GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

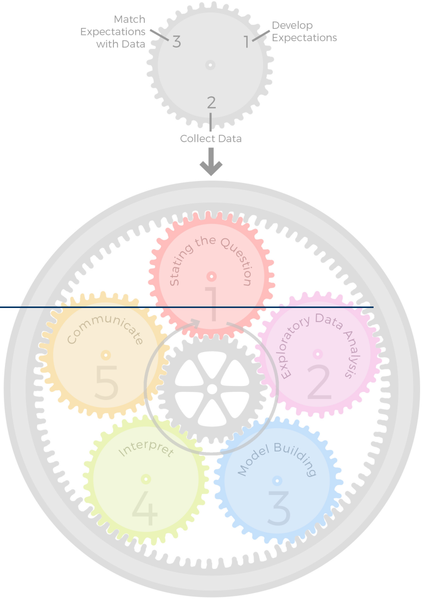
# Data Science – Basics

## Lecture 02 – How to approach a project


Fabian Sinz


15. April 2024


Institute for Computer Science – Campus Institute for Data Science (CIDAS)

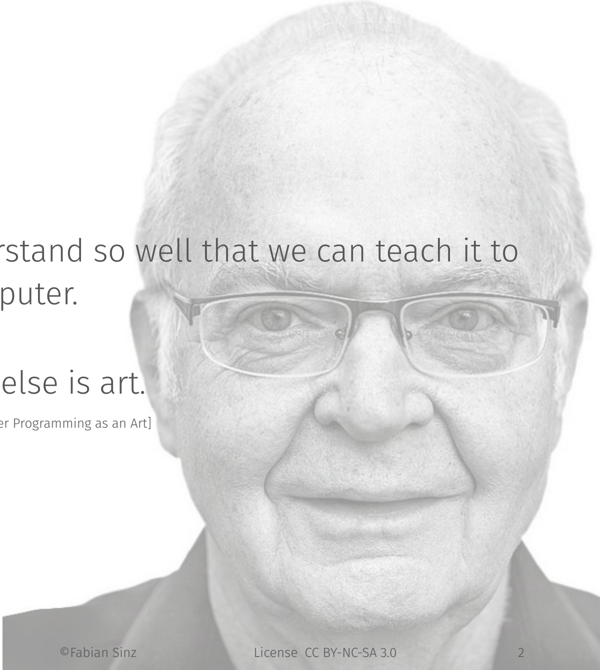




Epicycles of Analysis

Science is knowledge which we understand so well that we can teach it to a computer.

Everything else is art.

[Donald E. Knuth – Computer Programming as an Art]



# Epicycles of Analysis

# Lecture Schedule



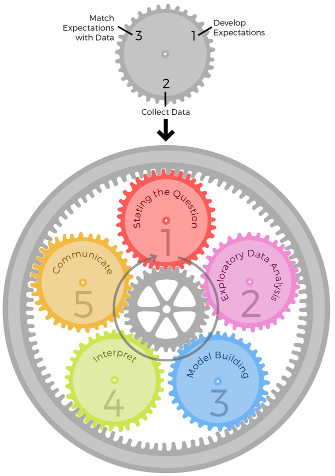
| | | |
|---|---|---|
| 1. | What is Data Science? | 08.04.2024 |
| 2. | How to approach a project | 15.04.2024 |
| 3. | Shell | 22.04.2024 |
| 4. | Data | 29.04.2024 |
| 5. | Visualization and Descriptive Statistics | 06.05.2024 |
| 6. | Clean Code | 13.05.2024 |
| 7. | Versioning | 27.05.2024 |
| 8. | Virtual Environments and Containerization | 03.06.2024 |
| 9. | Inferential Statistics | 10.06.2024 |
| 10. | Experimental Design | 17.06.2024 |
| 11. | Supervised Learning | 24.06.2024 |
| 12. | Unsupervised Learning | 01.07.2024 |
| 13. | Reporting and Time Management | 08.07.2024 |

