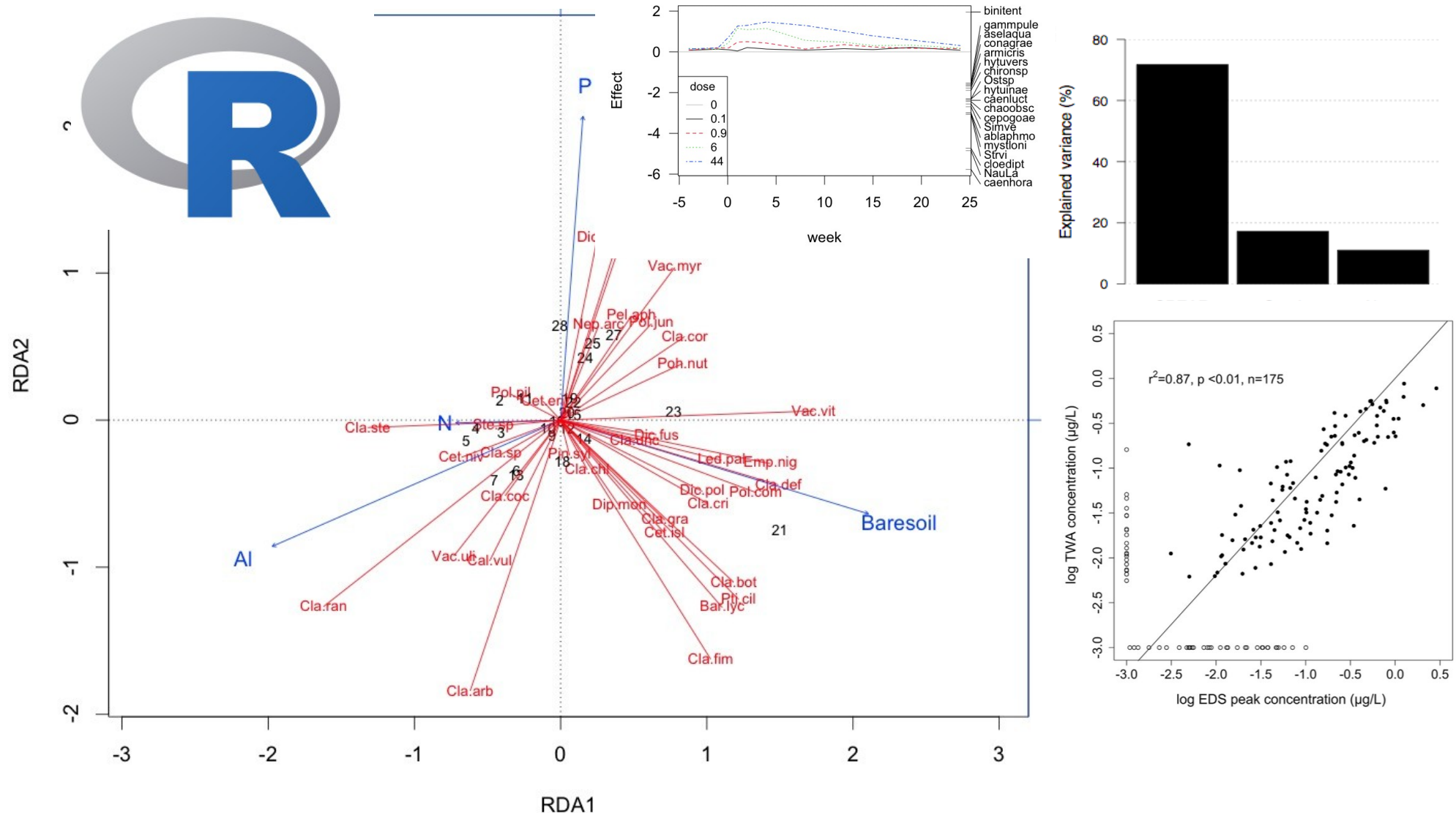


Tools for complex data analysis

University of Koblenz-Landau 2018/19



Ralf B. Schäfer


Short introduction

- Professor for Quantitative Landscape Ecology
- Current teaching: Data analysis (M.Sc.); GIS (B.Sc./M.Sc.); Environmental Modelling (B.Sc./M.Sc.); Environmental Philosophy (B.Sc.)
- Current research projects related to:
 - Community ecology of freshwater invertebrates and microorganisms
 - Response of freshwater ecosystems to different (anthropogenic) stressors (e.g. pollution)
 - Trophic linkages between aquatic & terrestrial systems
- Primarily field studies/experiments and data analyses/modelling
- Course assistants (Phd students): Stefan Kunz & Le Trong Dieu Hien (“Vicky”)



Using your own notebook

- We encourage use of your own notebook!
- install [R](#) and subsequently [RStudio](#)
- Run “0_install_packgs.R”, provided on github
- for installation of additional packages run `install.packages(“package to be installed”)`



The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It runs on a variety of UNIX platforms, Windows and MacOS. To [download R](#), please see the [mirror](#).

If you have questions about R like how to download and install the software, please read our [answers to frequently asked questions](#) before you search.

News

- [R version 3.2.3 \(Wooden Christmas-Tree\) prerelease versions](#) will be available from 11-30. Final release is scheduled for Thursday 2015-12-10.
- [R version 3.2.2 \(Fire Safety\)](#) has been released on 2015-08-14.
- [The R Journal Volume 7/1](#) is available.
- [R version 3.1.3 \(Smooth Sidewalk\)](#) has been released on 2015-03-05.
- [useR! 2015](#), took place at the University of Aalborg, Denmark, June 30 - July 1.
- [useR! 2014](#), took place at the University of California, Los Angeles, US.



The screenshot shows the RStudio website. The header includes the RStudio logo and navigation links: Home, RStudio IDE, Shiny, Training, Projects, About, and Blog. The main content area features a large blue R logo and the text "Welcome to RStudio" followed by "Software, education, and services for the R community". Below this, there are three columns of information:

- Powerful IDE for R**: RStudio IDE is a powerful and productive user interface for R. It's free and open source, and works great on Windows, Mac, and Linux. Buttons: [Download now](#), [Learn more](#).
- R training and education**: We've got hands-on courses for beginners and even R experts. Customize an on-site training or enroll in one of our public workshops. Buttons: [Request on-site](#), [View courses](#).
- Open source R packages**: Our developers and expert trainers are the authors of several popular R packages, including ggplot2, plyr, lubridate, and others. Button: [See projects](#).

At the bottom, the footer reads: © 2013 RStudio, Inc. [Follow @rstudioapp](#) | [Trademark](#) | [DMCA](#) | [Careers](#)

Course objectives: Learning targets

- Design a study and select corresponding tools for subsequent data analysis
- Select and apply techniques of data analysis for a research goal
- Classify, explain/interpret and critically evaluate different approaches to data analysis
- Programming simple to moderately complex data analysis tasks in R

Session overview

1.Framework for data analysis and research goals

2.Data exploration

3.Statistical modelling: Overview of techniques

Learning targets

- Explain the data analysis cycle
- Classify research goals
- Understand the role of data exploration and interpret and apply related tools
- Explain approaches to statistical modelling

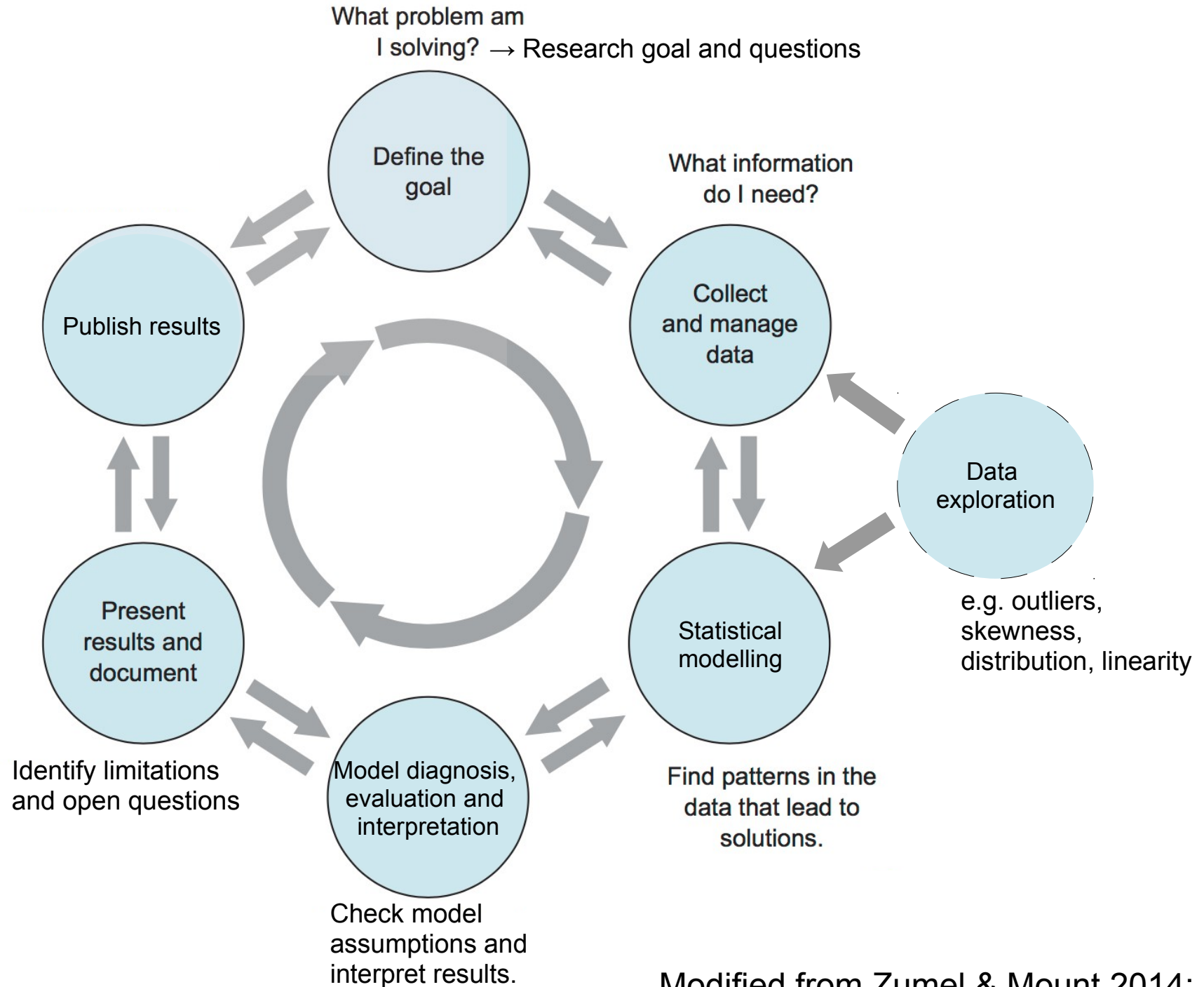
Learning targets and study questions

- Explain the data analysis cycle
 - Describe the steps of the data analysis cycle.
- Classify research goals
 - Distinguish different research goals and give an example for each.
 - Name a statistical technique for each research goal.
- Understand the role of data exploration and interpret and apply related tools
 - Describe the three roles of exploration.
 - Explain four tools for exploration including their domain of application in checking model assumptions.
 - What are quantiles? How are quantiles calculated?

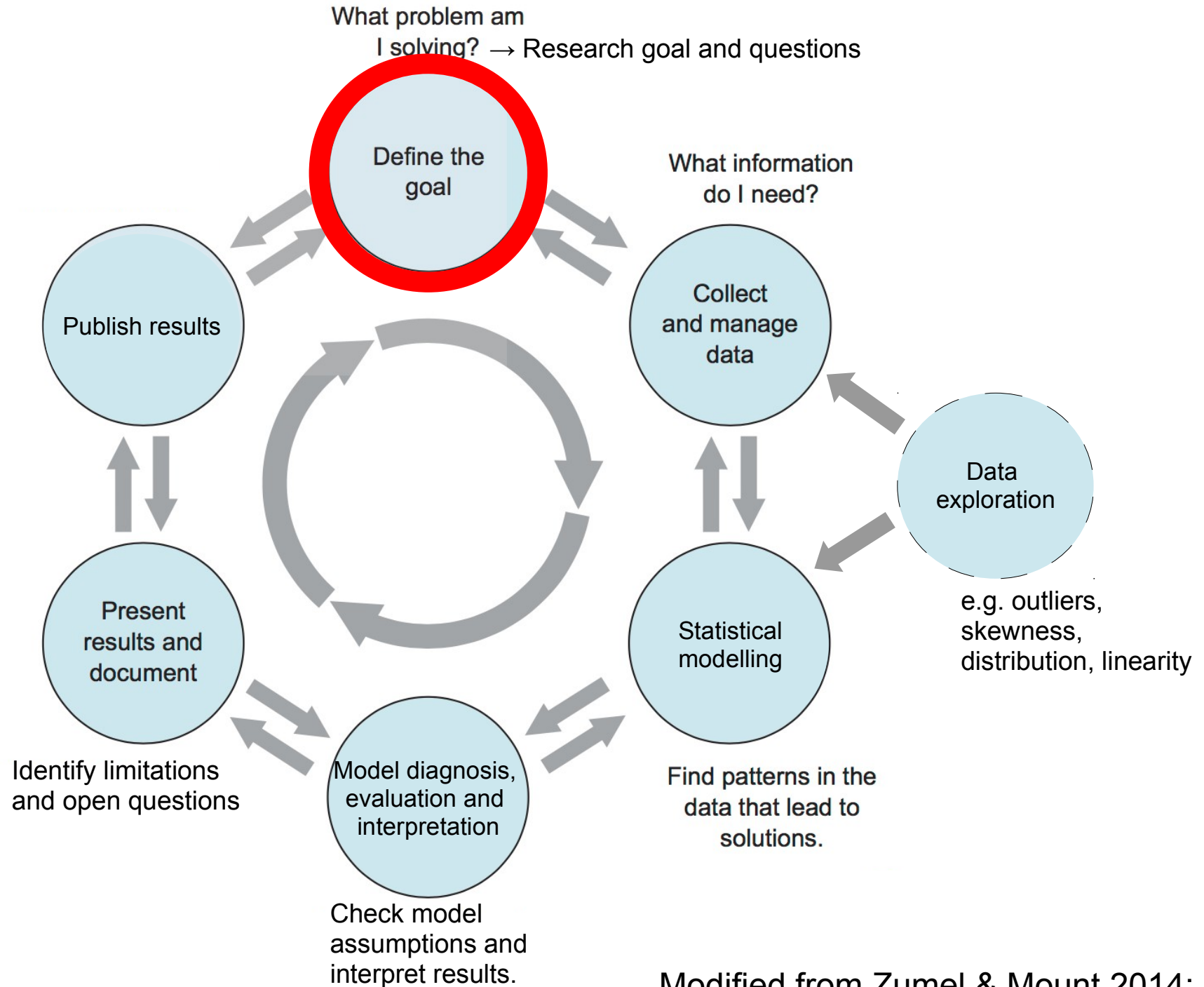
Learning targets and study questions

- Explain approaches to statistical modelling
 - Discuss the core differences of the two main approaches to statistical modelling (e.g. method, research goals, validation)

