TIPS & TRICKS FOR GETTING THE MOST VALUE OUT OF NUMERICAL DATA

Lars Ole Schwen

MEVIS Oberseminar, 2018-10-23

DatA Stats Focus

EDUCATION

Ten quick tips for getting the most scientific value out of numerical data

Lars Ole Schwen[1*], Sabrina Rueschenbaum[2*]

1 Fraunhofer MEVIS, Am Fallturm 1, Bremen, Germany, 2 Department of Internal Medicine 1, University Hospital Frankfurt, Goethe University, Theodor-Stern-Kai 7, Frankfurt (Main), Germany

* These authors contributed equally to this work.
* ole.schwen@mevis.fraunhofer.de

Abstract

Most studies in the life sciences and other disciplines involve generating and analyzing numerical data of some type as the foundation for scientific findings. Working with numerical data involves multiple challenges. These include reproducible data acquisition, appropriate data storage, computationally correct data analysis, appropriate reporting and presentation

Fraunhofer
MEVIS

TIPS & TRICKS FOR GETTING THE MOST VALUE OUT OF NUMERICAL DATA

A Data Analyst's Workplace

Ole Schwen: Getting the Most Value out of Numerical Data

1. Introduction

A Data Analyst's Workplace

usually: coffee $\longrightarrow$ Mathematician $\longrightarrow$ theorems

Fraunhofer
MEVIS

1. Introduction

A Data Analyst's Workplace

here: raw data } ⟶ [Data Analyst] ⟶ meaning and value
       tea

1. Introduction

A Data Analyst's Workplace

tools for Pomodoro technique,
cf. https://francescocirillo.com/pages/pomodoro-technique or ask Diego or Hendrik

A Data Analyst's Workplace

you all know what this is

A Data Analyst's Workplace

debugging tool, cf. http://www.rubberduckdebugging.com
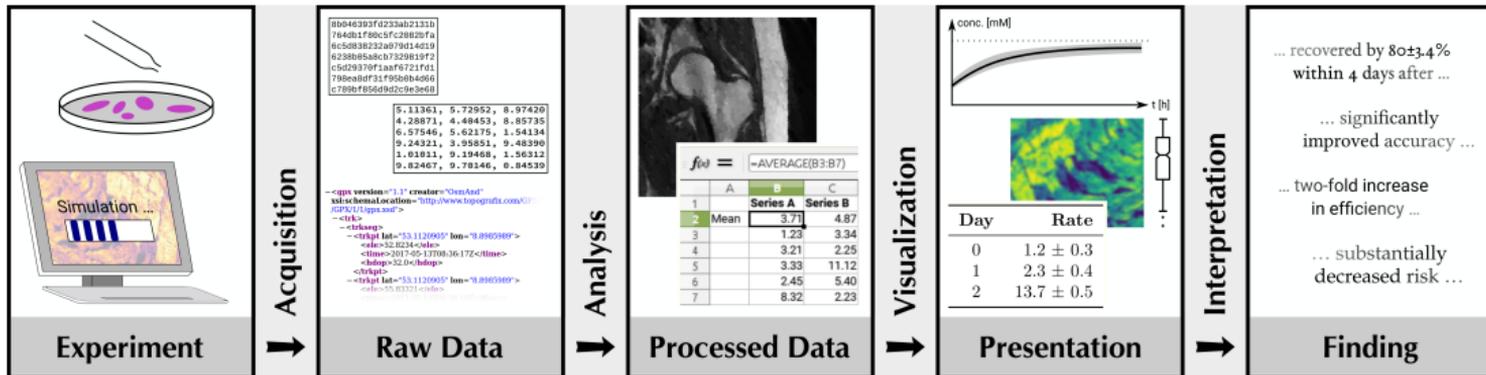

Fraunhofer
MEVIS

1. Introduction

A Data Analyst's Workplace

multifunctional penguin, here: plausibility checking tool and pointer to implausibilities

1. Introduction

Typical Scientific Workflow

Experiment		Raw Data		Processed Data		Presentation		Finding
	Acquisition		Analysis		Visualization		Interpretation	

Fraunhofer
MEVIS

Basics

- numerical data frequently basis for findings

1. Introduction
Basics

- numerical data frequently basis for findings
- acquisition, processing, and presentation need to be correct and reproducible

1. Introduction
Basics

- numerical data frequently basis for findings
- acquisition, processing, and presentation need to be correct and reproducible
- mistakes are annoying, but hard to avoid

1. Introduction
Basics

- numerical data frequently basis for findings
- acquisition, processing, and presentation need to be correct and reproducible
- mistakes are annoying, but hard to avoid
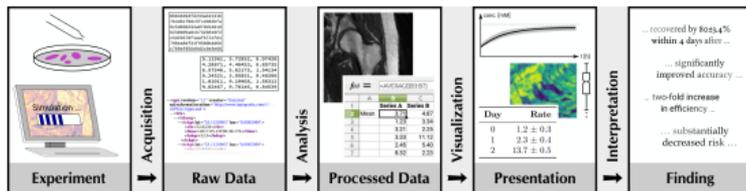- try to make life easy for yourself and others

2. **Keep Data Correct**

Contents

Experiment	➡	Raw Data	➡	Processed Data	➡	Presentation	➡	Finding

Tip 1: Keep Your Data Correct

Tip 1: Keep Your Data Correct

Watch out for surprising patterns

Plausibility Check!

Tip 1: Keep Your Data Correct

Watch out for surprising patterns

- two out of every seven values zero
  - weekend measurements missing?

Plausibility Check!

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Watch out for surprising patterns

- two out of every seven values zero
  - weekend measurements missing?
- count of colony-forming units 15, 46, 28, 46, 52, and 46

Plausibility Check!

Fraunhofer
MEVIS

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Watch out for surprising patterns
- two out of every seven values zero
  - weekend measurements missing?
- count of colony-forming units 15, 46, 28, 46, 52, and 46
  - mix-up in counts?

Plausibility Check!

