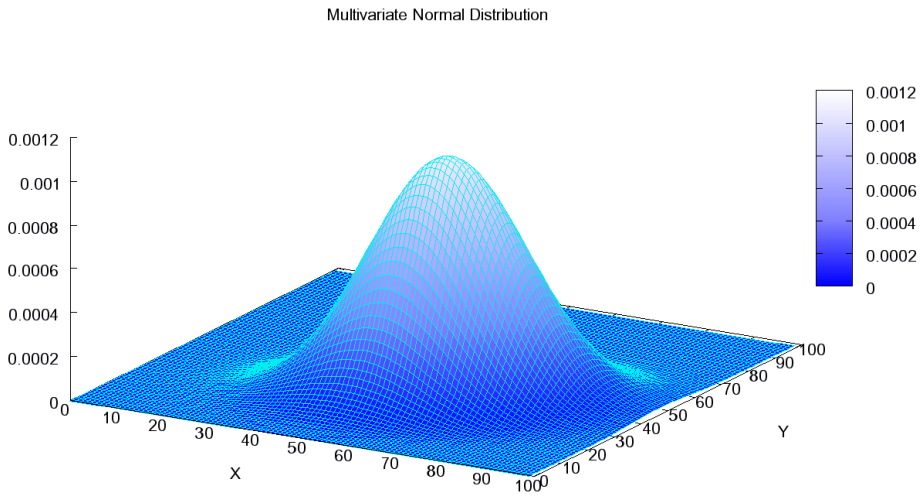
MISSING DATA

- **What to impute?**
- **Multiple imputation (MI)**
- **Results are pooled into one estimate $\bar{Q}$**
- **Variance *T* of the estimate is a combination of:**
  - Conventional sampling variance $\bar{U}$ (within-imputation)
  - Variance caused by the missing data *B* (between imputation)
- **Under the appropriate conditions, the pooled estimates are unbiased and have the correct statistical properties**
- **MI provides a mechanism for dealing with the inherent uncertainty of the imputations themselves**



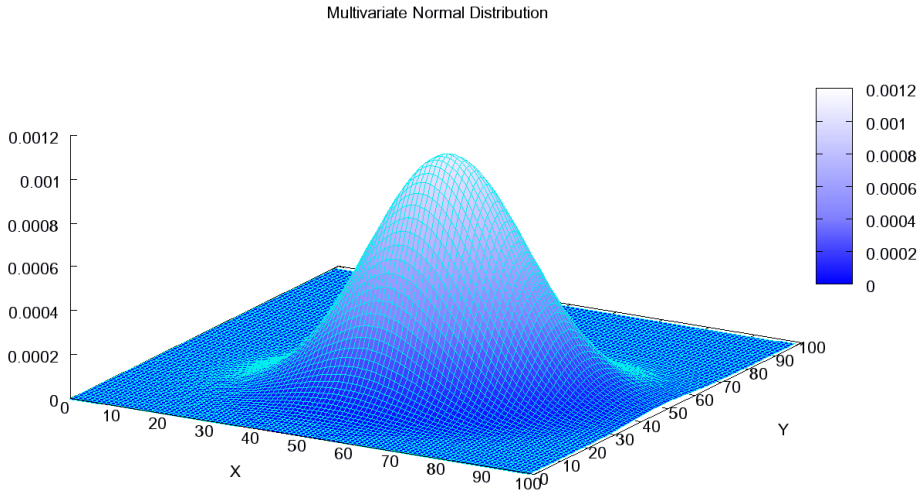
MISSING DATA

- What to impute?
- Multiple imputation
- It is possible to use MI in classification
- M predictions, combine by majority vote or average the prediction scores
- Uncertainty in the predictions can be dealt in a similar way as in the case of statistical analysis