Data analysis cycle



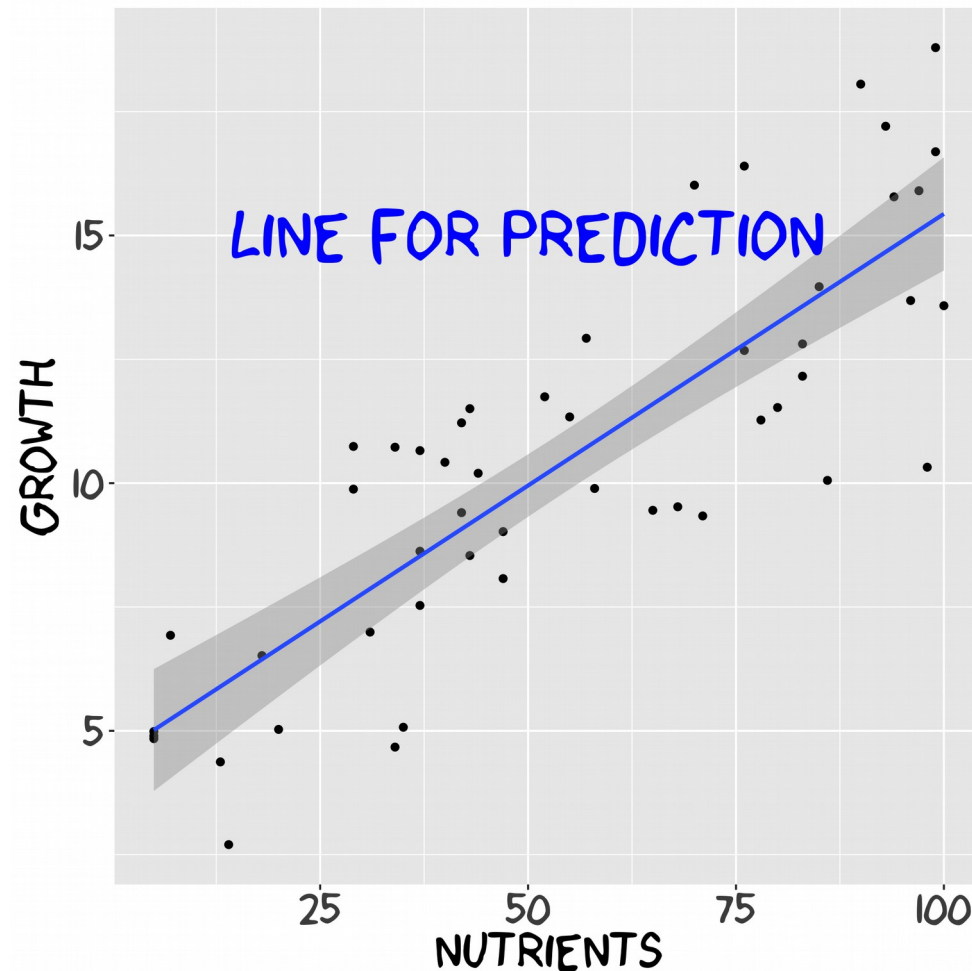
Data analysis cycle



Research goals with examples

1. Prediction

Example: Establish linear relationship between mean plant growth and nutrient concentrations from observations that allows for prediction of mean plant growth for non-observed nutrient concentrations



Research goals with examples

2. (Parameter) estimation

Example: Estimate body mass of koalas under poor habitat conditions



https://commons.wikimedia.org/wiki/File:Friendly_Female_Koala.JPG

Let the continuous random variable X be the body mass of koalas and the discrete random variable Y be the habitat condition. We define ω as an outcome/event:

$$Y(\omega) = \begin{cases} 0 & = \text{poor habitat} \\ 1 & = \text{good habitat} \end{cases}$$

Aim: Estimate expected value E of body mass in poor habitat:

$$E[X|Y=0] = \mu_{X|Y=0}$$

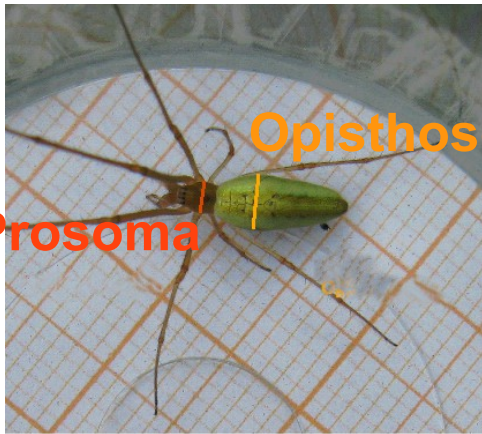
If we cannot sample all koalas, we estimate based on a sample:

$$\hat{\mu}_{X|Y=0} = \bar{x}_{X|Y=0}$$

Research goals with examples

3. Determination of probabilities and assessing hypotheses

Example: Does treatment of a spider with a pesticide at the typical application rate reduce the body size of the spider?



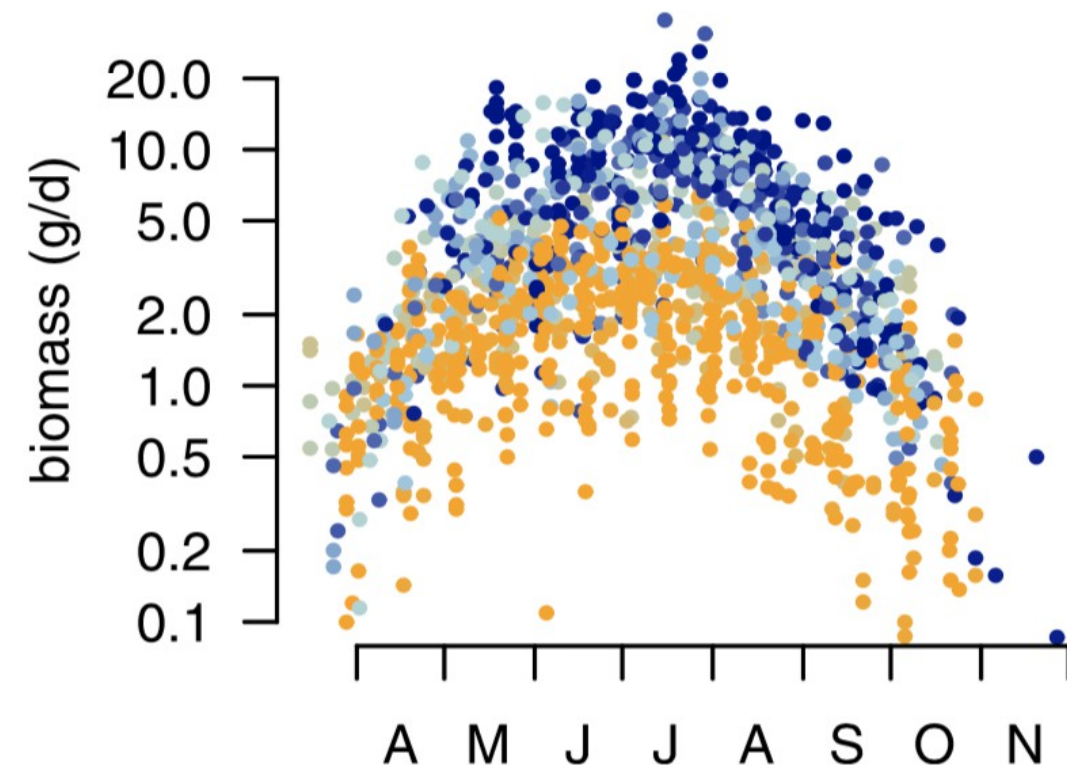
Scientific hypothesis: The pesticide requires the activation of energetically costly detoxication processes. This reduces the energy for growth and consequently the body size, measured as **prosomal** and **opisthosomal** width.

- Two perspectives:
 - How much evidence do the experimental results lend against the hypothesis of no pesticide effect? → Frequentist perspective
 - What is the probability that the pesticide has no effect in the light of the experimental results and prior beliefs or information? → Bayesian perspective

Research goals with examples

4. Explanation

Example: Which variables do best explain the reduction of insect biomass (in Germany)?



Hallmann et al. 2017 *PLOS One* 12: e0185809

Approach: Identification of the most parsimonious model for insect biomass and of the contribution of variables to the explanatory power.

Model 1: Biomass ~ Temp., Precipitation, Frost days ...

Model 2: Biomass ~ Arable land, Forest, Grassland ...

Model 3: Biomass ~ Arable land, Temp., Frost days ...

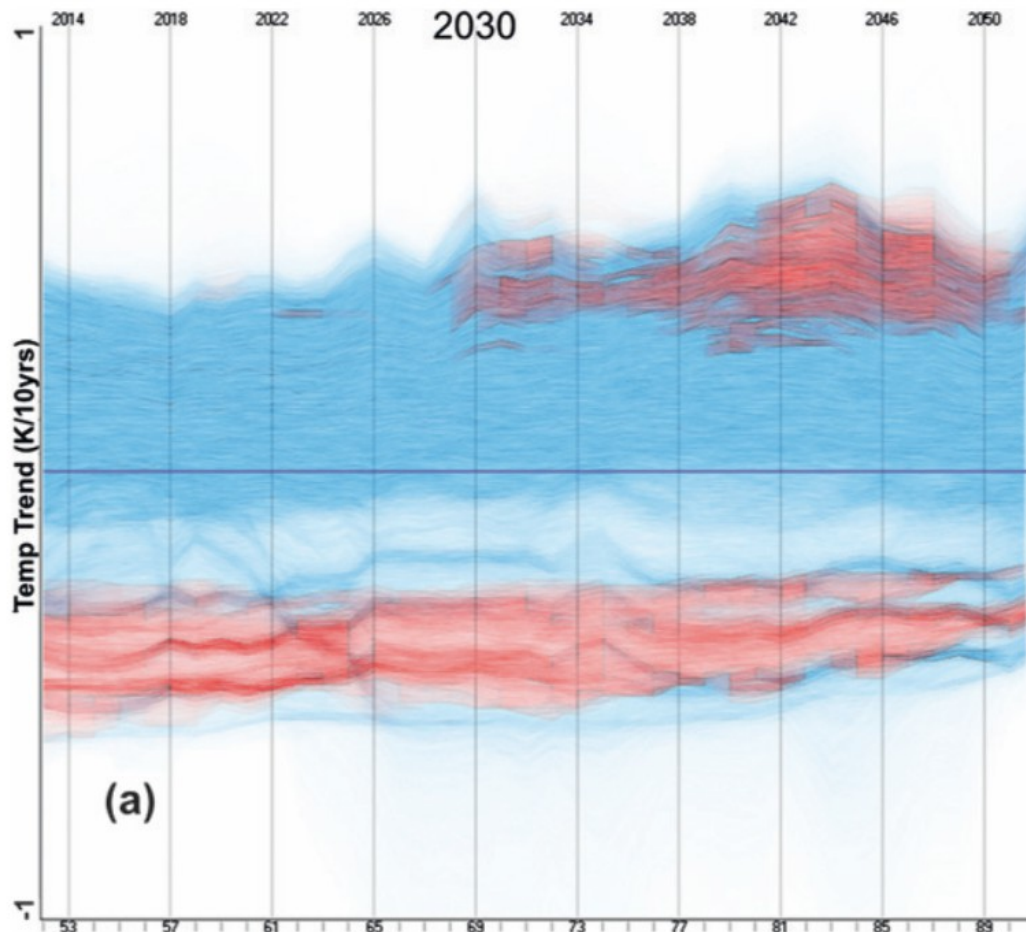
⋮

Model n : Biomass ~ Forest, Precipitation

Research goals with examples

5. Exploration and descriptive statistics

Example: Explore large climate data set to generate new ideas and hypotheses.



Approach: Visual exploration and calculation of descriptive statistics.

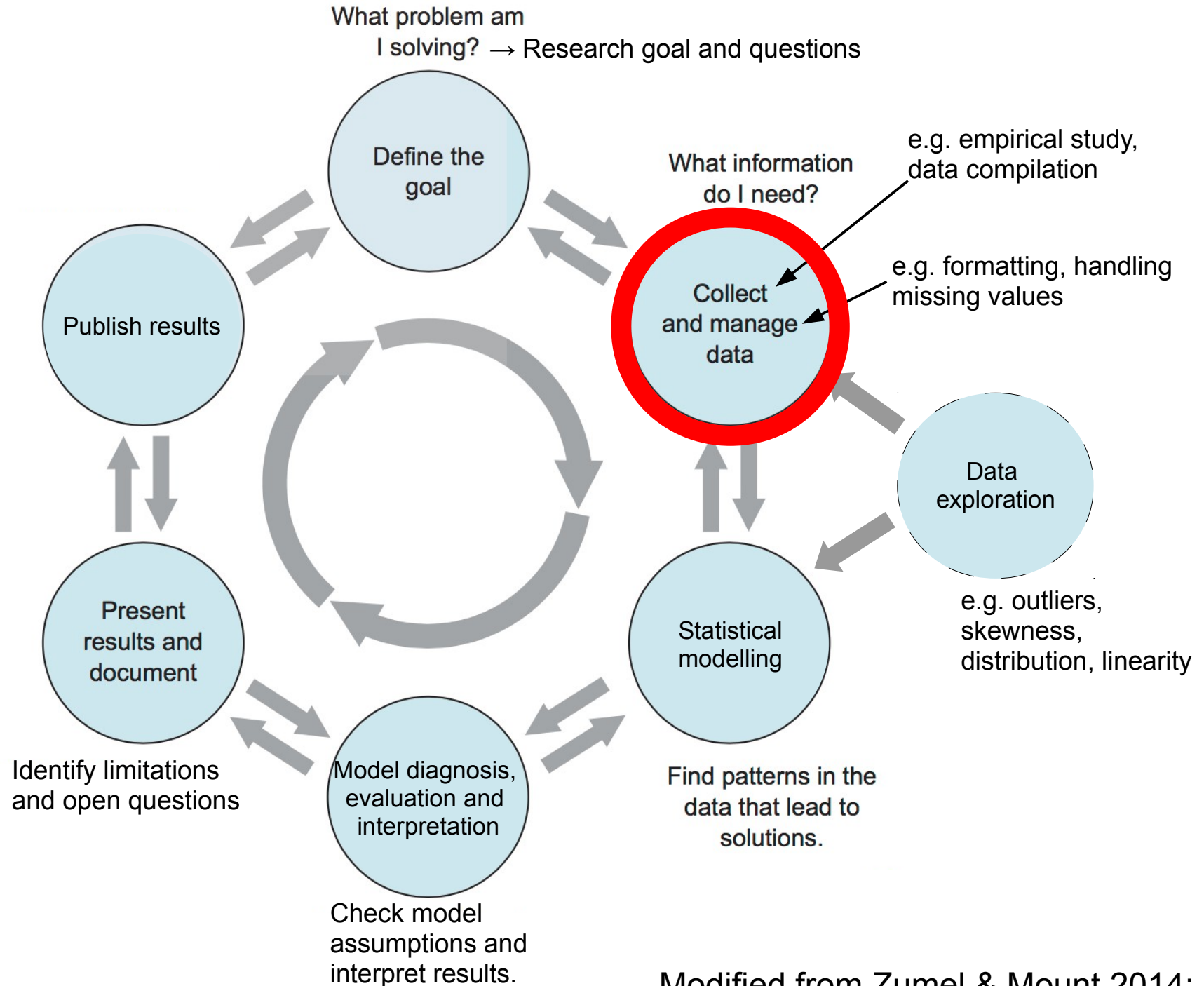
Observations used in exploration to generate hypotheses must not be used in assessing statistical hypotheses!