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Watch out for surprising patterns

- two out of every seven values zero
  - weekend measurements missing?
- count of colony-forming units 15, 46, 28, 46, 52, and 46
  - mix-up in counts?
- time points $t_1 = 16\,421.392\,\text{s}$ and $t_2 = 20\,021.392\,\text{s}$

Plausibility Check!

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Watch out for surprising patterns

- two out of every seven values zero
  - weekend measurements missing?
- count of colony-forming units 15, 46, 28, 46, 52, and 46
  - mix-up in counts?
- time points $t_1 = 16\,421.392\,\text{s}$ and $t_2 = 20\,021.392\,\text{s}$
  - measurement every hour or computed value?

Plausibility Check!

Fraunhofer
MEVIS

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Watch out for surprising patterns

- two out of every seven values zero
  - weekend measurements missing?
- count of colony-forming units 15, 46, 28, 46, 52, and 46
  - mix-up in counts?
- time points $t_1 = 16\,421.392\,\text{s}$ and $t_2 = 20\,021.392\,\text{s}$
  - measurement every hour or computed value?
- discuss data with other person

Plausibility Check!

Fraunhofer
MEVIS

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Watch out for surprising patterns

- two out of every seven values zero
  - weekend measurements missing?
- count of colony-forming units 15, 46, 28, 46, 52, and 46
  - mix-up in counts?
- time points $t_1 = 16\,421.392\,\text{s}$ and $t_2 = 20\,021.392\,\text{s}$
  - measurement every hour or computed value?
- discuss data with other person
  (or explain it to your penguin)

Plausibility Check!

Fraunhofer
MEVIS

Tip 1: Keep Your Data Correct

Practice guesstimation

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Practice guesstimation

- simplified assumptions and estimates
- back-of-the-envelope calculations
- examples below …
- Further reading: https://what-if.xkcd.com

Tip 1: Keep Your Data Correct

Retrospective example: Penguin growth

Fraunhofer
MEVIS

Tip 1: Keep Your Data Correct

Retrospective example: Penguin growth

- observed growth of new-born Humboldt penguins for one year.

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Retrospective example: Penguin growth

- observed growth of new-born Humboldt penguins for one year.
- Observation: 2 inches per week.

Fraunhofer
MEVIS

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Retrospective example: Penguin growth

- observed growth of new-born Humboldt penguins for one year.
- Observation: 2 inches per week.
- Plausible?

Plausibility Check!

Fraunhofer MEVIS

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Retrospective example: Penguin growth

- observed growth of new-born Humboldt penguins for one year.
- Observation: 2 inches per week.
- Plausible?

✗ 50 weeks × 5 cm/week = 2.5 m

Plausibility Check!

Fraunhofer
MEVIS

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Retrospective example: Penguin growth

- observed growth of new-born Humboldt penguins for one year.
- Observation: 2 inches per week.
- Plausible?

✗ 50 weeks × 5 cm/week = 2.5 m

Plausibility Check!

Other example:

One way ▾ 1 passenger ▾ Economy ▾				
📍 Toronto	⇄ 📍 Sydney YQY		📅 Fri, Nov 9	‹ ›
🍁	10:55 AM – 2:16 PM Air Canada · Operated by Air Canada Express - Sky Reg…	2h 21m YYZ–YQY	Nonstop	€155 ⌄

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Retrospective example: Penguin growth

- observed growth of new-born Humboldt penguins for one year.
- Observation: 2 inches per week.
- Plausible?

✗ 50 weeks × 5 cm/week = 2.5 m

Plausibility Check!

Other example:

One way ▾	1 passenger ▾	Economy ▾		
📍 Toronto	⇄	📍 Sydney YQY		📅 Fri, Nov 9 ‹ ›

	10:55 AM – 2:16 PM Air Canada · Operated by Air Canada Express - Sky Reg…	2h 21m YYZ–YQY	Nonstop	€155	⌄

Sydney, Nova Scotia, is probably nice—but not in Australia

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Prospective example: Amount of data for time-lapse video of new building

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Prospective example: Amount of data for time-lapse video of new building

- 2 years, i.e., 1000 days

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Prospective example: Amount of data for time-lapse video of new building

- 2 years, i.e., 1000 days
- image every minute, i.e., $25 \times 60 = 1500$ per day

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Prospective example: Amount of data for time-lapse video of new building

- 2 years, i.e., 1000 days
- image every minute, i.e., $25 \times 60 = 1500$ per day
- 2 MiB each

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Prospective example: Amount of data for time-lapse video of new building

- 2 years, i.e., 1000 days
- image every minute, i.e., $25 \times 60 = 1500$ per day
- 2 MiB each
- 3 TiB in total

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Prospective example: Amount of data for time-lapse video of new building

- 2 years, i.e., 1000 days
- image every minute, i.e., $25 \times 60 = 1500$ per day
- 2 MiB each
- 3 TiB in total
- upper limit

2. Keep Data Correct

Tip 1: Keep Your Data Correct

Prospective example: Amount of data for time-lapse video of new building

- 2 years, i.e., 1000 days
- image every minute, i.e., $25 \times 60 = 1500$ per day
- 2 MiB each
- 3 TiB in total
- upper limit

Disclaimer: Just guessed, I do not know what IT + Building team estimated.

2. Keep Data Correct

Tip 2: Track Your Sources

Tip 2: Track Your Sources

Pitfalls

Tip 2: Track Your Sources

Pitfalls

- Undocumented way how or in which context a number was measured

2. Keep Data Correct
Tip 2: Track Your Sources

Pitfalls

- Undocumented way how or in which context a number was measured
- Will you find "the snail paper by Smith" in two weeks?

2. Keep Data Correct

Tip 2: Track Your Sources

Pitfalls

- Undocumented way how or in which context a number was measured
- Will you find "the snail paper by Smith" in two weeks?
- Will you find $v_{mean} = 1.27$ millimeters per second if it was originally written as $V_{average} = 3$ inches per minute?