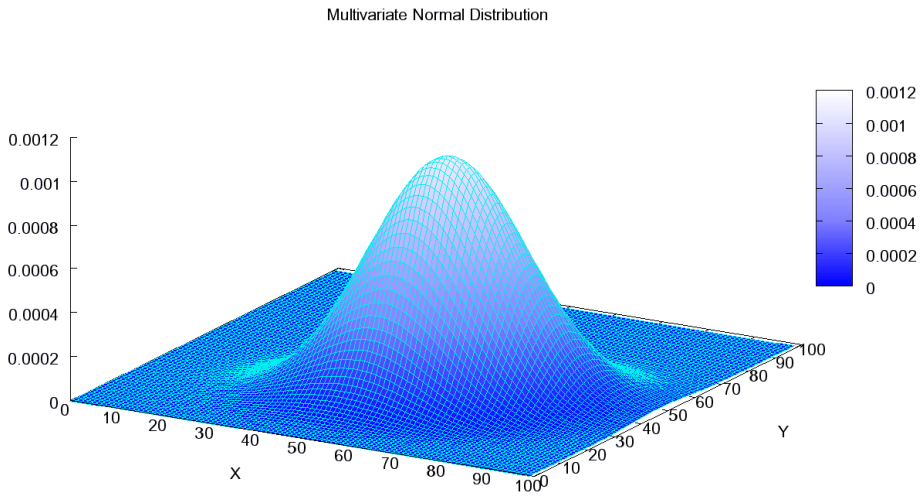
MISSING DATA

- **Maximum likelihood estimation**
- **Find the population parameter (e.g. mean) values that are most consistent with the raw data**
- **A likelihood function is used to quantify how well the data fits the estimated parameters**
  - Likelihoods are very small numbers
  - Taking the natural log makes the math more tractable
  - Log likelihood quantifies the relative probability of scores but on a logarithmic scale
- **Multivariate normal distribution is used as the starting point with continuous variables**



MISSING DATA

- **Maximum likelihood estimation**
- **The log likelihood for an entire sample is the sum of the individual log likelihoods**
- $logL = \sum_{i=1}^{N} logL_i$
- **ML is used to find the parameter values that best fit the sample**
  - i.e. has the highest log likelihood



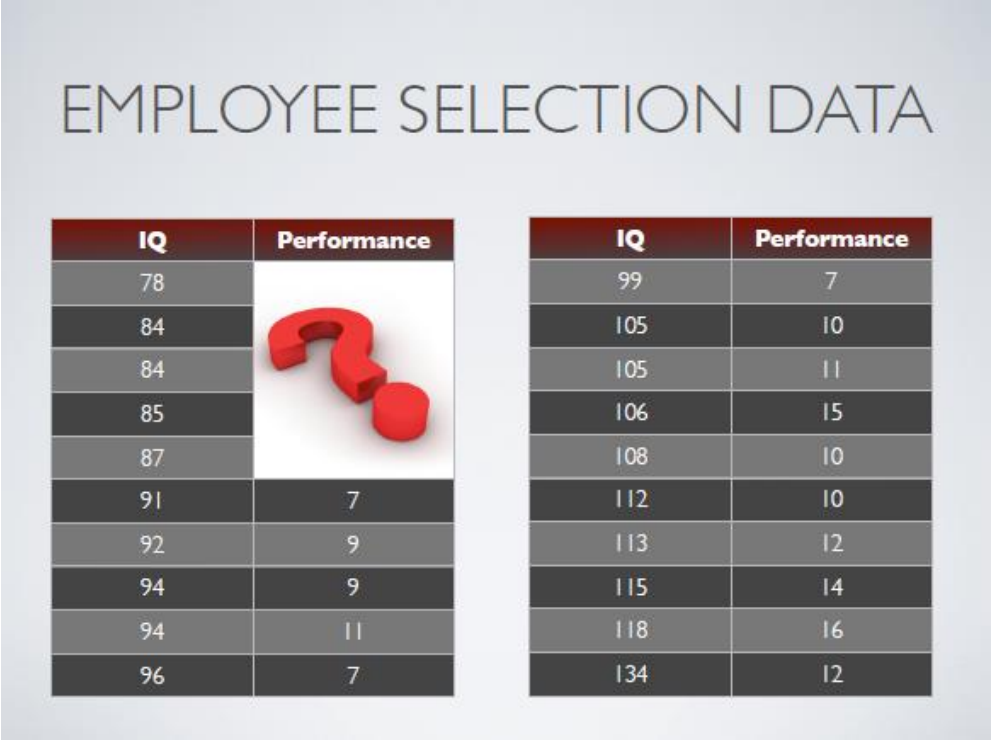
MISSING DATA

- Maximum likelihood estimation

- When some data values are missing, all available data is used to estimate the parameters
- Some samples contribute more information than others
- Missing values are not filled explicitly
- Information from the observed data is used to estimate the parameters of the incomplete cases
- This improves the accuracy of the parameter estimates for the whole data