Overview on research goals

1. Prediction
2. (Parameter) estimation
3. Determination of probabilities and assessing hypotheses
4. Explanation
5. Exploration and descriptive statistics

 **Inform study design and method selection in data analysis**

Data analysis cycle



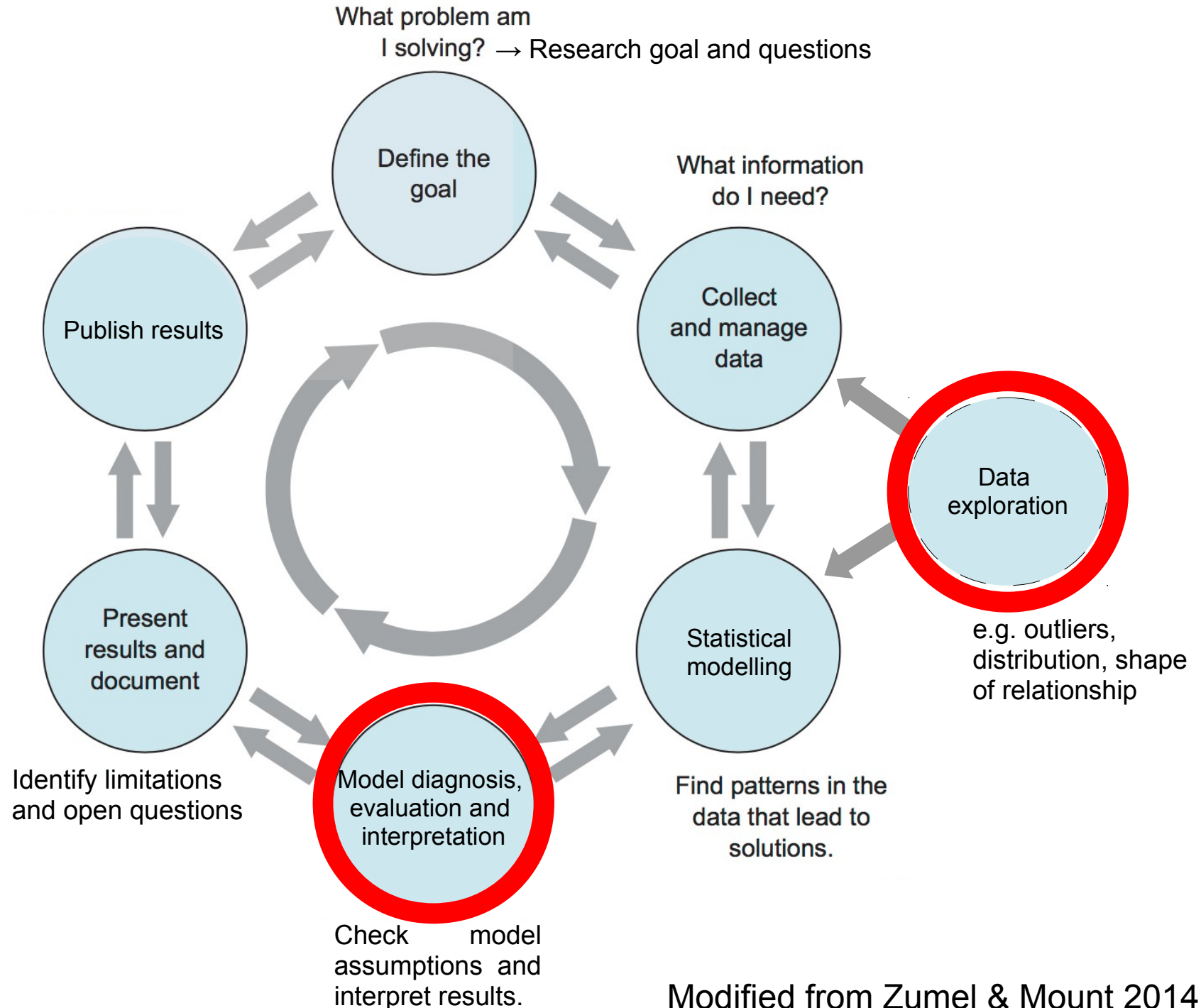
Session overview

1. Framework for data analysis and research goals


2. Data exploration

3. Statistical modelling: Overview of techniques

Data analysis cycle



Exploration and descriptive statistics

- Has at least three roles:
 - Suggest new ideas, understandings or hypotheses
→ stimulate additional analyses or follow-up studies
 - Facilitate model selection, checking of assumptions and identification of errors or outliers in the data
 *GIGA: Garbage in – Garbage out*
 - Provide tools for communicating the data

Tools for exploration: Model selection, assumptions, outliers and errors

Checking for:	Example for tool
Outliers and errors	Boxplot, Cleveland plot
Variance homogeneity	Conditional boxplot, Beanplot
Normal distribution	QQ-plot ¹ , Histogram
Shape of distribution	Histogram
(Double) zeros	Frequency plot or table
Collinearity	Scatterplot
Shape of relation between predictor and response variable	Scatterplot
Interactions	Coplots, Interaction plots
Spatial- or temporal autocorrelation	Variograms

¹QQ-plot = Quantile-Quantile plot

Quantiles

- Essential for several plots (e.g. QQ-plot, boxplot)
- Cut (ordered) data into subsets of equal size/probability

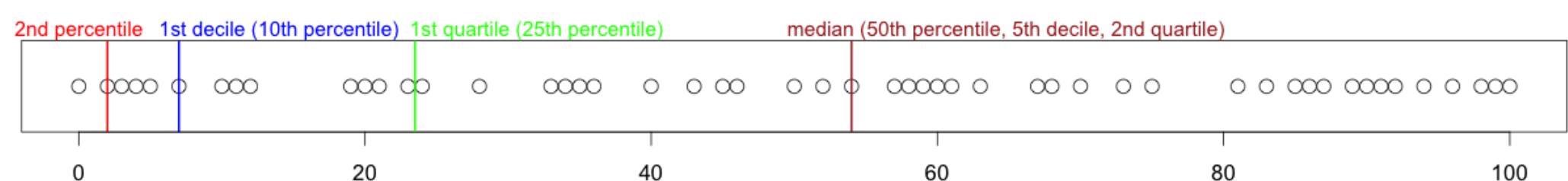
- Definition:

For data $x = \{x_1, x_2, x_3, \dots, x_n\}$, where the values are ordered as: $x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(n)}$ (\rightarrow order statistics), the q -quantiles are values that partition x into q subsets.

- E.g. 2-quantile = median, 4-quantiles = quartiles, 10-quantiles = deciles

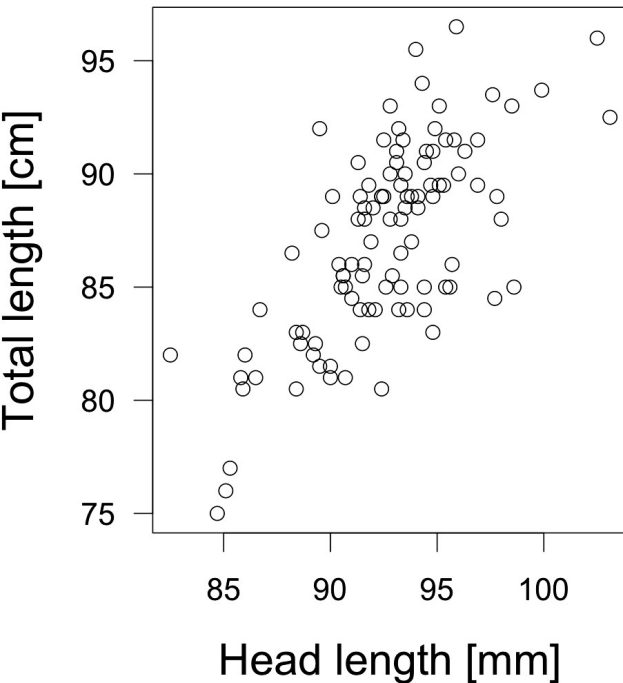
The x_k corresponding to the k -th q -quantile ($k \in \mathbb{N}_0 \wedge 0 < k < q$) cuts with: $P(X < x_k) \leq k/q$ and $P(X \geq x_k) \geq 1 - k/q$

Example: 51 random samples x (without replacement) with $x \in \mathbb{N}_0 \wedge x < 101$



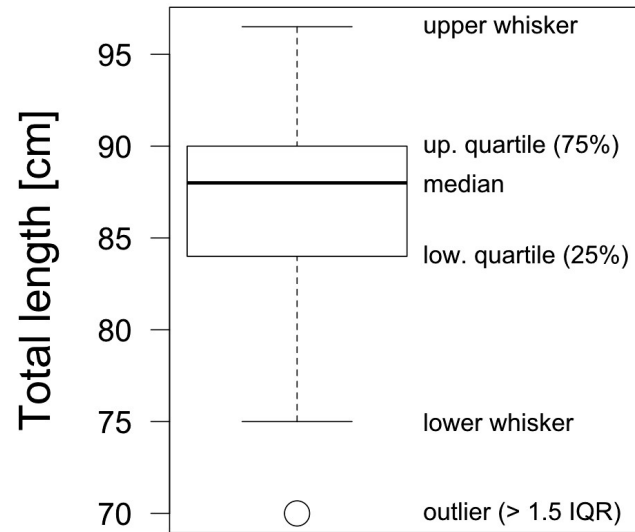
Common tools for exploration

Scatterplot



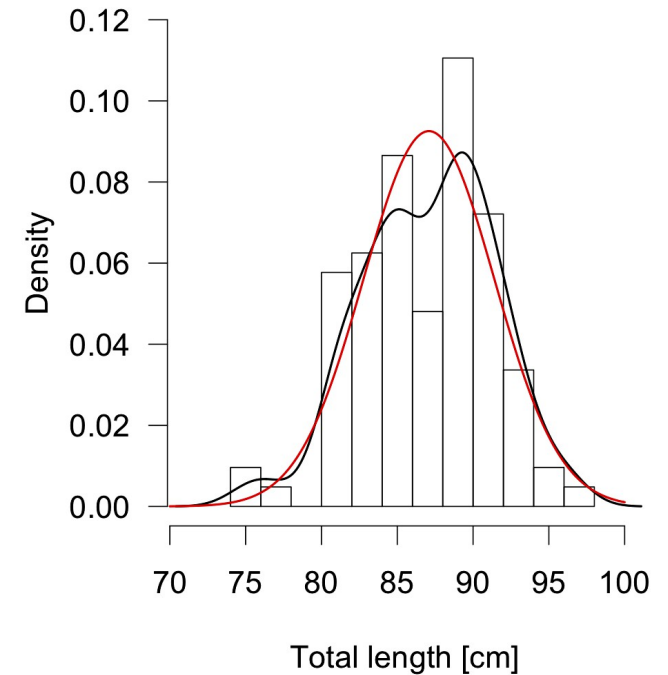
Shape of
relationship?
Collinearity?

Boxplot



Outliers and
errors? Variance
homogeneity?

**Histogram with density curve
and normal distribution**

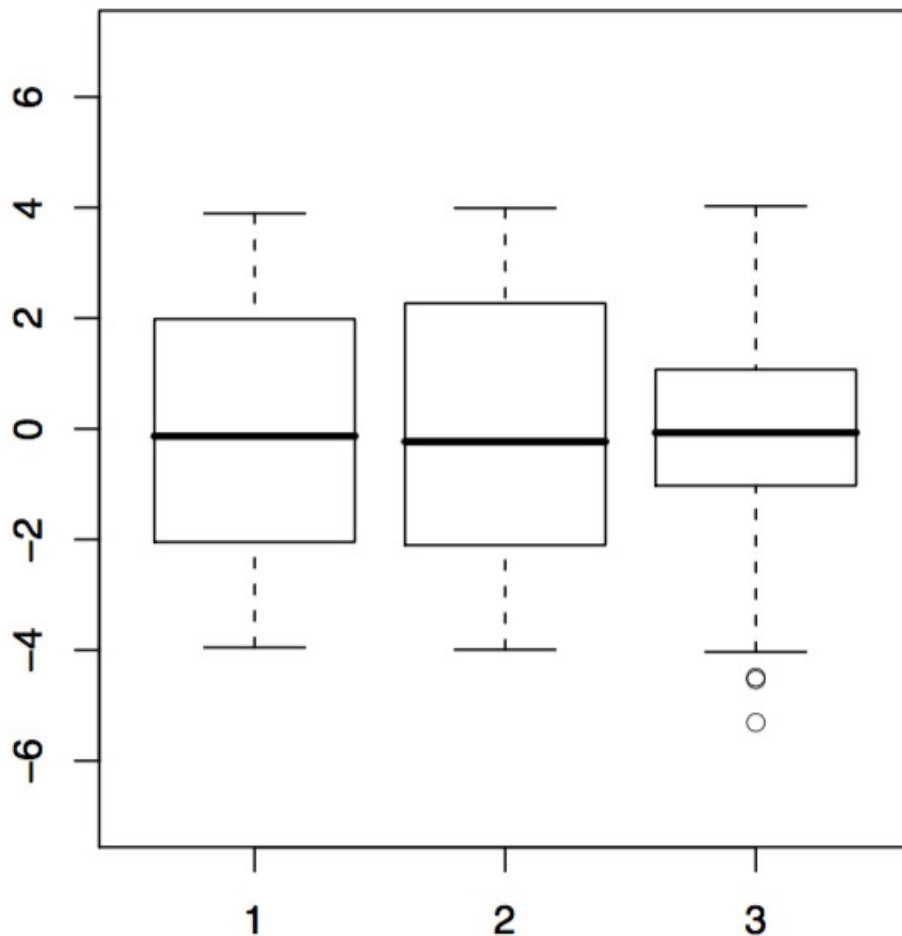


Shape of
distribution?
Normal distribution?