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did
- you might find mistakes

Fraunhofer
MEVIS

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes
- This is "like a version of Murphy's Law: Everything you do, you will probably have to do over again." (W. S. Noble: DOI 10.1371/journal.pcbi.1000424)

Fraunhofer
MEVIS

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure
- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes
- This is "like a version of Murphy's Law: Everything you do, you will probably have to do over again." (W. S. Noble: DOI 10.1371/journal.pcbi.1000424)

Separate data from formulas

Fraunhofer
MEVIS

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure

- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes
- This is "like a version of Murphy's Law: Everything you do, you will probably have to do over again." (W. S. Noble: DOI 10.1371/journal.pcbi.1000424)

Separate data from formulas

- scripts (R, Python, …) better than spreadsheets

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure
- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes
- This is "like a version of Murphy's Law: Everything you do, you will probably have to do over again." (W. S. Noble: DOI 10.1371/journal.pcbi.1000424)

Separate data from formulas
- scripts (R, Python, …) better than spreadsheets
- also avoids data overwriting and formula range problems

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure
- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes
- This is "like a version of Murphy's Law: Everything you do, you will probably have to do over again." (W. S. Noble: DOI 10.1371/journal.pcbi.1000424)

Separate data from formulas
- scripts (R, Python, …) better than spreadsheets
- also avoids data overwriting and formula range problems
- document what you did

Fraunhofer
MEVIS

2. Keep Data Correct

Tip 2: Track Your Sources

Automatize data analysis: evaluation should be one-step procedure
- fully documents what you did
- you might find mistakes
- advisors, co-authors, and reviewers tend to request changes
- This is "like a version of Murphy's Law: Everything you do, you will probably have to do over again." (W. S. Noble: DOI 10.1371/journal.pcbi.1000424)

Separate data from formulas
- scripts (R, Python, …) better than spreadsheets
- also avoids data overwriting and formula range problems
- document what you did
- version-control evaluation (and manuscripts and …)

Tip 3: Beware of Computational Pitfalls

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding
- Unavailable data points: properly left out? zero? counted or left out?

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding
- Unavailable data points: properly left out? zero? counted or left out?
- Integer division: "`84/96` cases OK" always produced zero

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding
- Unavailable data points: properly left out? zero? counted or left out?
- Integer division: "`84/96` cases OK" always produced zero
- Non-finite values (NaN)

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding
- Unavailable data points: properly left out? zero? counted or left out?
- Integer division: "84/96 cases OK" always produced zero
- Non-finite values (NaN)
  - always compare false

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding
- Unavailable data points: properly left out? zero? counted or left out?
- Integer division: "`84/96` cases OK" always produced zero
- Non-finite values (NaN)
  - always compare false
  - mess up sorting etc.

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding
- Unavailable data points: properly left out? zero? counted or left out?
- Integer division: "84/96 cases OK" always produced zero
- Non-finite values (NaN)
  - always compare false
  - mess up sorting etc.
  - Example: (standard Python `sort`)

Input data									
x	0.0	8.0	0.0	8.0	4.0	1.0	4.0	0.4	4.5
y	1.0	1.6	0.0	2.0	0.0	—	1.6	0.4	1.5

Fraunhofer
MEVIS

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding
- Unavailable data points: properly left out? zero? counted or left out?
- Integer division: "84/96 cases OK" always produced zero
- Non-finite values (NaN)
  - always compare false
  - mess up sorting etc.
  - Example: (standard Python `sort`)

Input data									
x	0.0	8.0	0.0	8.0	4.0	1.0	4.0	0.4	4.5
y	1.0	1.6	0.0	2.0	0.0	—	1.6	0.4	1.5
Analysis results (intermediate and final)									
Values x/y	0.0	5.0	nan	4.0	inf	nan	2.5	1.0	3.0

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding
- Unavailable data points: properly left out? zero? counted or left out?
- Integer division: "84/96 cases OK" always produced zero
- Non-finite values (NaN)
  - always compare false
  - mess up sorting etc.
  - Example: (standard Python `sort`)

Input data									
x	0.0	8.0	0.0	8.0	4.0	1.0	4.0	0.4	4.5
y	1.0	1.6	0.0	2.0	0.0	—	1.6	0.4	1.5
Analysis results (intermediate and final)									
Values *x/y*	0.0	5.0	nan	4.0	inf	nan	2.5	1.0	3.0
Sorted values	0.0	1.0	3.0	5.0	nan	4.0	inf	nan	2.5

2. Keep Data Correct

Tip 3: Beware of Computational Pitfalls

- Copying and pasting values: prevents one-step procedure, may introduce rounding
- Unavailable data points: properly left out? zero? counted or left out?
- Integer division: "84/96 cases OK" always produced zero
- Non-finite values (NaN)
  - always compare false
  - mess up sorting etc.
  - Example: (standard Python `sort`)

Input data									
x	0.0	8.0	0.0	8.0	4.0	1.0	4.0	0.4	4.5
y	1.0	1.6	0.0	2.0	0.0	—	1.6	0.4	1.5
Analysis results (intermediate and final)									
Values *x/y*	0.0	5.0	nan	4.0	inf	nan	2.5	1.0	3.0
Sorted values	0.0	1.0	3.0	5.0	nan	4.0	inf	nan	2.5
Maximum	3.0								
Mean	NaN								

Fraunhofer
MEVIS

3. Correctly Interpret Your Data
Contents

| Experiment | ➡ | Raw Data | ➡ | Processed Data | ➡ | Presentation | ➡ | Finding |

Tip 4: Treat Units With Respect

Tip 4: Treat Units With Respect

- What is "temperature in the 30s"?