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

## There are 5 core activities of data analysis:

1. Stating and refining the question

## Epicycle of Analysis




Epicycles of Analysis

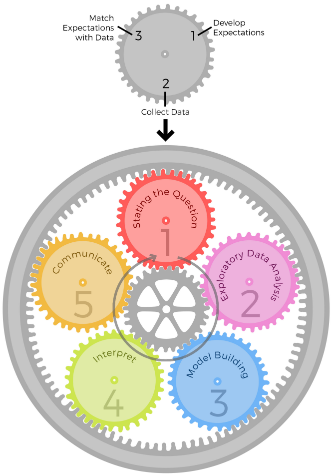
[Roger D. Peng & Elizabeth Matsui: The Art of Data Science]

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

There are 5 core activities of data analysis:

1. Stating and refining the question
2. Exploring the data

Epicycle of Analysis




Epicycles of Analysis

[Roger D. Peng & Elizabeth Matsui: The Art of Data Science]

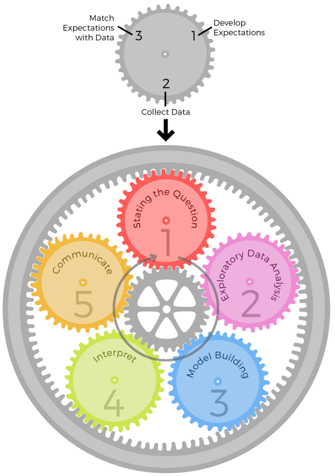
  

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

## There are 5 core activities of data analysis:

1. Stating and refining the question
2. Exploring the data
3. Building formal statistical models

## Epicycle of Analysis



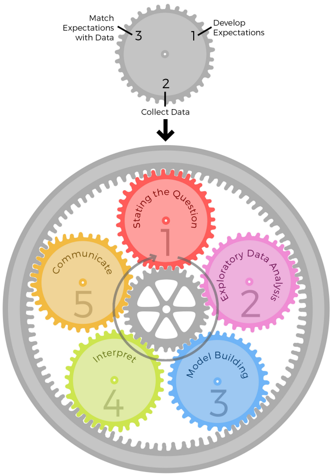

Epicycles of Analysis

[Roger D. Peng & Elizabeth Matsui: The Art of Data Science]

There are 5 core activities of data analysis:

1. Stating and refining the question
2. Exploring the data
3. Building formal statistical models
4. Interpreting the results

Epicycle of Analysis



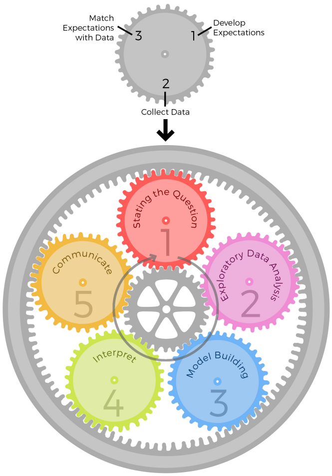

**Epicycles of Analysis**

[Roger D. Peng & Elizabeth Matsui: The Art of Data Science]

## There are 5 core activities of data analysis:

1. Stating and refining the question
2. Exploring the data
3. Building formal statistical models
4. Interpreting the results
5. Communicating the results

## Epicycle of Analysis




**Epicycles of Analysis**

[Roger D. Peng & Elizabeth Matsui: The Art of Data Science]

# (Epi)Cycles of Analysis

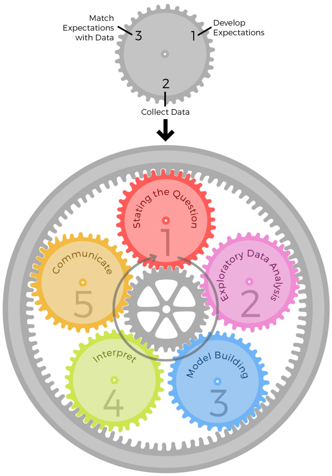
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