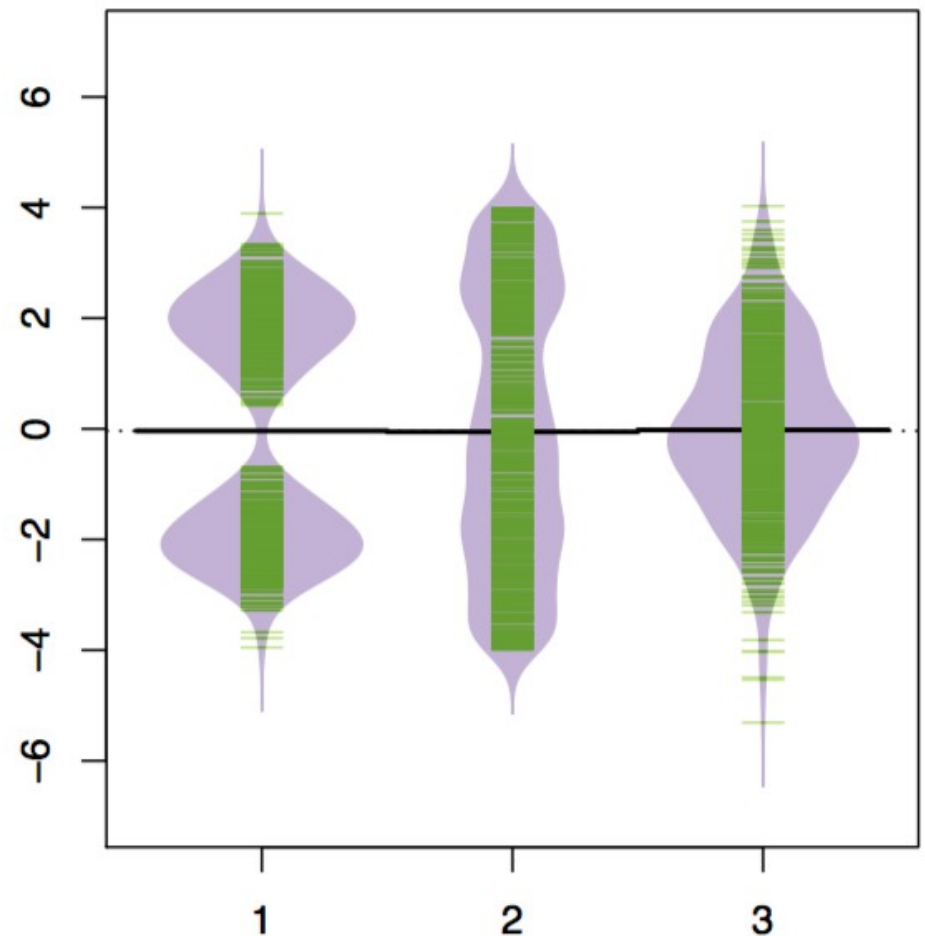
Tools for exploration: Beanplot

- More informative than boxplot, displays:
 - individual observations (green lines)
 - shape of distribution (could also be non-symmetric for sub-groups)
 - group mean and overall mean (instead of group median in boxplot)

boxplot



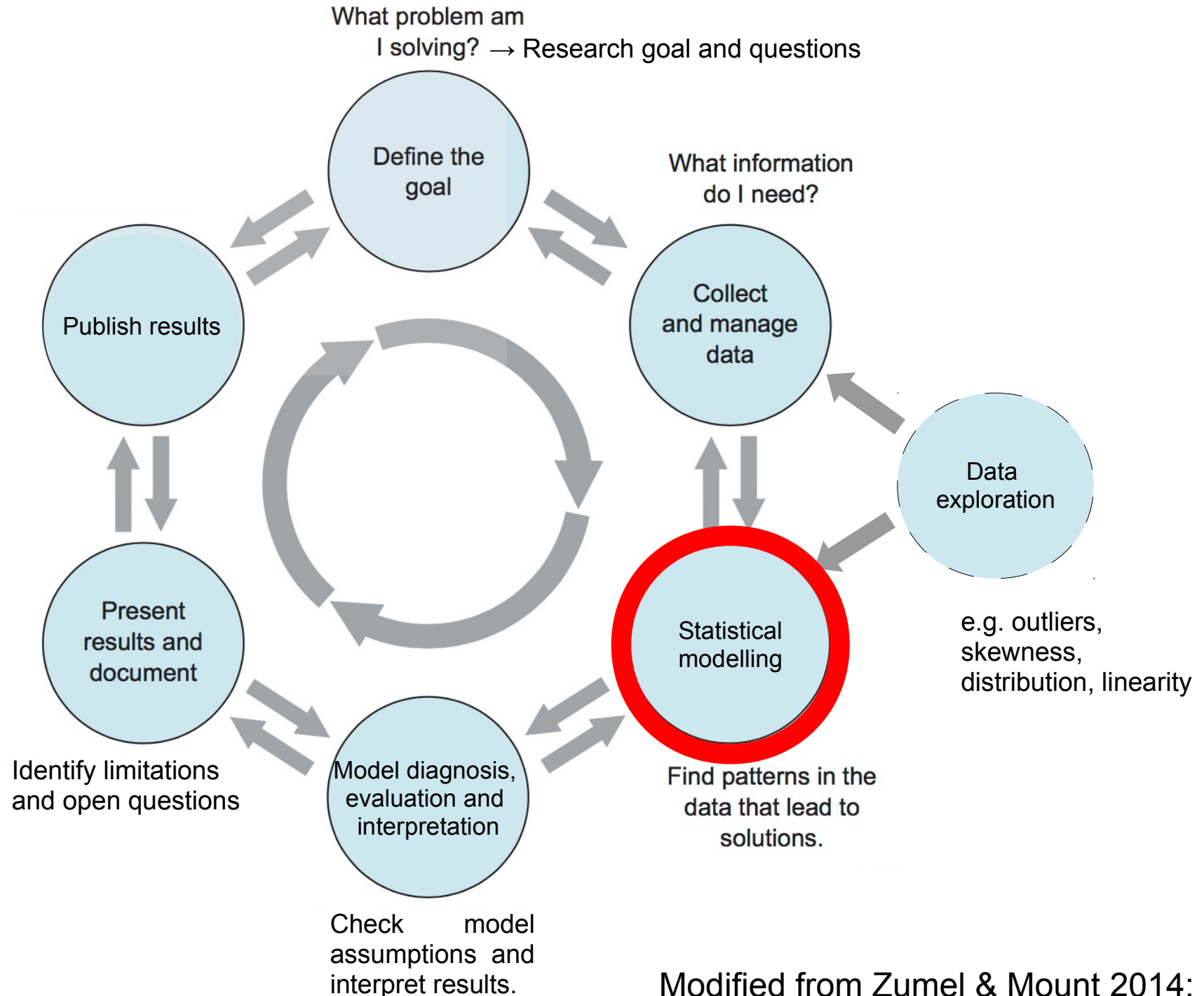
beanplot



Session overview

1. Framework for data analysis and research goals
2. Data exploration
- 3. Statistical modelling: Overview of techniques**

Data analysis cycle



Two incorrect ways of thinking about stats

1. Overconfidence: Statistics is like mathematics and provides a single, correct answer

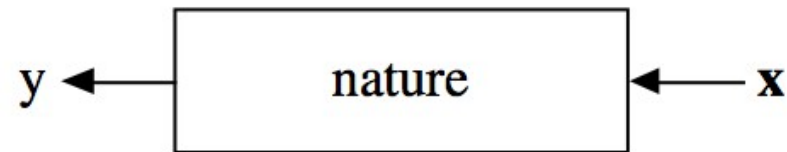
But: statistical thinking differs from mathematical thinking

2. Disbelief: Anything goes – statistics cannot be trusted

But: statistics provide quantitative support of the complete research process

Statistical modelling: The two cultures

Real world: Processes lead to association between X and Y



Examples for goals of statistical modelling: predict y from x , estimate relation between x and y

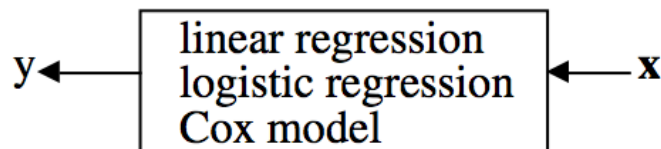
Data modelling culture
(classical statistics)

Algorithmic modelling culture
(machine learning)

Common data model

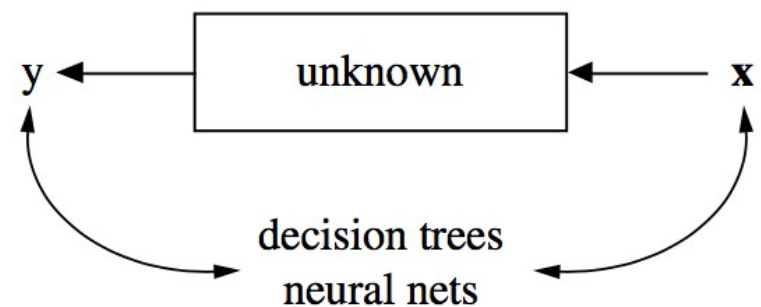
response variables = $f(\text{predictor variables, random noise, parameters})$

Estimate
parameters
from data



Model validation: Check model assumptions

Find algorithm that operates on x to predict y



Model validation: Predictive accuracy

Breiman 2001 *Statistical Science* 16: 199

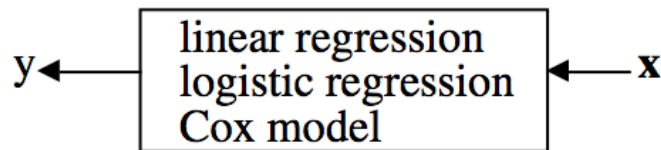
Statistical modelling: The two cultures

Data modelling culture (classical statistics)

Common data model

response variables = $f(\text{predictor variables, random noise, parameters})$

Estimate
parameters
from data



Model validation: Check model assumptions

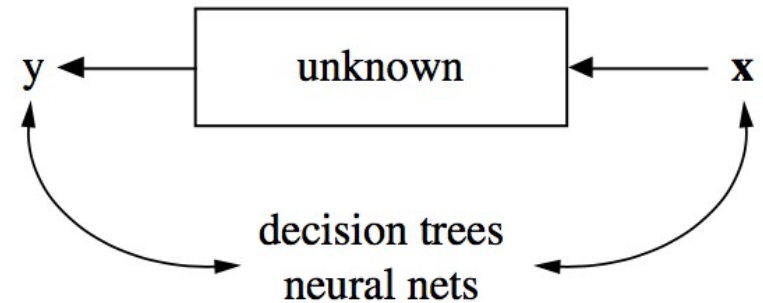
Methods appropriate for all research goals, but partly less powerful

Bayesian and frequentist frameworks

- Debates since decades (but also within frameworks)
 - Distinctions blurred in computational statistics
 - Inclusion of prior knowledge/beliefs in models only possible within Bayesian framework
- Will be discussed in more detail later

Algorithmic modelling culture (machine learning)

Find algorithm that operates on x to predict y



Model validation: Predictive accuracy

Methods primarily appropriate for prediction and exploration, and partly estimation

Overview univariate techniques of course

Type of analysis	Approach	Techniques	Research goals (this course)	Research goal (Paliy & Shankar 2016)
Association-based	Regression	Linear regression (Bayesian + frequentist), Regression trees	Prediction, Estimation, DPAH, Explanation	Identify variation in response explained by continuous variable(s)
	Correlation	Pearson correlation	Estimation, DPAH, Explanation, Exploration	Reveal relationship between variables
Group-based	Between-group comparisons	Analysis of variance (ANOVA), <i>t</i> -test	Prediction, Estimation, DPAH, Explanation	Identify variation in response explained by factor(s)
	Classification	Classification trees	Prediction, Estimation, DPAH, Explanation	Discriminate object classes based on values of measured variables

Much overlap:

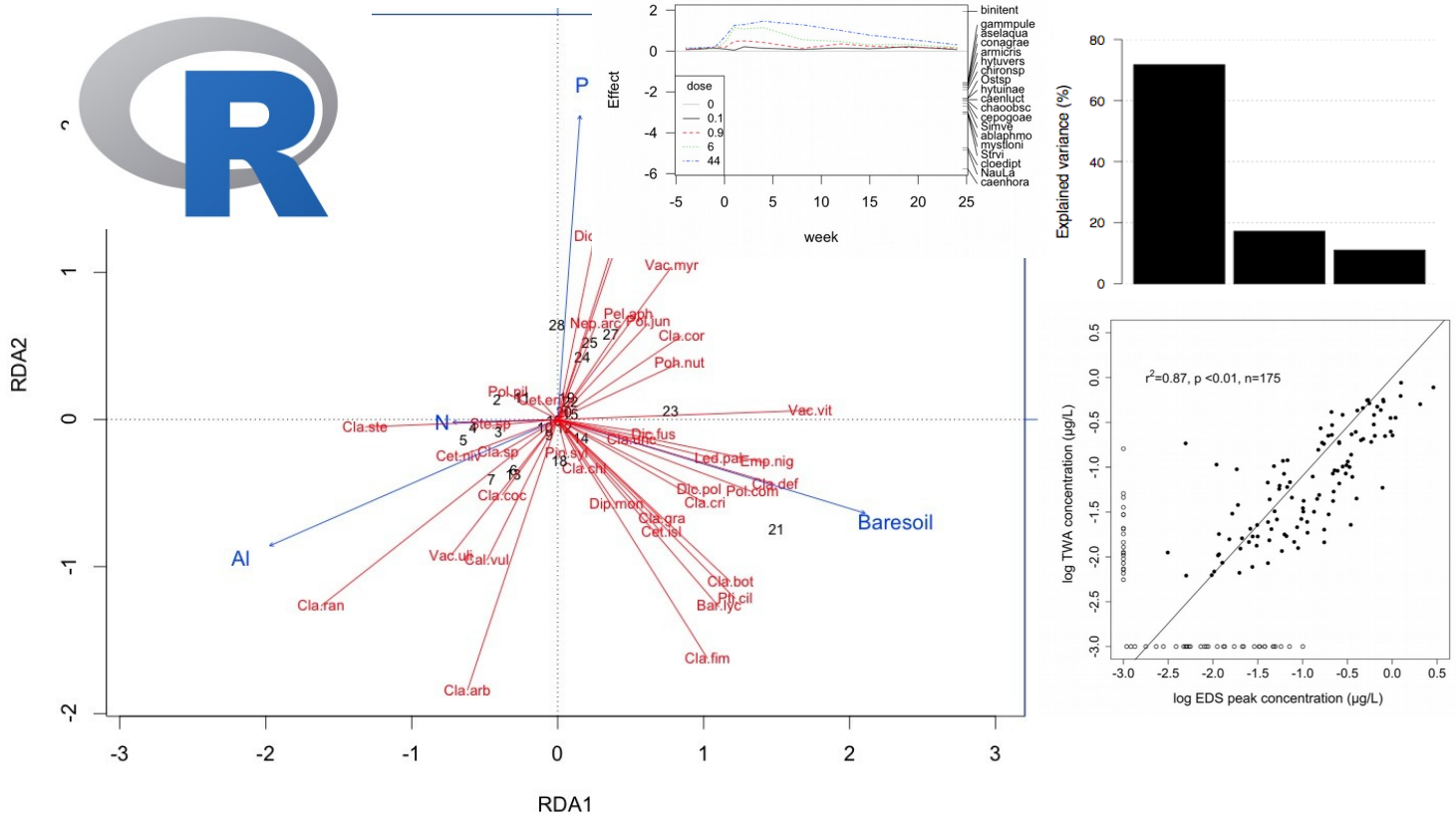
- e.g. Classification and Regression Trees (CART), Analysis of Covariance (ANCOVA)

Many extensions:

- e.g. Time series data, spatial data, special data structures

Tools for complex data analysis

University of Koblenz-Landau 2018/19



Ralf B. Schäfer

These slides and notes complement the lecture with exercises “Tools for complex data analysis” for ecotoxicologists and environmental scientists. Do not hesitate to contact me if you have any comments or you find any errors (slides, slide notes, or code): schaefer-ralf@uni-landau.de

While I made notes below the slides, some aspects are only mentioned in the R tutorials and demonstration associated with the lecture.