Fraunhofer
MEVIS
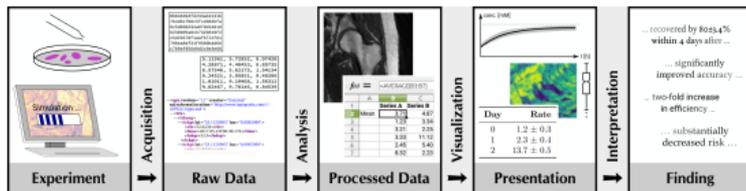
3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

- What is "temperature in the 30s"?
  - summer day at 30°C in Paris
  - expect snow at 30°F in Seattle
  - superconductor experiment at 30 K

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

- What is "temperature in the 30s"?
    - summer day at 30°C in Paris
    - expect snow at 30°F in Seattle
    - superconductor experiment at 30 K
- "A [mathematical] model without units is not a model." (Matthias König)

Tip 4: Treat Units With Respect

Use SI (Base) Units in computations

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use SI (Base) Units in computations

- Mistakes of factor 1000 easy to spot, 60 or 2.54 more difficult

3. Correctly Interpret Your Data
Tip 4: Treat Units With Respect

Use SI (Base) Units in computations

- Mistakes of factor 1000 easy to spot, 60 or 2.54 more difficult

or else …

- Mars Climate Orbiter crashed

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use SI (Base) Units in computations

- Mistakes of factor 1000 easy to spot, 60 or 2.54 more difficult

or else …

- Mars Climate Orbiter crashed
- reason: using pound seconds instead of Newton seconds as the unit of impulse

Tip 4: Treat Units With Respect

Use SI (Base) Units in computations

- Mistakes of factor 1000 easy to spot, 60 or 2.54 more difficult

or else …

- Mars Climate Orbiter crashed
- reason: using pound seconds instead of Newton seconds as the unit of impulse
- consequence: multimillion dollar mission failed

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m hard to read

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m hard to read
  - erythrocyte diameter is 7 µm

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m hard to read
  - erythrocyte diameter is 7 μm
- human water consumption of 3×10^{-8} m³/s difficult to interpret and misleading

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m hard to read
  - erythrocyte diameter is 7 μm
- human water consumption of 3×10^{-8} m³/s difficult to interpret and misleading
  - 2.6 liters per day more intuitive

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m hard to read
    - erythrocyte diameter is 7 μm
- human water consumption of 3×10^{-8} m³/s difficult to interpret and misleading
    - 2.6 liters per day more intuitive
- 8×10^{-8} m² fuel consumption of a car

Fraunhofer
MEVIS

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m hard to read
  - erythrocyte diameter is 7 µm
- human water consumption of 3×10^{-8} m³/s difficult to interpret and misleading
  - 2.6 liters per day more intuitive
- 8×10^{-8} m² fuel consumption of a car
  - correct SI unit and interesting to interpret, but hard to judge

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Use Intuitive Units for plausibility-checking and in presentation

- 0.000 007 m hard to read
  - erythrocyte diameter is 7 µm
- human water consumption of 3×10^{-8} m³/s difficult to interpret and misleading
  - 2.6 liters per day more intuitive
- 8×10^{-8} m² fuel consumption of a car
  - correct SI unit and interesting to interpret, but hard to judge
  - 8 liters per 100 kilometers (or fuel efficiency of 29.4 miles per gallon) better

Know SI Prefixes

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Know SI Prefixes

- Is pico 10^{-9}, 10^{-12}, or 10^{-15}?

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Know SI Prefixes
- Is pico 10^{-9}, 10^{-12}, or 10^{-15}?
- ppb: parts per billion, but how much is that?

Tip 4: Treat Units With Respect

Know SI Prefixes
- Is pico 10^{-9}, 10^{-12}, or 10^{-15}?
- ppb: parts per billion, but how much is that?
  - one billion = Eine Billion?

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Know SI Prefixes

- Is pico 10^{-9}, 10^{-12}, or 10^{-15}?
- ppb: parts per billion, but how much is that?
  - one billion = Eine Billion?
- calories vs. kcal

Fraunhofer
MEVIS

Units in Formulas

Units in Formulas

- help plausibility-checking and understanding

3. Correctly Interpret Your Data
Tip 4: Treat Units With Respect

Units in Formulas
- help plausibility-checking and understanding
- "two meters plus half a second" does not make sense

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Units in Formulas

- help plausibility-checking and understanding
- "two meters plus half a second" does not make sense
- differentiation corresponds to division by unit

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Units in Formulas

- help plausibility-checking and understanding
- "two meters plus half a second" does not make sense
- differentiation corresponds to division by unit
  - e.g., velocity in units meters per second is the temporal gradient (one over seconds) times position (meters)

Fraunhofer
MEVIS

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Units in Formulas

- help plausibility-checking and understanding
- "two meters plus half a second" does not make sense
- differentiation corresponds to division by unit
  - e.g., velocity in units meters per second is the temporal gradient (one over seconds) times position (meters)
- integration corresponds to multiplication by unit

Interpreting Formulas via Units

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Interpreting Formulas via Units

- influence of CYP enzymes on the metabolism of theobromine via Michaelis–Menten equation

$$\frac{\mathrm{d}\,c(t)}{\mathrm{d}\,t} = \dot{c}(t) = -\frac{V \cdot c(t)}{K_m + c(t)} \tag{1}$$

Fraunhofer
MEVIS

3. Correctly Interpret Your Data
Tip 4: Treat Units With Respect

Interpreting Formulas via Units

- influence of CYP enzymes on the metabolism of theobromine via Michaelis–Menten equation

$$\frac{\mathrm{d}\,c(t)}{\mathrm{d}\,t} = \dot{c}(t) = -\frac{V \cdot c(t)}{K_m + c(t)} \tag{1}$$

- concentration c either mass or molar amount per volume

Fraunhofer
MEVIS

3. Correctly Interpret Your Data
Tip 4: Treat Units With Respect

Interpreting Formulas via Units

- influence of CYP enzymes on the metabolism of theobromine via Michaelis–Menten equation

$$\frac{d\,c(t)}{d\,t} = \dot{c}(t) = -\frac{V \cdot c(t)}{K_m + c(t)} \tag{1}$$

- concentration c either mass or molar amount per volume
- addition in the denominator implies that K_m is a constant also in units of concentration.