## There are 5 core activities of data analysis:

1. Stating and refining the question
2. Exploring the data
3. Building formal statistical models
4. Interpreting the results
5. Communicating the results

## Epicycle of Analysis

1. Setting Expectations




**Epicycles of Analysis**

[Roger D. Peng & Elizabeth Matsui: The Art of Data Science]

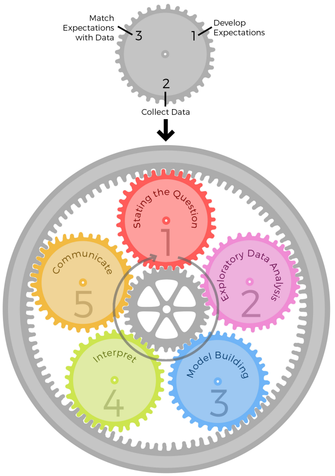
# (Epi)Cycles of Analysis

## There are 5 core activities of data analysis:

1. Stating and refining the question
2. Exploring the data
3. Building formal statistical models
4. Interpreting the results
5. Communicating the results

## Epicycle of Analysis

1. Setting Expectations
2. Collecting information (data), comparing the data to your expectations, and if the expectations don't match




**Epicycles of Analysis**

[Roger D. Peng & Elizabeth Matsui: The Art of Data Science]

# (Epi)Cycles of Analysis

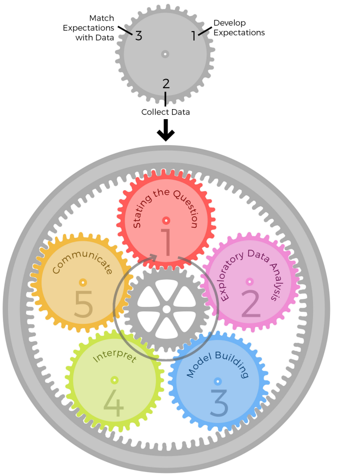
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

## There are 5 core activities of data analysis:

1. Stating and refining the question
2. Exploring the data
3. Building formal statistical models
4. Interpreting the results
5. Communicating the results

## Epicycle of Analysis

1. Setting Expectations
2. Collecting information (data), comparing the data to your expectations, and if the expectations don't match
3. Revising your expectations or fixing the data so your data and your expectations match.




**Epicycles of Analysis**

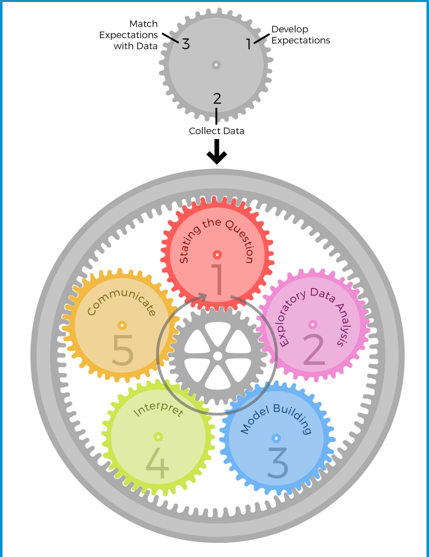
[Roger D. Peng & Elizabeth Matsui: The Art of Data Science]

## There are 5 core activities of data analysis:

1. State and refine the question
2. Explore the data
3. Build a model/code/analysis
4. Interpret the results
5. Communicate the results

## Epicycle of Analysis

1. Set expectations
2. Collect information/data
3. Compare the data to your expectations
4. If they don’t match: Revise your expectations or fix the code/analysis/model/data.