Short introduction


- Professor for Quantitative Landscape Ecology
- Current teaching: Data analysis (M.Sc.); GIS (B.Sc./M.Sc.); Environmental Modelling (B.Sc./M.Sc.); Environmental Philosophy (B.Sc.)
- Current research projects related to:
 - Community ecology of freshwater invertebrates and microorganisms
 - Response of freshwater ecosystems to different (anthropogenic) stressors (e.g. pollution)
 - Trophic linkages between aquatic & terrestrial systems
- Primarily field studies/experiments and data analyses/modelling
- Course assistants (Phd students): Stefan Kunz & Le Trong Dieu Hien (“Vicky”)



www.landscapeecology.uni-landau.de

Using your own notebook

- We encourage use of your own notebook!
- install [R](#) and subsequently [RStudio](#)
- Run “0_install_packgs.R”, provided on github
- for installation of additional packages run `install.packages(“package to be installed”)`



The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It runs on a variety of UNIX platforms, Windows and MacOS. To [download R](#), please use the [CRAN mirror](#).

If you have questions about R like how to download and install the software, please read our [answers to frequently asked questions](#) before you search.

News

- [R version 3.2.3 \(Wooden Christmas-Tree\)](#) prerelease versions will be available from 2015-11-11-30. Final release is scheduled for Thursday 2015-12-10.
- [R version 3.2.2 \(Fire Safety\)](#) has been released on 2015-08-14.
- [The R Journal Volume 7/1](#) is available.
- [R version 3.1.3 \(Smooth Sidewalk\)](#) has been released on 2015-03-05.
- [useR! 2015](#), took place at the University of Aalborg, Denmark, June 30 - July 4, 2015.
- [useR! 2014](#), took place at the University of California, Los Angeles, US, June 1-5, 2014.



Welcome to RStudio

Software, education, and services for the R community

Powerful IDE for R

RStudio IDE is a powerful and productive user interface for R. It's free and open source, and works great on Windows, Mac, and Linux.

[Download now](#) [Learn more](#)

R training and education

We've got hands-on courses for beginners and even R experts. Customize an on-site training or enroll in one of our public workshops.

[Request on-site](#) [View courses](#)

Open source R packages

Our developers and expert trainers are the authors of several popular R packages, including ggplot2, plyr, lubridate, and others.

[See projects](#)

© 2013 RStudio, Inc. [Follow @rstudioapp](#) | [Trademark](#) | [DMCA](#) | [Careers](#)

Course objectives: Learning targets

- Design a study and select corresponding tools for subsequent data analysis
- Select and apply techniques of data analysis for a research goal
- Classify, explain/interpret and critically evaluate different approaches to data analysis
- Programming simple to moderately complex data analysis tasks in R

4

The module hand book gives the following targeted learning outcomes: The students are able to design a study and select corresponding tools for subsequent data analysis. They can link scientific questions to methods of data analysis. The students are familiar with different approaches to data analysis including frequentist and bayesian statistics as well as machine learning approaches. The students are able to process research data and apply data analysis tools in a software environment. They know the advantages and disadvantages of the different methods.

4

Session overview

1. Framework for data analysis and research goals

2. Data exploration

3. Statistical modelling: Overview of techniques

Learning targets

- Explain the data analysis cycle
- Classify research goals
- Understand the role of data exploration and interpret and apply related tools
- Explain approaches to statistical modelling

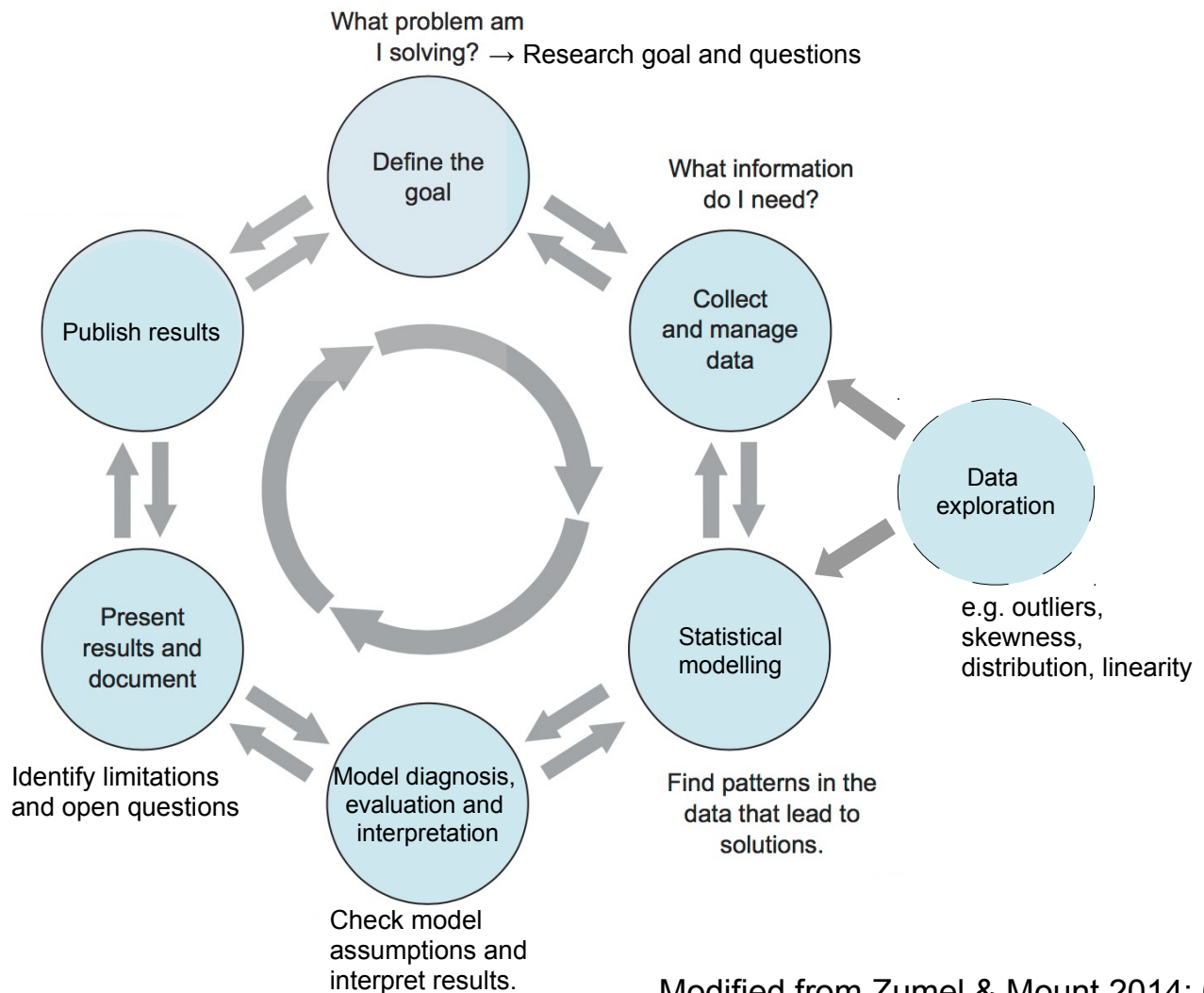
Learning targets and study questions

- Explain the data analysis cycle
 - Describe the steps of the data analysis cycle.
- Classify research goals
 - Distinguish different research goals and give an example for each.
 - Name a statistical technique for each research goal.
- Understand the role of data exploration and interpret and apply related tools
 - Describe the three roles of exploration.
 - Explain four tools for exploration including their domain of application in checking model assumptions.
 - What are quantiles? How are quantiles calculated?

Learning targets and study questions

- Explain approaches to statistical modelling
 - Discuss the core differences of the two main approaches to statistical modelling (e.g. method, research goals, validation)

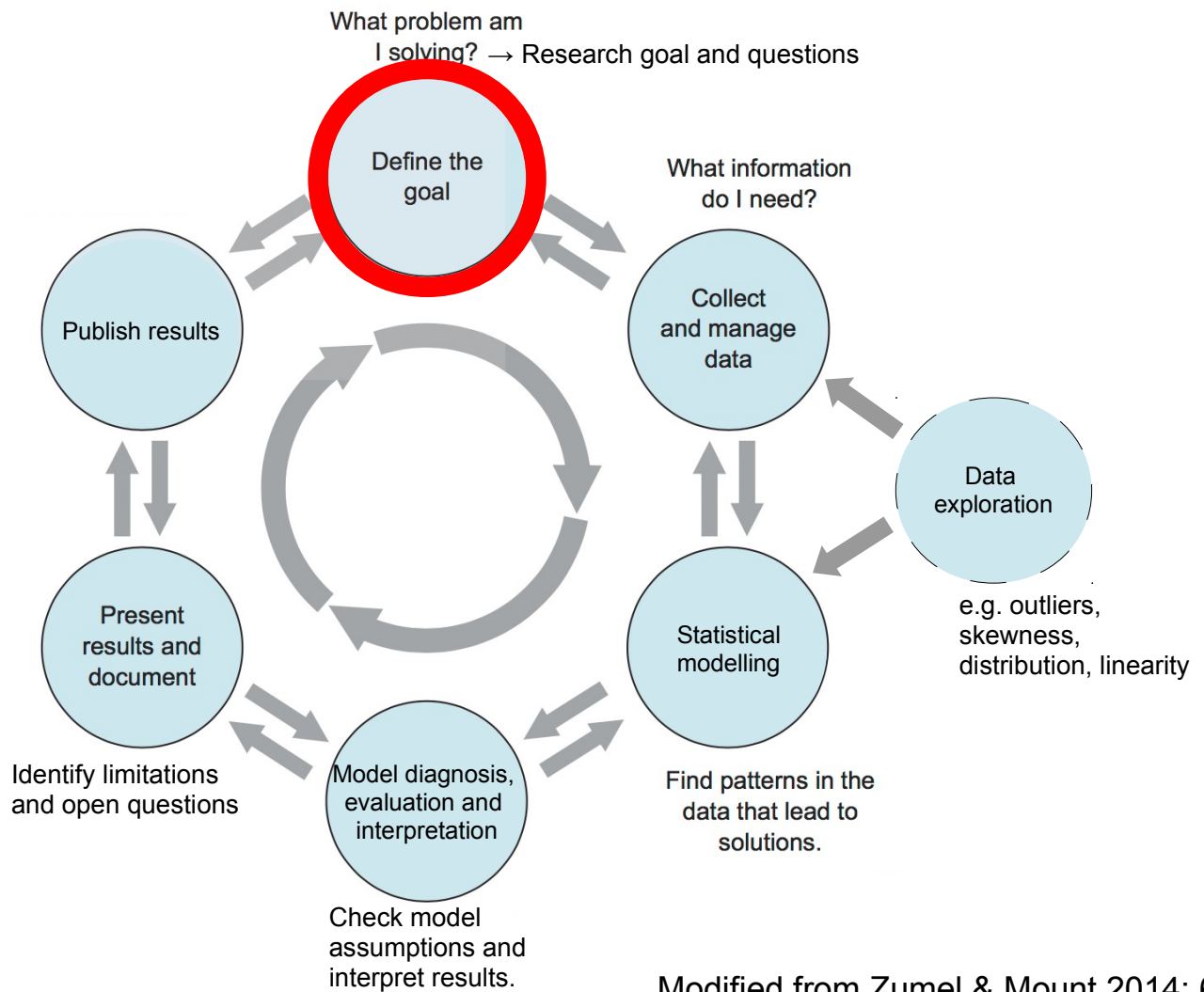
Data analysis cycle



Zumel N. & Mount J. (2014) Practical data science with R. Manning Publications Co, Shelter Island, NY.

Data exploration visualised with dashed line as it will depend on the research context if and when data exploration is conducted. However, most frequently data exploration (e.g. descriptive statistics such as data summaries) is employed before statistical modelling to aid in model selection and to identify errors or outliers. Moreover, the goal of some studies is exploration and eventually no statistical modelling is done. In this case, data exploration would directly lead to the presentation of results. Conversely, in case that a clear research hypothesis has been established before data collection and the data is known to be free from outliers or errors, data exploration may be unnecessary before statistical modelling. Still, the tools related to data exploration will be needed to check model assumptions.