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Interpreting Formulas via Units

- influence of CYP enzymes on the metabolism of theobromine via Michaelis–Menten equation

$$\frac{\mathrm{d}\,c(t)}{\mathrm{d}\,t} = \dot{c}(t) = -\frac{V \cdot c(t)}{K_m + c(t)} \tag{1}$$

- concentration c either mass or molar amount per volume
- addition in the denominator implies that K_m is a constant also in units of concentration.
- factor c (concentration) in numerator implies that V has units of concentration per time, i.e., a rate of concentration change.

Tip 4: Treat Units With Respect

Separate Unit Conversion and Formula Logic

3. Correctly Interpret Your Data

Tip 4: Treat Units With Respect

Separate Unit Conversion and Formula Logic, or else …

$$\dot{c}(t) = -\frac{\frac{1\,\text{min}\cdot 1\,\text{mol}\cdot\tilde{V}}{60\,\text{s}\cdot 180.164\,\text{g}} \cdot c(t)}{\frac{1\,\text{min}\cdot 1\,\text{mol}\cdot\tilde{K}_m}{60\,\text{s}\cdot 180.164\,\text{g}} + c(t)} \tag{2}$$

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Asymptotic Behavior: what happens in formula $f(x)$ if

- $x \to \pm\infty$ or limits of useful ranges

Tip 5: Verify Your Formulas

Check Asymptotic Behavior: what happens in formula $f(x)$ if

- $x \to \pm\infty$ or limits of useful ranges
- $x \to 0$

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Asymptotic Behavior: what happens in formula $f(x)$ if

- $x \to \pm\infty$ or limits of useful ranges
- $x \to 0$
- differences $\to 0$

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Asymptotic Behavior: what happens in formula $f(x)$ if

- $x \rightarrow \pm\infty$ or limits of useful ranges
- $x \rightarrow 0$
- differences $\rightarrow 0$
- denominators $\rightarrow 0$

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Asymptotic Behavior: what happens in formula $f(x)$ if

- $x \to \pm\infty$ or limits of useful ranges
- $x \to 0$
- differences $\to 0$
- denominators $\to 0$
- What is the monotonicity in these cases?

Check Individual Constants and Variables

Fraunhofer
MEVIS

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Individual Constants and Variables

Recall Michaelis–Menten example:

$$\dot{c}(t) = -\frac{V \cdot c(t)}{K_m + c(t)} \tag{3}$$

V and K_m are positive constants.

- $c > 0$: negative (i.e., decrease)

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Individual Constants and Variables

Recall Michaelis–Menten example:

$$\dot{c}(t) = -\frac{V \cdot c(t)}{K_m + c(t)} \tag{3}$$

V and K_m are positive constants.

- $c > 0$: negative (i.e., decrease)
- $c = 0$: no change

Fraunhofer
MEVIS

3. Correctly Interpret Your Data
Tip 5: Verify Your Formulas

Check Individual Constants and Variables

Recall Michaelis–Menten example:

$$\dot{c}(t) = -\frac{V \cdot c(t)}{K_m + c(t)} \tag{3}$$

V and K_m are positive constants.

- $c > 0$: negative (i.e., decrease)
- $c = 0$: no change
- c large: obtain V (i.e., maximal rate)

Fraunhofer
MEVIS

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Individual Constants and Variables

Recall Michaelis–Menten example:

$$\dot{c}(t) = -\frac{V \cdot c(t)}{K_m + c(t)} \tag{3}$$

V and K_m are positive constants.

- $c > 0$: negative (i.e., decrease)
- $c = 0$: no change
- c large: obtain V (i.e., maximal rate)
- larger V: larger decay if all other values stay the same

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Check Individual Constants and Variables

Recall Michaelis–Menten example:

$$\dot{c}(t) = -\frac{V \cdot c(t)}{K_m + c(t)} \tag{3}$$

V and K_m are positive constants.

- $c > 0$: negative (i.e., decrease)
- $c = 0$: no change
- c large: obtain V (i.e., maximal rate)
- larger V: larger decay if all other values stay the same
- K_m more difficult to interpret: $K_m = c$ implies half maximal rate

Tip 5: Verify Your Formulas

Scaling
Example: testing interaction of n peripheral devices (pairwise interaction)

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Scaling

Example: testing interaction of *n* peripheral devices (pairwise interaction)

- *n* times procurement and customs bureaucracy

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Scaling

Example: testing interaction of *n* peripheral devices (pairwise interaction)

- *n* times procurement and customs bureaucracy
- $n(n-1)/2 = (n^2 - n)/2$ interactions to be tested

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Scaling

Example: testing interaction of n peripheral devices (pairwise interaction)

- n times procurement and customs bureaucracy
- $n(n-1)/2 = (n^2 - n)/2$ interactions to be tested
- eventually, n^2 effort will dominate

Respect Monotonicity

Tip 5: Verify Your Formulas

Respect Monotonicity

Suppose you quantify soil contamination based on presence of metallophytes.

Respect Monotonicity

Suppose you quantify soil contamination based on presence of metallophytes.

✗ soil quality index (high presence $\not\Rightarrow$ high quality)

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Respect Monotonicity

Suppose you quantify soil contamination based on presence of metallophytes.

✘ soil quality index (high presence $\not\Rightarrow$ high quality)

✔ soil pollution index (high presence $\Rightarrow$ high pollution)

3. Correctly Interpret Your Data

Tip 5: Verify Your Formulas

Respect Monotonicity

Suppose you quantify soil contamination based on presence of metallophytes.