MISSING DATA

- **Maximum likelihood estimation**



- **This example is from Texas A&M Summer Statistics Workshop held in June, 2012**
- **Original author**

Craig K. Enders
Arizona State University
Department of Psychology

MISSING DATA

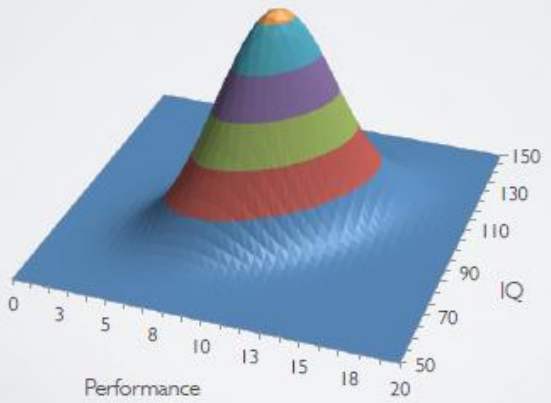
– Maximum likelihood estimation

ESTIMATION EXAMPLE

- The true job performance mean is $\mu = 10.35$
- Deleting the incomplete cases produced $\mu = 10.67$
- ML improves the estimate of the job performance mean by including the IQ scores for the five incomplete cases
- The normal distribution is the key to understanding how ML missing data handling works

BIVARIATE NORMAL DISTRIBUTION

- ML assumes that IQ and performance ratings are normally distributed



MISSING DATA

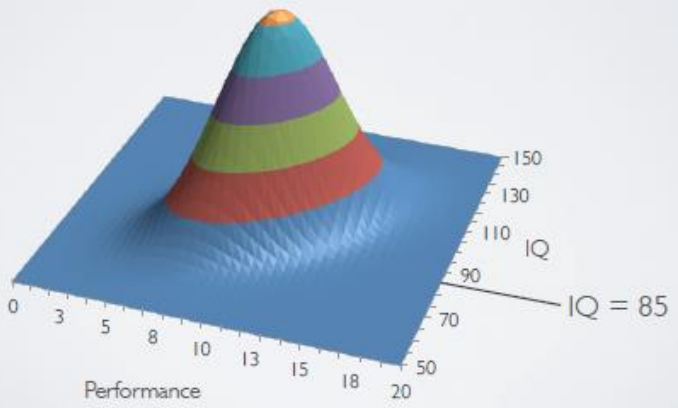
– Maximum likelihood estimation

IMPLICATIONS OF NORMALITY

- The IQ scores provide information about the missing performance ratings
- For a given IQ value, some job performance ratings are more plausible than others
- The normal distribution effectively constrains the plausible range of score values for the missing data

EXAMPLE 1: IQ = 85, PERFORMANCE = ?

- Consider the incomplete case with IQ = 85

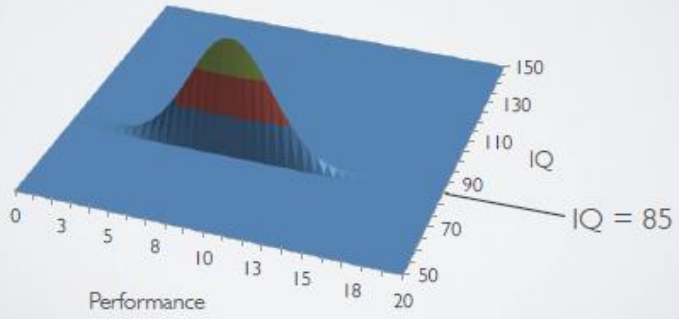


MISSING DATA

– **Maximum likelihood estimation**

NORMAL CURVE SLICE IQ = 85

• Job performance distribution, conditional on IQ = 85



WHAT DO WE KNOW ABOUT THE MISSING VALUE?

• Given that IQ = 85, the most likely value of the missing performance rating is ≈ 9

MISSING DATA

– Maximum likelihood estimation

WHAT HAPPENS TO THE ML ESTIMATE OF μ ?

- The 15 complete cases produced an average of μ = 10.67
- A case with an IQ = 85 would likely have a performance rating of approximately 9
- Based on this information, ML adjusts the job performance mean downward to account for the plausible (but missing) performance rating
- This adjustment is based solely on the observed IQ value!