Epicycles of Analysis

[Roger D. Peng & Elizabeth Matsui: The Art of Data Science]

# Stating and Refining the Question

# Types of Questions



1. **Descriptive:** summarize data, descriptive statistics


[https://imgur.com/gallery/cYZVjTl]

# Types of Questions

1. **Descriptive:** summarize data, descriptive statistics
2. **Exploratory:** hypothesis-generating


[https://imgur.com/gallery/cYZVjTl]

# Types of Questions

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

1. **Descriptive:** summarize data, descriptive statistics
2. **Exploratory:** hypothesis-generating
3. **Inferential:** question about a particular hypothesis


[https://imgur.com/gallery/cYZVjTl]

# Types of Questions



1. **Descriptive:** summarize data, descriptive statistics
2. **Exploratory:** hypothesis-generating
3. **Inferential:** question about a particular hypothesis
4. **Predictive:** question about an unknown quantity without interest in the causes


[https://imgur.com/gallery/cYZVjTl]

# Types of Questions

1. **Descriptive:** summarize data, descriptive statistics
2. **Exploratory:** hypothesis-generating
3. **Inferential:** question about a particular hypothesis
4. **Predictive:** question about an unknown quantity without interest in the causes
5. **Causal:** question about causes


[https://imgur.com/gallery/cYZVjTl]

# Types of Questions

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

1. **Descriptive:** summarize data, descriptive statistics
2. **Exploratory:** hypothesis-generating
3. **Inferential:** question about a particular hypothesis
4. **Predictive:** question about an unknown quantity without interest in the causes
5. **Causal:** question about causes
6. **Mechanistic:** "how"-type questions


[https://imgur.com/gallery/cYZVjTl]

# Characteristics of a Good Question



A good question should be

- of interest to the audience (motivation)


[https://imgur.com/gallery/cYZVjTl]

# Characteristics of a Good Question



A good question should be

- of interest to the audience (motivation)
- not already be answered (gap in knowledge)


[https://imgur.com/gallery/cYZVjTl]

# Characteristics of a Good Question

A good question should be

- of interest to the audience (motivation)
- not already be answered (gap in knowledge)
- stem from a plausible framework (rationale)


[https://imgur.com/gallery/cYZVjTl]

# Characteristics of a Good Question

A good question should be

- of interest to the audience (motivation)
- not already be answered (gap in knowledge)
- stem from a plausible framework (rationale)
- answerable (feasibility)


[https://imgur.com/gallery/cYZVjTl]

# Characteristics of a Good Question

A good question should be

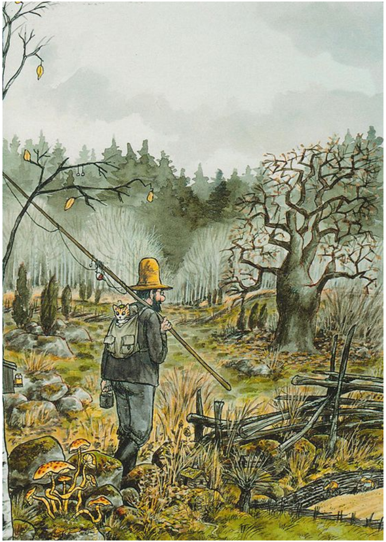
- of interest to the audience (motivation)
- not already be answered (gap in knowledge)
- stem from a plausible framework (rationale)
- answerable (feasibility)
- specific


[https://imgur.com/gallery/cYZVjTl]

# Translating a question into a data problem



- How will a good answer look like?

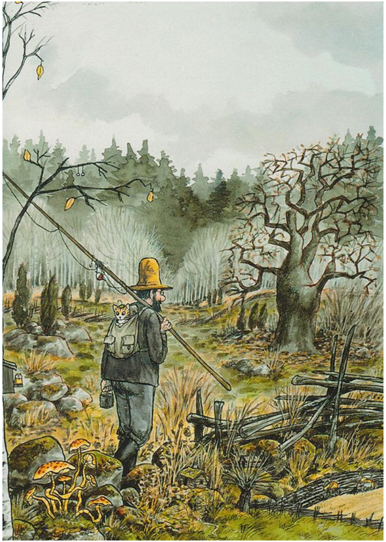


[https://www.flickr.com/photos/allthebeautifulthings/15518893546]