- ✘ soil quality index (high presence $\not\Rightarrow$ high quality)
- ✔ soil pollution index (high presence $\Rightarrow$ high pollution)

In our work

- error measure (to be minimized)
  vs.
- quality measure (to be maximized)

Tip 6: Know Thy Statistical Methods

Correlation Coefficients
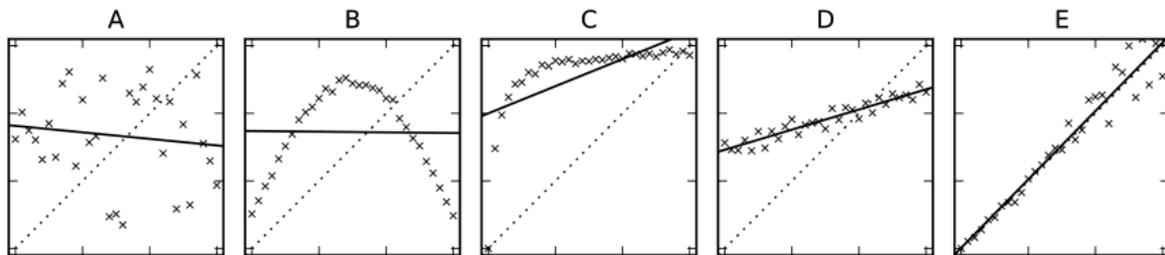
Correlation Coefficients

Tip 6: Know Thy Statistical Methods

Correlation Coefficients
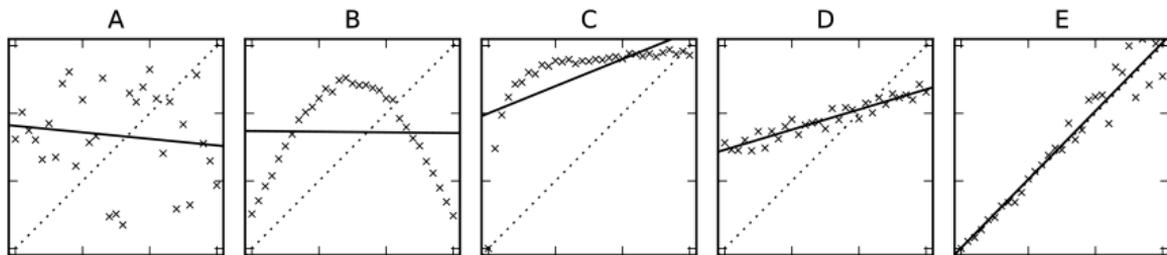
Data Series	Spearman CC	Pearson CC	Concordance CC
(a) noise	−0.108	−0.128	−0.120
(b) non-monotonic	−0.019	−0.013	−0.012
(c) monotonic, non-linear	**0.955**	0.639	0.280
(d) linear	0.951	**0.948**	0.459
(e) identical	0.971	0.972	**0.971**

Correlation is not Causation

Tip 6: Know Thy Statistical Methods

Correlation is not Causation

- decreasing number of pirates and global warming
  [B. Henderson, The Gospel of the Flying Spaghetti Monster]

Tip 6: Know Thy Statistical Methods

Correlation is not Causation

- decreasing number of pirates and global warming
  [B. Henderson, The Gospel of the Flying Spaghetti Monster]
- Nicholas Cage movies and people drowning in swimming pools

Tip 6: Know Thy Statistical Methods

Correlation is not Causation

- decreasing number of pirates and global warming
  [B. Henderson, The Gospel of the Flying Spaghetti Monster]
- Nicholas Cage movies and people drowning in swimming pools
- German steel production and SPD results in elections

Statistical Significance is

Fraunhofer
MEVIS

Tip 6: Know Thy Statistical Methods

Statistical Significance is

✔ a high probability that an effect (e.g., difference) is not just due to randomness

Tip 6: Know Thy Statistical Methods

Statistical Significance is

✔ a high probability that an effect (e.g., difference) is not just due to randomness

✘ not proof of the above

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Statistical Significance is

✔ a high probability that an effect (e.g., difference) is not just due to randomness

✘ not proof of the above

✘ not a measure how big the effect is

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Statistical Significance is

✔ a high probability that an effect (e.g., difference) is not just due to randomness

✘ not proof of the above

✘ not a measure how big the effect is

✘ not a check whether the test is applicable in the first place

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Statistical Significance is

✔ a high probability that an effect (e.g., difference) is not just due to randomness

✘ not proof of the above

✘ not a measure how big the effect is

✘ not a check whether the test is applicable in the first place

✘ lack of significance $\neq$ no meaningful difference

Tip 6: Know Thy Statistical Methods

Statistical Significance is

✔ a high probability that an effect (e.g., difference) is not just due to randomness

✘ not proof of the above

✘ not a measure how big the effect is

✘ not a check whether the test is applicable in the first place

✘ lack of significance $\not\Rightarrow$ no meaningful difference

✔ statistical significance

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Statistical Significance is

✔ a high probability that an effect (e.g., difference) is not just due to randomness

✘ not proof of the above

✘ not a measure how big the effect is

✘ not a check whether the test is applicable in the first place

✘ lack of significance $\not\Rightarrow$ no meaningful difference

✔ statistical significance

✘ not general/biological/… significance

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Statistical Significance is

✔ a high probability that an effect (e.g., difference) is not just due to randomness

✘ not proof of the above

✘ not a measure how big the effect is

✘ not a check whether the test is applicable in the first place

✘ lack of significance $\neq$ no meaningful difference

✔ statistical significance

✘ not general/biological/… significance

✘ not a measure of importance/meaning/relevance

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database (Mann–Whitney u tests)

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database (Mann–Whitney u tests)

Description	Group A			Group B			approx. p-value
	group	n	Med	group	n	Med	
Birth Month b-CS	June			Sep.			1.32×10^{-4}

Med: Median, b-CS: batting/caught stealing, f-GS: fielding/games started, b-SH: batting/sacrifice hits, b-H: batting/hits, b-G: batting/games, CA: California, PA: Pennsylvania, MA: Massachusetts

Fraunhofer
MEVIS

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database (Mann–Whitney u tests)

Description	group	Group A n	Med	group	Group B n	Med	approx. p-value
Birth Month b-CS	June	277	9.0	Sep.	348	5.0	1.32×10^{-4}

Med: Median, b-CS: batting/caught stealing, f-GS: fielding/games started, b-SH: batting/sacrifice hits, b-H: batting/hits, b-G: batting/games, CA: California, PA: Pennsylvania, MA: Massachusetts

Fraunhofer
MEVIS

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database (Mann–Whitney u tests)

Description	Group A			Group B			approx. p-value
	group	n	Med	group	n	Med	
Birth Month							
b-CS	June	277	9.0	Sep.	348	5.0	1.32×10^{-4}
f-GS	June	471	76.0	Sep.	638	43.5	1.95×10^{-3}