Data analysis cycle



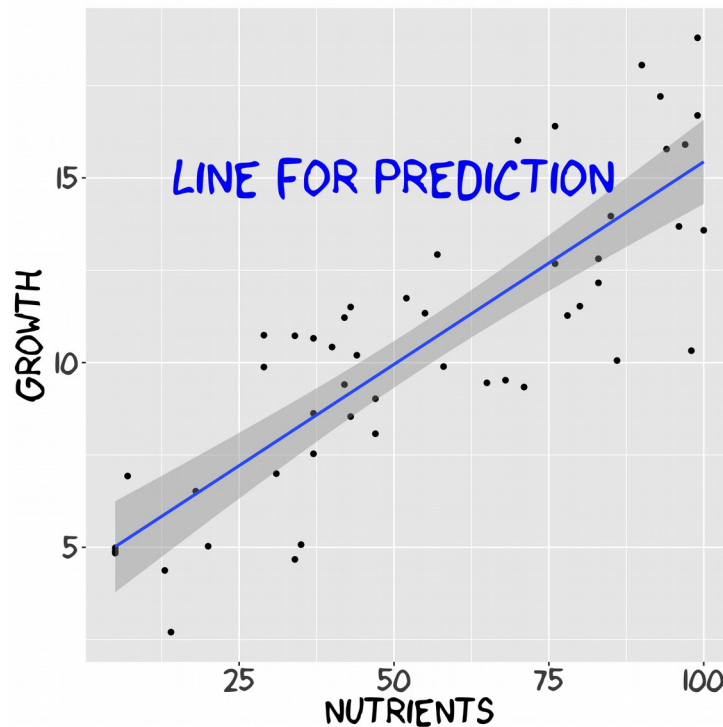
10

Zumel N. & Mount J. (2014) Practical data science with R. Manning Publications Co, Shelter Island, NY.

Research goals with examples

1. Prediction

Example: Establish linear relationship between mean plant growth and nutrient concentrations from observations that allows for prediction of mean plant growth for non-observed nutrient concentrations



11

We need to briefly clarify the difference between estimation and prediction. Estimation relates to the derivation of a parameter (e.g. the slope of the line), prediction to the value of the random variable that we predict using the regression line. Note that in the statistical population, the parameter has a precise true value, whereas the random variable that we predict carries uncertainty, deriving from the fact that it features a probability distribution.

Research goals with examples

2. (Parameter) estimation

Example: Estimate body mass of koalas under poor habitat conditions



https://commons.wikimedia.org/wiki/File:Friendly_Female_Koala.JPG

Let the continuous random variable X be the body mass of koalas and the discrete random variable Y be the habitat condition. We define ω as an outcome/event:

$$Y(\omega) = \begin{cases} 0 & = \text{poor habitat} \\ 1 & = \text{good habitat} \end{cases}$$

Aim: Estimate expected value E of body mass in poor habitat:

$$E[X|Y=0] = \mu_{X|Y=0}$$

If we cannot sample all koalas, we estimate based on a sample:

$$\hat{\mu}_{X|Y=0} = \bar{x}_{X|Y=0}$$

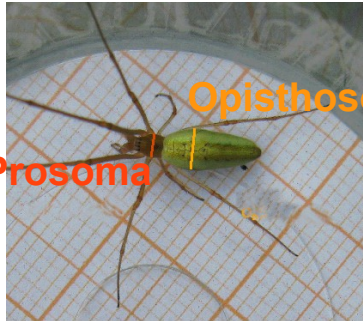
12

For example, a study may have measured the body mass of koalas under good and poor habitat conditions and found that the average body mass of koalas is 12 kg under good habitat conditions and 9 kg under poor habitat conditions. For the sake of simplicity of this example, we ignore that males and females exhibit considerable differences in body mass that would need to be accounted for.

Research goals with examples

3. Determination of probabilities and assessing hypotheses

Example: Does treatment of a spider with a pesticide at the typical application rate reduce the body size of the spider?



Scientific hypothesis: The pesticide requires the activation of energetically costly detoxication processes. This reduces the energy for growth and consequently the body size, measured as **prosomal** and **opisthosomal** width.

- Two perspectives:
 - How much evidence do the experimental results lend against the hypothesis of no pesticide effect? → Frequentist perspective
 - What is the probability that the pesticide has no effect in the light of the experimental results and prior beliefs or information? → Bayesian perspective

13

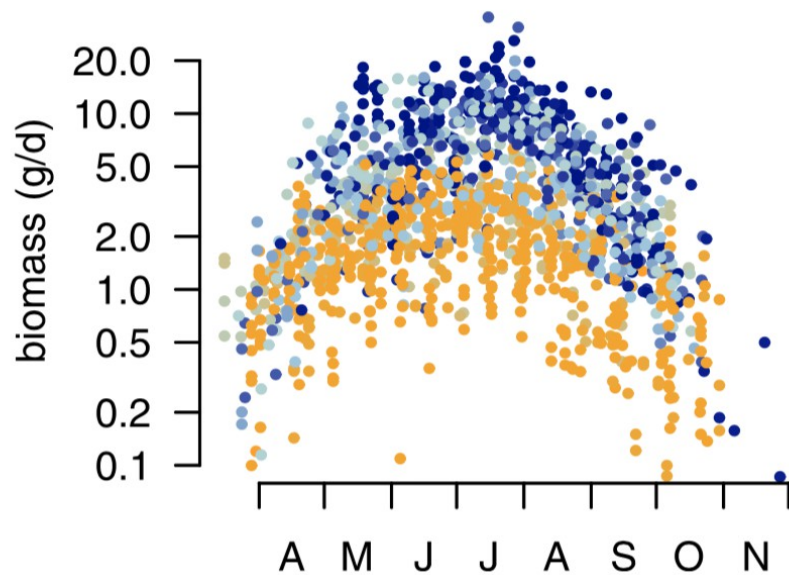
The determination of probabilities is typically an element of assessing a hypothesis. The more widely employed term is *hypothesis testing*. We will discuss later in the lecture, why we avoid this term. Moreover, another widely employed term is *statistical inference*. The term inference refers to the broader context, where the goal is to infer an unknown property of a population (e.g. the most likely value of the mean) based on a random sample from that population. You should be familiar with the term, though we selected another term since it is more self-explanatory.

We will discuss Bayesian and frequentist approaches to inference in detail later in the course.

Research goals with examples

4. Explanation

Example: Which variables do best explain the reduction of insect biomass (in Germany)?



Hallmann et al. 2017 *PLOS One* 12: e0185809

Approach: Identification of the most parsimonious model for insect biomass and of the contribution of variables to the explanatory power.

Model 1: Biomass ~ Temp., Precipitation, Frost days ...

Model 2: Biomass ~ Arable land, Forest, Grassland ...

Model 3: Biomass ~ Arable land, Temp., Frost days ...

⋮

Model n : Biomass ~ Forest, Precipitation

14

The figure shows the biomass over the seasons from 1989 to 2016, with a colour gradient from blue (1989) to orange (2016).

The paper is freely available under:
<https://doi.org/10.1371/journal.pone.0185809>

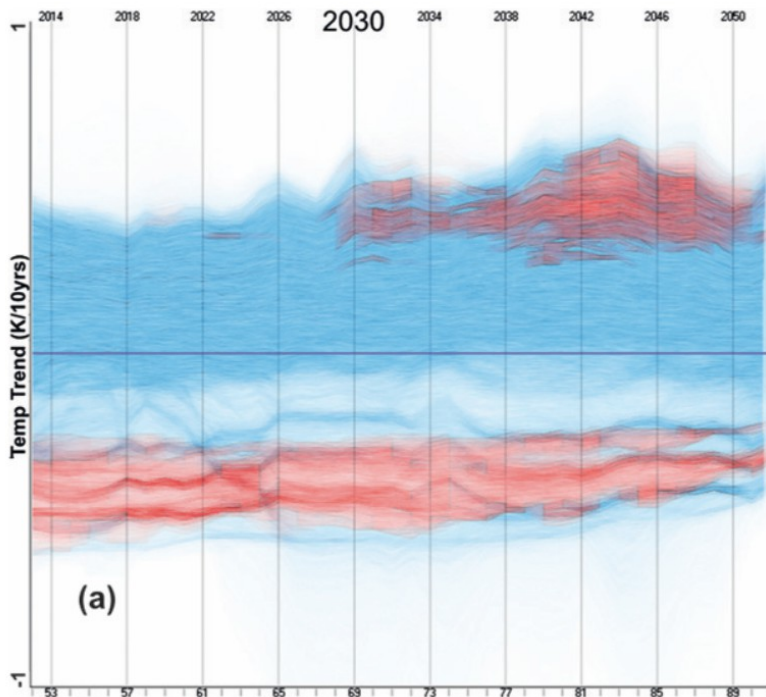
Note that the paper contains several technical errors (not only) and should not serve as a template for data analysis.

We consider “Explanation” as a research goal independent from the other goals despite some overlap. For example, model selection can also be framed as assessing the evidence for different hypotheses, where each hypothesis specifies a model. Nevertheless, decisions during model selection can differ between research goals (e.g. parsimony more important if the research goal is explanation than prediction). In addition, some approaches (e.g. the LASSO for simultaneous variable selection and estimation of variable importance) strongly differ from the approaches related to the classical framework employed when assessing hypotheses.

Research goals with examples

5. Exploration and descriptive statistics

Example: Explore large climate data set to generate new ideas and hypotheses.



Approach: Visual exploration and calculation of descriptive statistics.

Observations used in exploration to generate hypotheses must not be used in assessing statistical hypotheses!

Ladstädter et al. 2009 *J. Atmos. Oceanic Atmosph.* 27: 667

15

With large data sets continuously becoming available, explorative studies can be valuable to generate new ideas, understandings and hypotheses. These can be scrutinised in follow-up analyses of different data sets (or data that has been withheld, which will be discussed later in the context of training and test data), experiments or modelling studies. Though researchers always need some theoretical framework to run data analyses (e.g. to identify response variables), exploratory studies do not start from *a priori* hypotheses and do not focus on the estimation of specific parameters.

Overview on research goals

1. Prediction

2. (Parameter) estimation

3. Determination of probabilities and assessing hypotheses

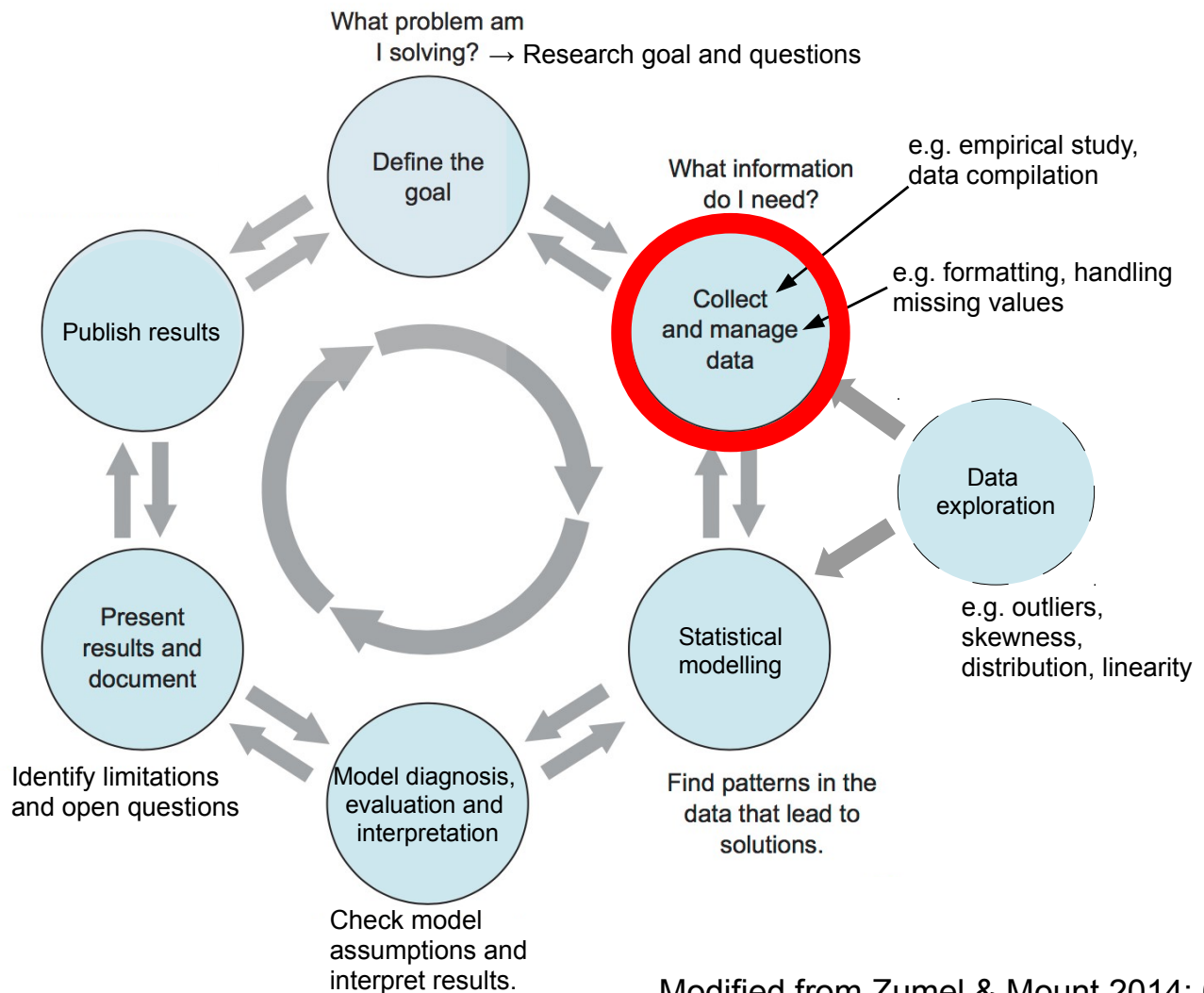
4. Explanation

5. Exploration and descriptive statistics



Inform study design and method selection in data analysis

Data analysis cycle



The topic of data management and pre-processing is touched during some of the practical parts, though a thorough treatment of the topic is beyond the scope of this course. Baumer et al. (2017: 63-131) and Wickham & Gronemund (2016: 43-261) provide guidance (and computer code) on organising data, with a main focus on data that has been imported into R. A brief overview on importing data into R is given in a blog post:

<https://www.datacamp.com/community/tutorials/r-data-import-tutorial>

A general overview on tools for managing biological data including relational databases and file processing is provided by Haddock & Dunn (2011).

Baumer, B., Kaplan, D., Horton, N.J., 2017. Modern data science with R, Chapman & Hall/CRC texts in statistical science series. CRC Press, Boca Raton.

Haddock, S.H.D., Dunn, C.W., 2011. Practical computing for biologists, Sinauer Associates, Sunderland, Mass.

Wickham, H., Grolemond, G., 2016. R for data science: import, tidy, transform, visualize, and model data, First edition. ed. O'Reilly, Sebastopol, CA.

Zumel N. & Mount J. (2014) Practical data science with R. Manning Publications Co, Shelter Island, NY.