# Translating a question into a data problem



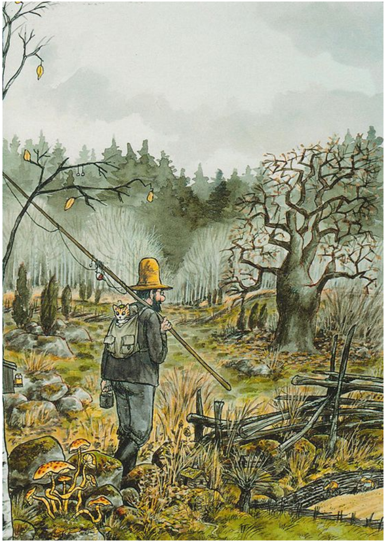
- How will a good answer look like?
- Will the answer be clear cut?



[https://www.flickr.com/photos/allthebeautifulthings/15518893546]

# Translating a question into a data problem

- How will a good answer look like?
- Will the answer be clear cut?
- Do I have the data to answer the question?

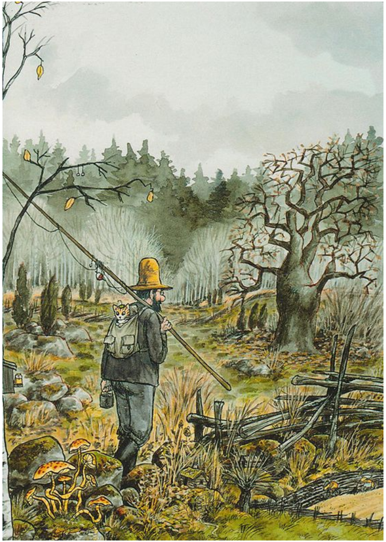


[https://www.flickr.com/photos/allthebeautifulthings/15518893546]

# Translating a question into a data problem



- How will a good answer look like?
- Will the answer be clear cut?
- Do I have the data to answer the question?
- Is the data biased?



[https://www.flickr.com/photos/allthebeautifulthings/15518893546]

# To the notebook!

## Goals

- Check whether there are problems with the dataset


[https://imgur.com/gallery/cYZVjTl]

# Exploratory Data Analysis



## Goals

- Check whether there are problems with the dataset
- Check whether the question can be answered with the dataset


[https://imgur.com/gallery/cYZVjTl]

# Exploratory Data Analysis



## Goals

- Check whether there are problems with the dataset
- Check whether the question can be answered with the dataset
- Develop a prototype of the answer/solution


[https://imgur.com/gallery/cYZVjTl]

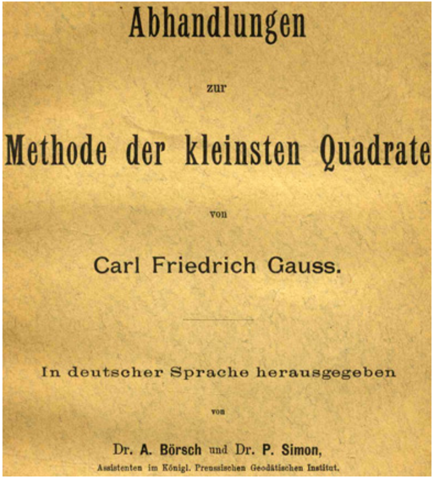
# Models & Inference

# Why models?



## Models can

- compress your data (the easiest model is no model)



Abhandlungen

zur

Methode der kleinsten Quadrate

von

Carl Friedrich Gauss.

In deutscher Sprache herausgegeben

von

Dr. A. Börsch und Dr. P. Simon,

Assistenten im Königl. Preussischen Geodätischen Institut.