Med: Median, b-CS: batting/caught stealing, f-GS: fielding/games started, b-SH: batting/sacrifice hits, b-H: batting/hits, b-G: batting/games, CA: California, PA: Pennsylvania, MA: Massachusetts

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database (Mann–Whitney u tests)

Description	Group A			Group B			approx. p-value
	group	n	Med	group	n	Med	
Birth Month							
b-CS	June	277	9.0	Sep.	348	5.0	1.32×10^{-4}
f-GS	June	471	76.0	Sep.	638	43.5	1.95×10^{-3}
Death Day							
b-SH	day 9	168	19.5	day 27	168	6.0	3.02×10^{-5}
b-H	day 9	248	56.5	day 27	250	23.0	3.95×10^{-5}

Med: Median, b-CS: batting/caught stealing, f-GS: fielding/games started, b-SH: batting/sacrifice hits, b-H: batting/hits, b-G: batting/games, CA: California, PA: Pennsylvania, MA: Massachusetts

Fraunhofer
MEVIS

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database (Mann–Whitney *u* tests)

Description	group	Group A n	Med	group	Group B n	Med	approx. p-value
Birth Month							
b-CS	June	277	9.0	Sep.	348	5.0	1.32×10^{-4}
f-GS	June	471	76.0	Sep.	638	43.5	1.95×10^{-3}
Death Day							
b-SH	day 9	168	19.5	day 27	168	6.0	3.02×10^{-5}
b-H	day 9	248	56.5	day 27	250	23.0	3.95×10^{-5}
Birth State							
b-G	CA	2182	121.5	PA	660	33.0	6.74×10^{-27}
b-G	CA	2182	121.5	MA	1410	52.0	1.54×10^{-25}

Med: Median, b-CS: batting/caught stealing, f-GS: fielding/games started, b-SH: batting/sacrifice hits, b-H: batting/hits, b-G: batting/games, CA: California, PA: Pennsylvania, MA: Massachusetts

Fraunhofer
MEVIS

Analysis of the Lahman Baseball Database: Lessons learned

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database: Lessons learned

- I do not know what these baseball terms mean.
- ⇒ Also domain non-experts[1] can find seemingly "highly significant" results

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database: Lessons learned

- I do not know what these baseball terms mean.
- ⇒ Also domain non-experts[1] can find seemingly "highly significant" results

  [1] i.e., any idiot

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database: Lessons learned

- I do not know what these baseball terms mean.
- ⇒ Also domain non-experts[1] can find seemingly "highly significant" results

    [1] i.e., any idiot, e.g., me

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database: Lessons learned

- I do not know what these baseball terms mean.
- ⇒ Also domain non-experts[1] can find seemingly "highly significant" results

    [1] i.e., any idiot, e.g., me

- This is misuse of reasonable statistical methods for "data dredging" or "*p*-value hacking"

3. Correctly Interpret Your Data

Tip 6: Know Thy Statistical Methods

Analysis of the Lahman Baseball Database: Lessons learned

- I do not know what these baseball terms mean.
- ⇒ Also domain non-experts[1] can find seemingly "highly significant" results

  [1] i.e., any idiot, e.g., me

- This is misuse of reasonable statistical methods for "data dredging" or "*p*-value hacking"
- At least, Bonferroni correction should be applied (rendering most results above non-significant)

3. Correctly Interpret Your Data
Tip 7: Keep Track of Accuracy

- 68.432 853 percent of all statistics pretend to be more accurate than they actually are

Error Sources and Propagation

Fraunhofer
MEVIS

3. Correctly Interpret Your Data
Tip 7: Keep Track of Accuracy

Error Sources and Propagation

- systematic errors, limited resolution, small samples,
- random errors (focus here)

3. Correctly Interpret Your Data

Tip 7: Keep Track of Accuracy

Error Sources and Propagation

- systematic errors, limited resolution, small samples,
- random errors (focus here)
- errors propagate through calculations

Error Sources and Propagation

- systematic errors, limited resolution, small samples,
- random errors (focus here)
- errors propagate through calculations
- results should not be based on noise

Example of Uncertainty Quantification

Tip 7: Keep Track of Accuracy

Example of Uncertainty Quantification

(low-tech) measurements of ground speed of white storks

Tip 7: Keep Track of Accuracy

Example of Uncertainty Quantification

(low-tech) measurements of ground speed of white storks

- distance $1\,km \pm 10\,\%$
- time $100\,s \pm 1\,\%$

Plausibility Check!

3. Correctly Interpret Your Data

Tip 7: Keep Track of Accuracy

Example of Uncertainty Quantification

(low-tech) measurements of ground speed of white storks

- distance 1 km ± 10 %
- time 100 s ± 1 %
- velocity uncertainty $\sqrt{10^2 + 1^2} = 10.05$ %, assuming no correlation and normal distribution
- distance error dominates

3. Correctly Interpret Your Data

Tip 7: Keep Track of Accuracy

Example of Uncertainty Quantification

(low-tech) measurements of ground speed of white storks

- distance 1 km ± 10 %
- time 100 s ± 1 %
- velocity uncertainty $\sqrt{10^2 + 1^2} = 10.05$ %, assuming no correlation and normal distribution
- distance error dominates

Monte-Carlo Simulations

- repeat data evaluation many times with input data sampled from distribution estimating uncertainty

3. Correctly Interpret Your Data

Tip 7: Keep Track of Accuracy

Example of Uncertainty Quantification

(low-tech) measurements of ground speed of white storks

- distance 1 km ± 10 %
- time 100 s ± 1 %
- velocity uncertainty $\sqrt{10^2 + 1^2} = 10.05 \%$, assuming no correlation and normal distribution
- distance error dominates

Monte-Carlo Simulations

- repeat data evaluation many times with input data sampled from distribution estimating uncertainty