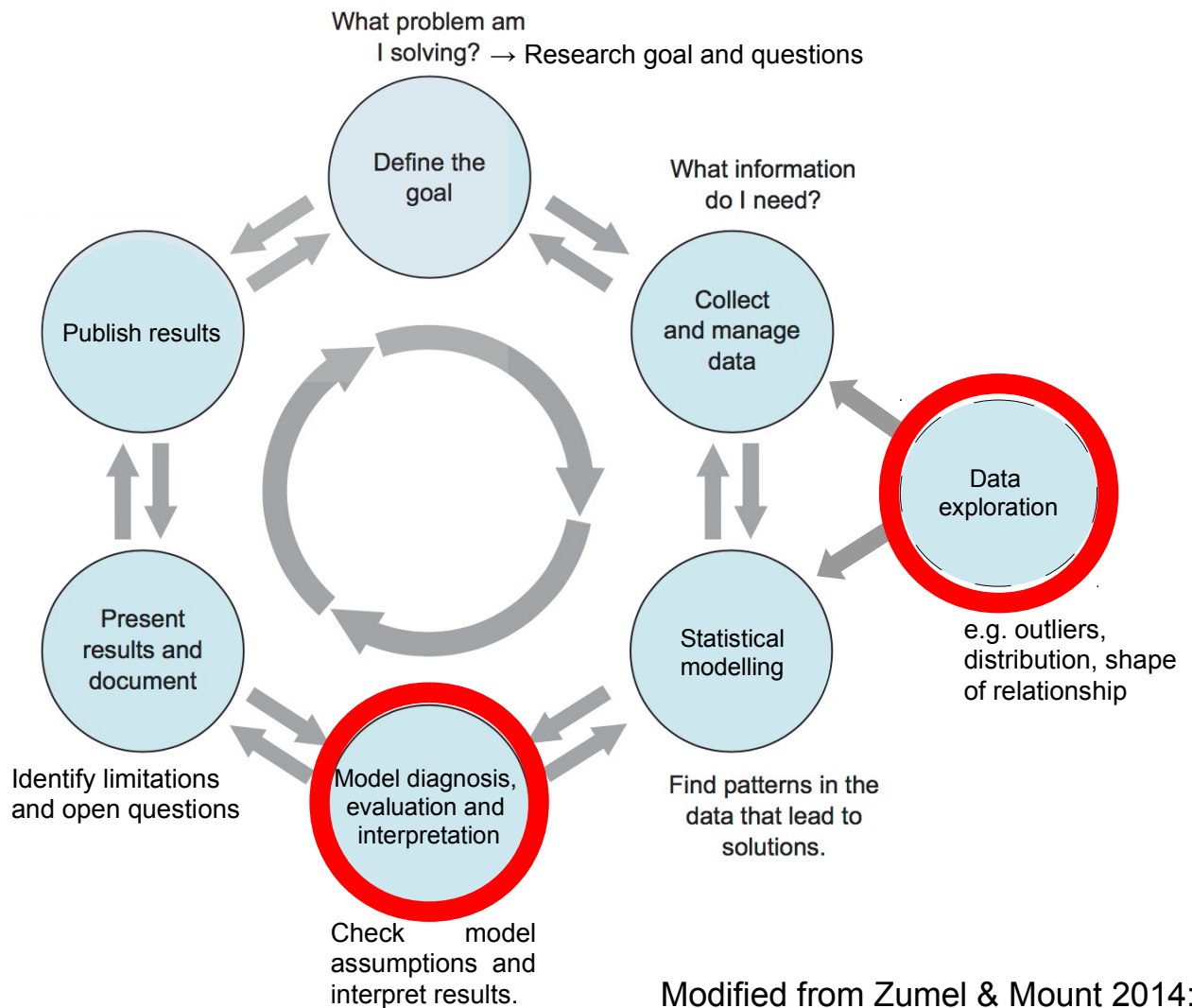
Session overview

1. Framework for data analysis and research goals

2. Data exploration


3. Statistical modelling: Overview of techniques

Data analysis cycle



Zumel N. & Mount J. (2014) Practical data science with R. Manning Publications Co, Shelter Island, NY.

Exploration and descriptive statistics

- Has at least three roles:
 - Suggest new ideas, understandings or hypotheses
→ stimulate additional analyses or follow-up studies
 - Facilitate model selection, checking of assumptions and identification of errors or outliers in the data
-  *GIGA: Garbage in – Garbage out*
- Provide tools for communicating the data

20

Maindonald & Brown (2010: 44) list four roles that I consider related. For example, they differentiate between the role of generating new ideas and understanding, the role of revealing additional information to stimulate new research avenues and the role of challenging current theoretical understanding. For me, these roles are strongly related and will in most cases use the same tools, and can only be differentiated based on the outcome of the research process.

To give a detailed overview on exploratory analysis, is beyond the scope of this course. We only discuss a few central tools for exploration, but several books are available that provide examples with code that can easily be adjusted to your purpose. For example, Wickham & Golemund (2016: 81-111) provide an overview on tools for exploratory data analysis. Baumer et al. (2017: 9-63) discuss pitfalls when visualising data and give guidance on visual data exploration. Gohil (2015), Rahlf (2017) and Wickham (2016) provide many examples for graphs with computer code. Also visit the R graph gallery for an overview on ways to visualise data: <https://www.r-graph-gallery.com/> or the <https://www.data-to-viz.com> project.

Baumer, B., Kaplan, D., Horton, N.J., 2017. Modern data science with R, Chapman & Hall/CRC texts in statistical science series. CRC Press, Boca Raton.
Gohil, A., 2015. R data visualization cookbook: over 80 recipes to analyze data and create stunning visualizations with R. Packt Publishing, Birmingham.
Rahlf, T., 2017. Data visualisation with R: 100 examples. Springer, Cham. (also available in german).
Wickham, H., 2016. Ggplot2: elegant graphics for data analysis, 2nd ed, Springer, New York.
Wickham, H., Golemund, G., 2016. R for data science: import, tidy, transform, visualize, and model data, O'Reilly, Sebastopol, CA.

Tools for exploration: Model selection, assumptions, outliers and errors

Checking for:	Example for tool
Outliers and errors	Boxplot, Cleveland plot
Variance homogeneity	Conditional boxplot, Beanplot
Normal distribution	QQ-plot ¹ , Histogram
Shape of distribution	Histogram
(Double) zeros	Frequency plot or table
Collinearity	Scatterplot
Shape of relation between predictor and response variable	Scatterplot
Interactions	Coplots, Interaction plots
Spatial- or temporal autocorrelation	Variograms

¹QQ-plot = Quantile-Quantile plot

21

Zuur et al. 2009 *Meth. Ecol. Evol.* 1: 3

Outliers are observations that deviate strongly from the other observations. An outlier is an influential point if it influences a statistic, parameter or model. For example, an outlier typically influences the mean but not the mode or median, i.e. would be an influential point for the mean but not for the mode or median. Errors, i.e. values in the data that do not match the real measurement (e.g. putting a decimal point at the wrong position when typing raw data into the computer), are often outliers, unless a relevant fraction of the data are wrong or the error results in a systematic bias. Imagine the case that two data sets have been merged ignoring that the same variable had been measured on different units (e.g. *g* and *mg*). Rather than in outliers, this would result in a joint data set with a conspicuous bi-modal distribution, which could be spotted with a Cleveland plot or beanplot (but harder to spot with a boxplot). A very readable discussion on how to spot and deal with outliers is given in Ieno & Zuur (2015: Chapter 2).

All other assumptions mentioned (e.g. variance homogeneity, collinearity) will be discussed in detail later in the course.

A recommended read on data exploration is the highly cited paper by Zuur et al (2009).

Ieno E.N. & Zuur A.F. (2015) A beginner's guide to data exploration and visualization with R. Highland Statistics Ltd, Newburgh.

Zuur, A.F.; Ieno, E.N.; Elphick, C.S (2009): A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* 1: 3–14. Free to download at: <https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/j.2041-210X.2009.00001.x>

21

Quantiles

- Essential for several plots (e.g. QQ-plot, boxplot)
- Cut (ordered) data into subsets of equal size/probability

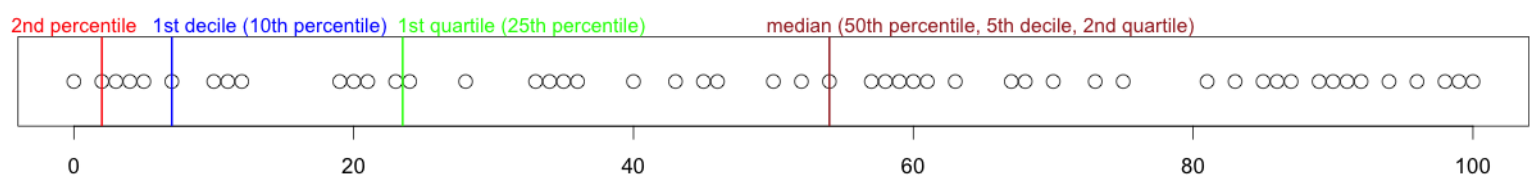
- Definition:

For data $x = \{x_1, x_2, x_3, \dots, x_n\}$, where the values are ordered as: $x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(n)}$ (\rightarrow order statistics), the q -quantiles are values that partition x into q subsets.

- E.g. 2-quantile = median, 4-quantiles = quartiles, 10-quantiles = deciles

The x_k corresponding to the k -th q -quantile ($k \in \mathbb{N}_0 \wedge 0 < k < q$) cuts with: $P(X < x_k) \leq k/q$ and $P(X \geq x_k) \geq 1 - k/q$

Example: 51 random samples x (without replacement) with $x \in \mathbb{N}_0 \wedge x < 101$



22

The 2nd, 10th and 50th percentile corresponds to specific observations x (i.e. 2, 7 and 54), whereas the 25th percentile (23.5) represents an interpolation between two data values.

Quartiles are used to construct boxplots. QQ-plots typically extract the sample quantiles from the assumed (population) probability distribution to compare the values for the sample quantiles to those from the theoretical distribution.

Example: For the 5-quantiles, $k = \{1, 2, 3, 4\}$ with:

$$P(X < x_1) \leq 1/5 \text{ and } P(X \geq x_1) \geq 1 - 1/5$$

$$P(X < x_2) \leq 2/5 \text{ and } P(X \geq x_2) \geq 1 - 2/5$$

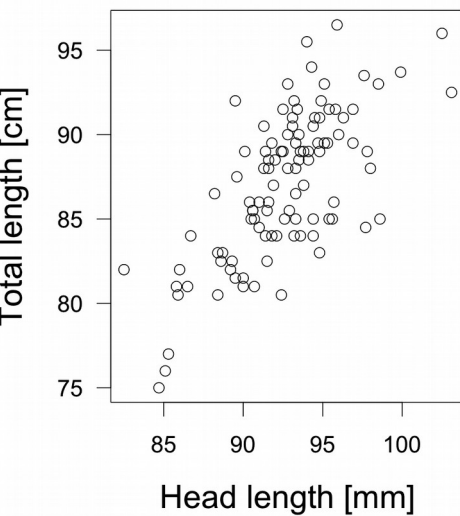
$$P(X < x_3) \leq 3/5 \text{ and } P(X \geq x_3) \geq 1 - 3/5$$

$$P(X < x_4) \leq 4/5 \text{ and } P(X \geq x_4) \geq 1 - 4/5$$

22

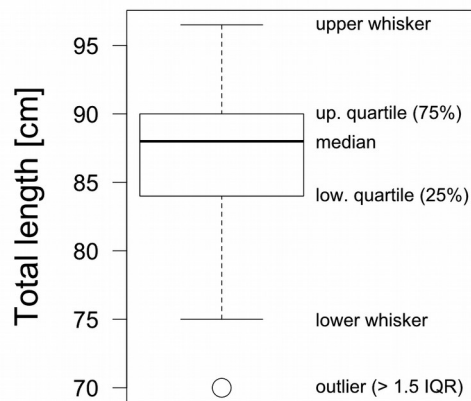
Common tools for exploration

Scatterplot



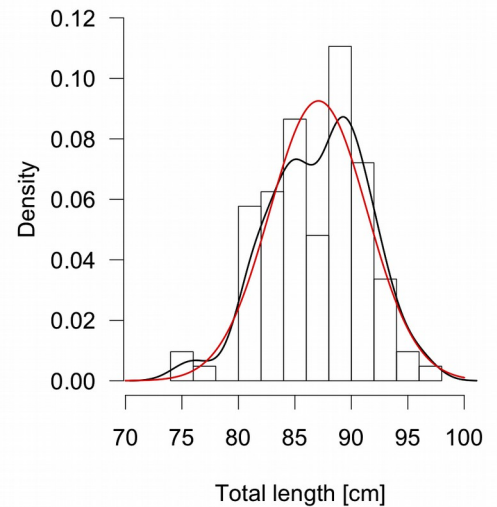
Shape of
relationship?
Collinearity?

Boxplot



Outliers and
errors? Variance
homogeneity?

Histogram with density curve
and normal distribution

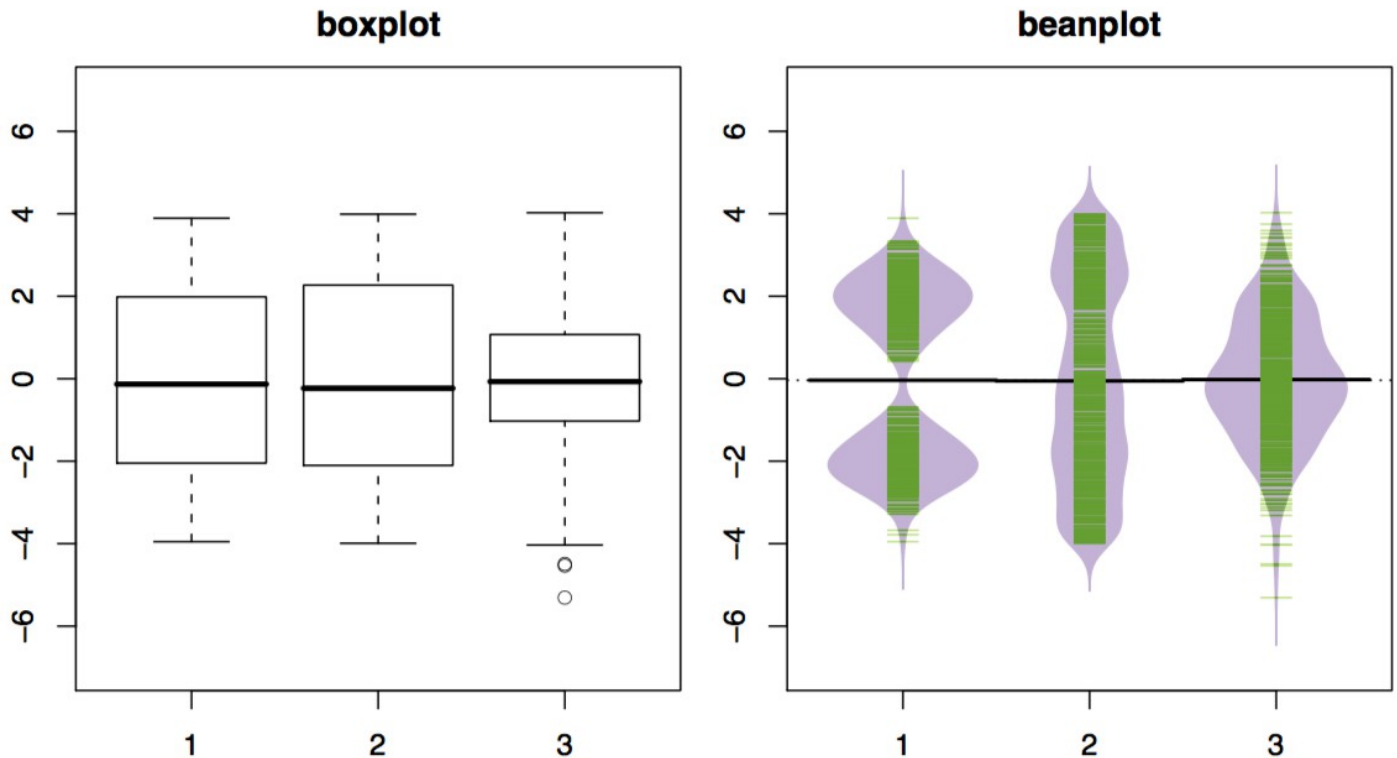


Shape of
distribution?
Normal distribution?