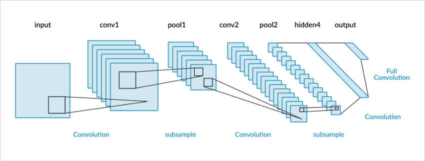
Epicycle of expectation-data-comparison applies as well!

# Why models?

## Models can

- compress your data (the easiest model is no model)
- allow you to extrapolate (predict)



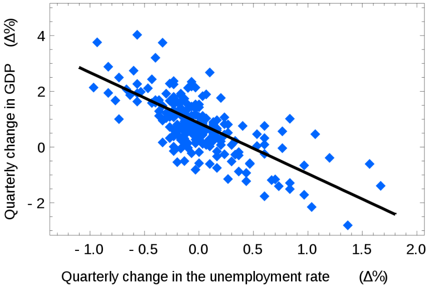

[https://medium.com/analytics-vidhya/your-handbook-to-convolutional-neural-networks-628782b68f7e]

**Epicycle of expectation-data-comparison applies as well!**

Models can

- compress your data (the easiest model is no model)
- allow you to extrapolate (predict)
- allow you to identify interpretable parameters (not all of them though)




[https://en.wikipedia.org/wiki/Simple_linear_regression]

Epicycle of expectation-data-comparison applies as well!

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Models can

- compress your data (the easiest model is no model)
- allow you to extrapolate (predict)
- allow you to identify interpretable parameters (not all of them though)
- allow you to deal with uncertainty

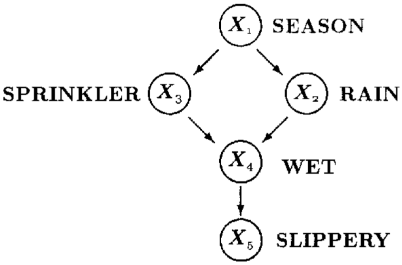
[Rev. Thomas Bayes (1701-1761)]

Epicycle of expectation-data-comparison applies as well!

# Why models?

Models can

- compress your data (the easiest model is no model)
- allow you to extrapolate (predict)
- allow you to identify interpretable parameters (not all of them though)
- allow you to deal with uncertainty
- can be a statistical description how your data was generated (expectations)




[Pearl 1997]

Epicycle of expectation-data-comparison applies as well!

# Why models?



Models can

- compress your data (the easiest model is no model)
- allow you to extrapolate (predict)
- allow you to identify interpretable parameters (not all of them though)
- allow you to deal with uncertainty
- can be a statistical description how your data was generated (expectations)
- it allows you to do inference (estimate an unknown quantity)

Epicycle of expectation-data-comparison applies as well!

# Models con't

We will talk more about models and inference later in the lecture, including questions like

- When is a model "good enough"?
- How to find parameters of a model?
- How to choose model?
- How to make decisions with models?
- What model to choose for which questions (predictive, inferential, causal, mechanistic)?

# Elements of the model epicycle

## Expectations

- Sketch a plot of what you expect

## Data

## Comparison and adjustment

# Elements of the model epicycle

## Expectations

- Sketch a plot of what you expect
- Think about what a model should and should not be able to do.

## Data

## Comparison and adjustment

# Elements of the model epicycle

## Expectations

- Sketch a plot of what you expect
- Think about what a model should and should not be able to do.

## Data

- Come up with toy data for which you have clear expectations.

## Comparison and adjustment

# Elements of the model epicycle

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

## Expectations

- Sketch a plot of what you expect
- Think about what a model should and should not be able to do.

## Data

- Come up with toy data for which you have clear expectations.
- Run model on smaller dataset first (faster).

## Comparison and adjustment

# Elements of the model epicycle

## Expectations

- Sketch a plot of what you expect
- Think about what a model should and should not be able to do.

## Data

- Come up with toy data for which you have clear expectations.
- Run model on smaller dataset first (faster).
- Deliberately “destroy” certain information in your data (shuffeling).