ESTIMATION RESULTS

- Including the five incomplete cases produced parameter values that better approximate the complete-data estimates

Method	IQ	Performance
Complete	100.00	10.35
Deletion	105.47	10.67
ML	100.00	10.28

- ML borrowed information from the observed IQ scores to derive the job performance parameters

MISSING DATA

- **Maximum likelihood estimation**

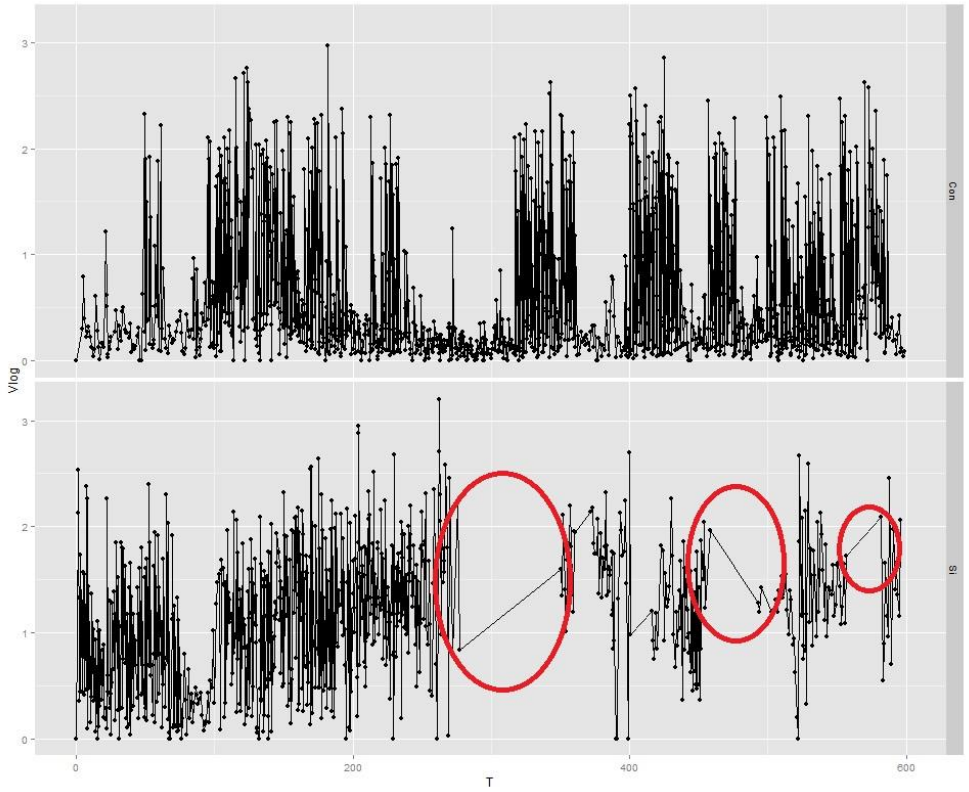
- **Expectation Maximization (EM) algorithm is often used to find the ML estimates**
  - Expectation step (E step)
    - Calculate the expected value of the log likelihood based on current estimation of the parameters
  - Maximization step (M step)
    - Find parameter values that maximize the expected value of the log likelihood
  - Iterate until the algorithm converges
- **Others: sweep operator, Newton-Raphson, Bayesian simulation**
- **These methods are knowns as full information maximum likelihood (FIML) methods**