There are several rules of thumb as to what can be regarded as an outlier – but it remains more or less a subjective decision. John Tukey suggested to define x as an outlier if: $x < (Q1 - 1.5 \text{ IQR})$ or $x > (Q3 + 1.5 \text{ IQR})$, where $Q1$ denotes the lower quartile, $Q3$ denotes the upper quartile, and $\text{IQR} = (Q3 - Q1)$ denotes the interquartile range. In practice, the type of data, number of observations and knowledge about the data should be taken into account when deciding about the classification and how to deal with an outliers.

Tools for exploration: Beanplot

- More informative than boxplot, displays:
 - individual observations (green lines)
 - shape of distribution (could also be non-symmetric for sub-groups)
 - group mean and overall mean (instead of group median in boxplot)



24

Kampstra 2008 *J. Stat. Softw.* 9:1

Note that the groups 1 and 2 look fairly similar in the boxplot, whereas their differences are obvious in the beanplot. The overall mean is barely visible for the displayed data in the beanplot due to overlap with the individual mean, see Kampstra (2008) for plots where the overall mean is better visible.

Tools such as boxplots are commonly used and most academics are familiar with their interpretation. Nevertheless, new tools such as the beanplot can be more informative than the tools commonly used in exploration. New tools can be found in the scientific literature or websites devoted to graphs such as the R graph gallery <https://www.r-graph-gallery.com/>. Gohil (2015), Rahlf (2017) and Wickham (2016) provide many examples for graphs along with computer code.

Gohil, A., 2015. R data visualization cookbook: over 80 recipes to analyze data and create stunning visualizations with R. Packt Publishing, Birmingham.

Kampstra P. (2008) Beanplot: A Boxplot Alternative for Visual Comparison of Distribution. Journal of Statistical Software 28, 1–9. Freely available at <http://www.jstatsoft.org/v28/c01/>

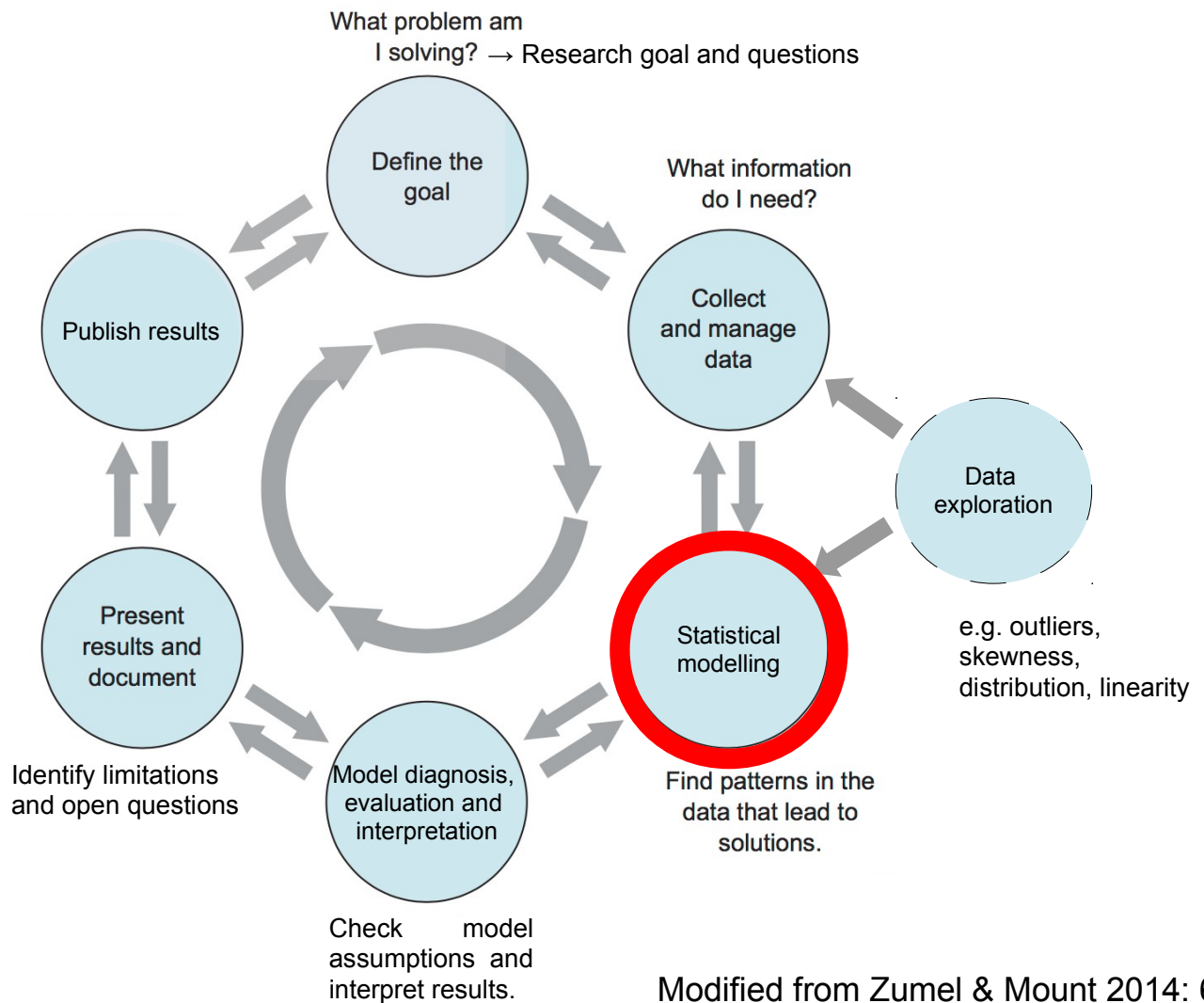
Rahlf, T., 2017. Data visualisation with R: 100 examples. Springer, Cham. (also available in german).

Wickham, H., 2016. Ggplot2: elegant graphics for data analysis, 2nd ed, Springer, New York.

Session overview

1. Framework for data analysis and research goals
2. Data exploration
- 3. Statistical modelling: Overview of techniques**

Data analysis cycle



26

Zumel N. & Mount J. (2014) Practical data science with R. Manning Publications Co, Shelter Island, NY.

Two incorrect ways of thinking about stats

1. Overconfidence: Statistics is like mathematics and provides a single, correct answer

But: statistical thinking differs from mathematical thinking

2. Disbelief: Anything goes – statistics cannot be trusted

But: statistics provide quantitative support of the complete research process

Compared to mathematics, particularly subfields such as abstract algebra or number theory, statistics is strongly linked to its application. Issues such as study design, data quality and research methods influence the choice of the statistical method as well as the validity in terms of credibility and reliability in terms of reproducibility of the answer it provides. In particular, it is a misconception that larger sample sizes automatically result in less biased parameter estimates and more reliable results.

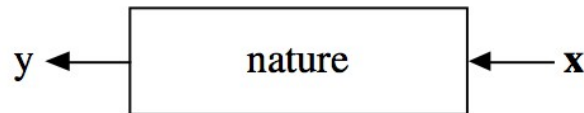
Conversely, statistics (and data analysis) are an essential part of the research process. As for any part of the research process, scientists may disagree in some cases, for example on the choice and interpretation of methods of data analysis. But this does not mean that *anything goes* or that *there are lies, damned lies and statistics*. In the end, transparency and reproducibility of statistical methods and results is pivotal to create trust (see Gandrud (2016) for tools for reproducible research and the ROpen Sci project: <https://ropensci.org/about/>).

Gandrud C. (2016) Reproducible research with R and R Studio. 2nd ed. CRC Press/Taylor & Francis Group, Boca Raton. Free to download at: <https://englianhui.files.wordpress.com/2016/01/reproducible-research-with-r-and-studio-2nd-edition.pdf>

Tintle N., Chance B., Cobb G., Roy S., Swanson T. & VanderStoep J. (2015) Combating Anti-Statistical Thinking Using Simulation-Based Methods Throughout the Undergraduate Curriculum. *The American Statistician* 69, 362–370.

Statistical modelling: The two cultures

Real world: Processes lead to association between X and Y



Examples for goals of statistical modelling: predict y from x , estimate relation between x and y

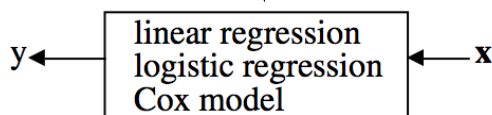
Data modelling culture
(classical statistics)

Algorithmic modelling culture
(machine learning)

Common data model

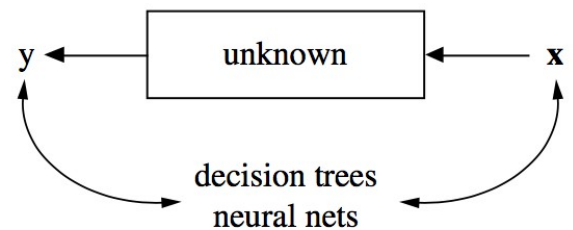
response variables = $f(\text{predictor variables, random noise, parameters})$

Estimate
parameters
from data



Model validation: Check model assumptions

Find algorithm that operates on x
to predict y



Model validation: Predictive accuracy

Breiman 2001 *Statistical Science* 16: 199

Breiman L. (2001) Statistical modeling: The two cultures. *Statistical Science* 16, 199–215.

The very readable debate is available at:

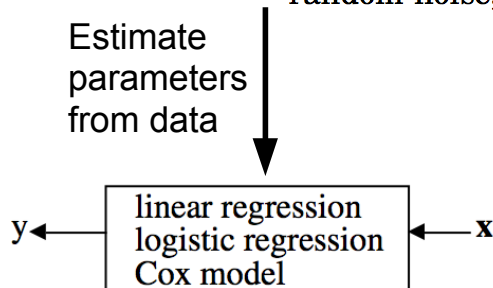
https://projecteuclid.org/download/pdf_1/euclid.ss/1009213726

Statistical modelling: The two cultures

Data modelling culture (classical statistics)

Common data model

response variables = $f(\text{predictor variables, random noise, parameters})$



Model validation: Check model assumptions

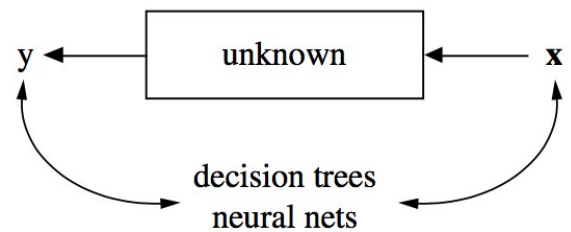
Methods appropriate for all research goals, but partly less powerful

Bayesian and frequentist frameworks

- Debates since decades (but also within frameworks)
 - Distinctions blurred in computational statistics
 - Inclusion of prior knowledge/beliefs in models only possible within Bayesian framework
- Will be discussed in more detail later

Algorithmic modelling culture (machine learning)

Find algorithm that operates on x to predict y



Model validation: Predictive accuracy

Methods primarily appropriate for prediction and exploration, and partly estimation

Breiman 2001 *Statistical Science* 16: 199

For how the distinctions between Bayesian and frequentist frameworks are blurred, see Efron & Hastie (2016).

Moreover, the distinction between the classical statistical and machine learning community is also becoming obsolete with the advent of many techniques that build on both statistics and machine learning. For details see Ryo & Rillig (2017). Moreover, we will meet several techniques in the course that fuse data models with algorithmic modelling.

Breiman L. (2001) Statistical modeling: The two cultures. *Statistical Science* 16, 199–215.

Efron B. & Hastie T. (2016) Computer age statistical inference: algorithms, evidence, and data science. Cambridge University Press, New York, NY.

Ryo M. & Rillig M.C. (2017) Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere* 8, e01976.

Overview univariate techniques of course

Type of analysis	Approach	Techniques	Research goals (this course)	Research goal (Paliy & Shankar 2016)
Association-based	Regression	Linear regression (Bayesian + frequentist), Regression trees	Prediction, Estimation, DPAH, Explanation	Identify variation in response explained by continuous variable(s)
	Correlation	Pearson correlation	Estimation, DPAH, Explanation, Exploration	Reveal relationship between variables
Group-based	Between-group comparisons	Analysis of variance (ANOVA), <i>t</i> -test	Prediction, Estimation, DPAH, Explanation	Identify variation in response explained by factor(s)
	Classification	Classification trees	Prediction, Estimation, DPAH, Explanation	Discriminate object classes based on values of measured variables

Much overlap:

- e.g. Classification and Regression Trees (CART), Analysis of Covariance (ANCOVA)

Many extensions:

- e.g. Time series data, spatial data, special data structures

30

Here we provide an overview on the techniques introduced in this course, classified by the type of analysis and the general approach to data analysis. In addition, we indicate for which research goals the techniques may be appropriate. The right column gives the research goal according to the classification of Paliy & Shankar (2016).

We abbreviated “Determination of probabilities and assessing hypotheses” with “DPAH” in the respective column.

Several techniques combine approaches. For example, Classification and Regression trees combine regression and classification. Moreover, analysis of covariance links regression and between-group comparisons.

Paliy O. & Shankar V. (2016) Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology* 25, 1032–1057.