## Comparison and adjustment

# Elements of the model epicycle

## Expectations

- Sketch a plot of what you expect
- Think about what a model should and should not be able to do.

## Data

- Come up with toy data for which you have clear expectations.
- Run model on smaller dataset first (faster).
- Deliberately “destroy” certain information in your data (shuffeling).

## Comparison and adjustment

- Use the simplest model that does the job.

# Elements of the model epicycle



## Expectations

- Sketch a plot of what you expect
- Think about what a model should and should not be able to do.

## Data

- Come up with toy data for which you have clear expectations.
- Run model on smaller dataset first (faster).
- Deliberately “destroy” certain information in your data (shuffeling).

## Comparison and adjustment

- Use the simplest model that does the job.
- Always test on unseen data.

# Elements of the model epicycle



## Expectations

- Sketch a plot of what you expect
- Think about what a model should and should not be able to do.

## Data

- Come up with toy data for which you have clear expectations.
- Run model on smaller dataset first (faster).
- Deliberately “destroy” certain information in your data (shuffeling).

## Comparison and adjustment

- Use the simplest model that does the job.
- Always test on unseen data.
- Don’t “p-hack” (test until you have a significant result).

# Interpreting the results

# Principle of interpretation

- Revisit your original question

# Principle of interpretation



- Revisit your original question
- Check the nature of the result: directionality, magnitude, and uncertainty

# Principle of interpretation



- Revisit your original question
- Check the nature of the result: directionality, magnitude, and uncertainty
- Put your result into the context about what is known about the subject (a single result is next to meaningless)

# Principle of interpretation

- Revisit your original question
- Check the nature of the result: directionality, magnitude, and uncertainty
- Put your result into the context about what is known about the subject (a single result is next to meaningless)
- Think about possible controls (imagine your results wrong and try to explain it)

# Principle of interpretation

- Revisit your original question
- Check the nature of the result: directionality, magnitude, and uncertainty
- Put your result into the context about what is known about the subject (a single result is next to meaningless)
- Think about possible controls (imagine your results wrong and try to explain it)
- Think about implications and what actions (if any) need to be taken to answer your original question.

# Communication

# Elements of communication

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

- Communication craftsmanship (talks, writing, plots, ...)

# Elements of communication



- Communication craftsmanship (talks, writing, plots, ...)
- Routine communication as tool for data analysis (data/feedback acquisition)

# Elements of communication



- Communication craftsmanship (talks, writing, plots, ...)
- Routine communication as tool for data analysis (data/feedback acquisition)
- Communication of your results

# Communication as tool for data analysis

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

1. Select the right **audience** for the type of feedback you need.

Rachel Miller
@AllthingsIC

Love the #badstockphotosofmyjob hashtag. Here’s some for Communication professionals. Can’t remember the last time I used a megaphone in a client meeting, or anywhere! 🙈
8:31 AM - May 5, 2018
107 23 people are talking about this

[https://tinyurl.com/vrkj5wwn]

# Communication as tool for data analysis

1. Select the right **audience** for the type of feedback you need.
2. Make sure the **content** is concise and sufficient for the audience to understand the information you are presenting.


[https://tinyurl.com/vrkj5wwn]

# Communication as tool for data analysis



1. Select the right **audience** for the type of feedback you need.
2. Make sure the **content** is concise and sufficient for the audience to understand the information you are presenting.
3. **Style:** Avoid jargon and focus technical details/issues to technical audience only.


[https://tinyurl.com/vrkj5wwn]

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

1. Select the right **audience** for the type of feedback you need.
2. Make sure the **content** is concise and sufficient for the audience to understand the information you are presenting.
3. **Style:** Avoid jargon and focus technical details/issues to technical audience only.
4. Make sure to have a collaborative **attidude** and be open to constructive feedback.


[https://tinyurl.com/vrkj5wwn]

Thanks for listening.
Questions?


[https://tinyurl.com/vrkj5wwn]

# References

# Source for this lecture