MISSING DATA

- **Maximum likelihood estimation**

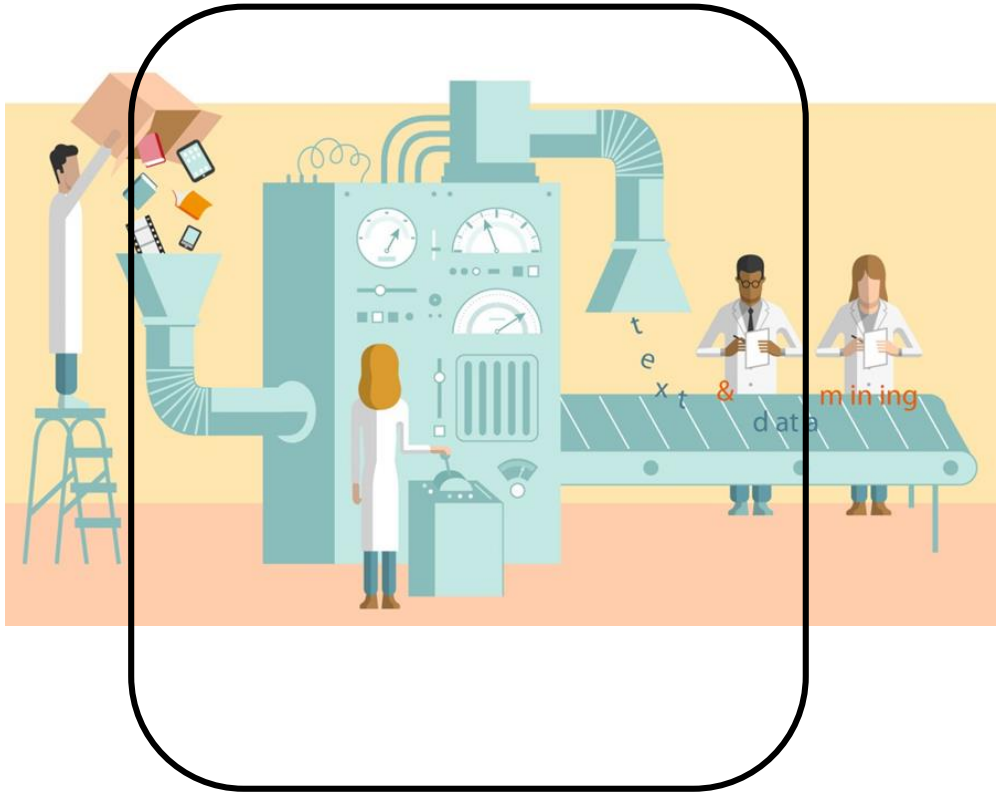
- **Expectation Maximization (EM) algorithm**
  - Well suited only for numerical data sets

- **ML methods tend to underestimate standard errors**
  - Multiple imputation is better in this regard



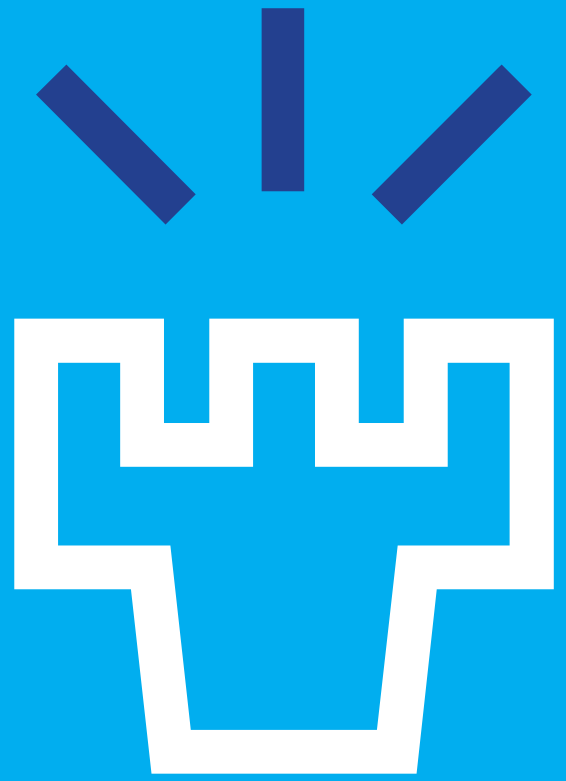
MISSING DATA

- **Time series / longitudinal data**
- **Deletion messes with the timeline / autocorrelation**
- **Last observation carried forward (LOCF)**
  - Often used in clinical trials
  - Can yield biased estimates even under MCAR
  - Analysis following imputation should distinguish between real and imputed data
- **Another variation is baseline observation carried forward (BOCF)**
- **Multiple imputation is possible with longitudinal data, too**



CONCLUSIONS

- Real world data sets contain problems like missing values
- Be aware of these
- There are no hard rules on how to handle MVs, it always depends
- Get to know your data and find out why there are missing values, this will lead you into right direction on what to do
- Multiple imputation makes it possible to deal with the inherent uncertainty regarding missing data imputations



UNIVERSITY
OF OULU