3. Correctly Interpret Your Data

Tip 7: Keep Track of Accuracy

Example of Uncertainty Quantification

(low-tech) measurements of ground speed of white storks

- distance 1 km ± 10 %
- time 100 s ± 10 %
- velocity uncertainty $\sqrt{10^2 + 10^2} = 14.36\,\%$, assuming no correlation and normal distribution
- distance error dominates

Monte-Carlo Simulations

- repeat data evaluation many times with input data sampled from distribution estimating uncertainty
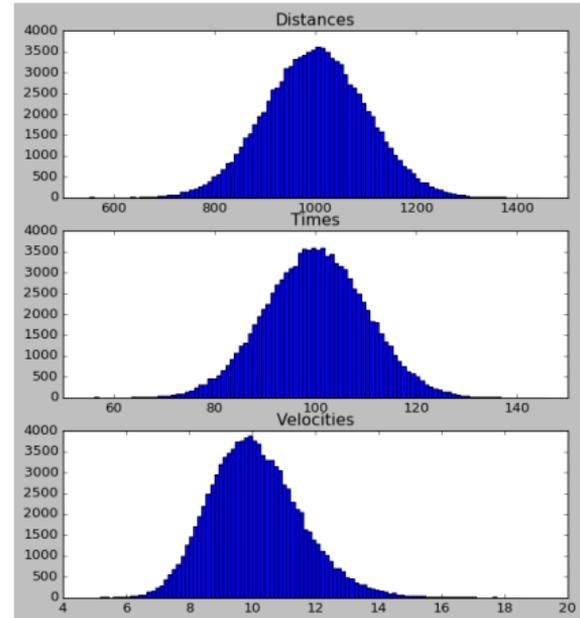
Tip 8: Don't Rely on Numbers Alone

3. Correctly Interpret Your Data

Tip 8: Don't Rely on Numbers Alone

(x, y) data series A to D:

	Series A	Series B	Series C	Series D
$\bar{x}$	9.0	9.0	9.0	9.0
s_x^2	11.0	11.0	11.0	11.0
$\bar{y}$	7.5	7.5	7.5	7.5
s_y^2	4.127	4.128	4.122	4.123
$\rho_{x,y}$	0.816	0.816	0.816	0.817

- These (x, y) data series are pretty much the same.

Tip 8: Don't Rely on Numbers Alone

(x, y) data series A to D:

	Series A	Series B	Series C	Series D
$\bar{x}$	9.0	9.0	9.0	9.0
s_x^2	11.0	11.0	11.0	11.0
$\bar{y}$	7.5	7.5	7.5	7.5
s_y^2	4.127	4.128	4.122	4.123
$\rho_{x,y}$	0.816	0.816	0.816	0.817

- These (x, y) data series are pretty much the same.

Or are they?

Tip 8: Don't Rely on Numbers Alone
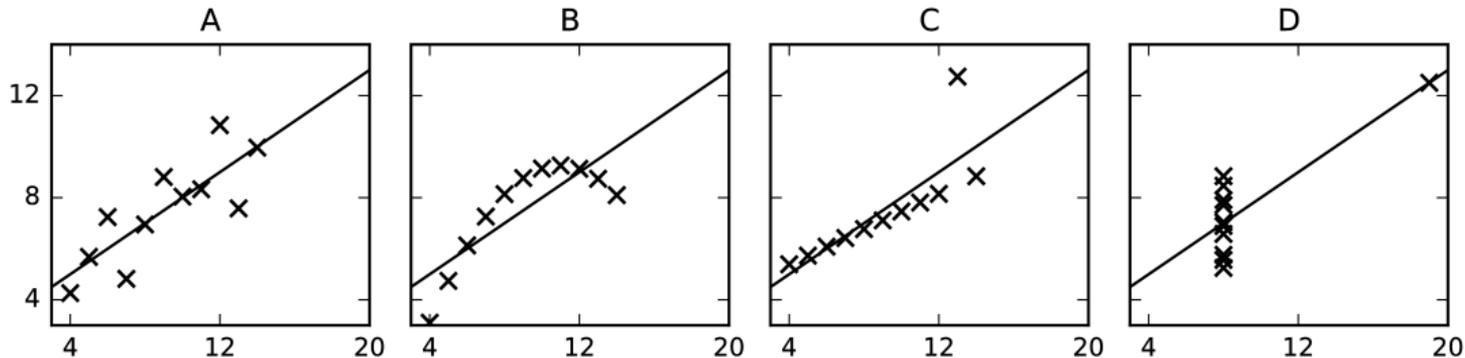
- Anscombe's quartet [Anscombe, 1973]

3. Correctly Interpret Your Data

Tip 8: Don't Rely on Numbers Alone

Let's count dandelions. Automatically, of course.

Tip 8: Don't Rely on Numbers Alone

Let's count dandelions. Automatically, of course.

- Training/Calibration: This is one flower.
  Black rod (flashlight) is a length reference

3. Correctly Interpret Your Data

Tip 8: Don't Rely on Numbers Alone

Let's count dandelions. Automatically, of course.

- Training/Calibration: This also is one flower.

Tip 8: Don't Rely on Numbers Alone

Let's count dandelions. Automatically, of course.

Continue training and validation, maybe cross-validation …

3. Correctly Interpret Your Data

Tip 8: Don't Rely on Numbers Alone

Let's count dandelions. Automatically, of course.

- Analysis result: 4 and 5 dandelion flowers, respectvely.

3. Correctly Interpret Your Data

Tip 8: Don't Rely on Numbers Alone

Let's count dandelions. Automatically, of course.

- Analysis result: 4 and 5 dandelion flowers, respectvely.

Lessons Learned so far

3. Correctly Interpret Your Data

Tip 8: Don't Rely on Numbers Alone

Lessons Learned so far

- don't blindly trust in numbers or algorithmic results
- assess the actual data
- summarizing data is simplification and loses information

Tip 8: Don't Rely on Numbers Alone

Lessons Learned so far

- don't blindly trust in numbers or algorithmic results
- assess the actual data
- summarizing data is simplification and loses information

Tips:

- quick & dirty plots OK for data assessment
- choose the right way to look at it (scatter plots, time series, spatial data)

Don't Over-Interpret Numbers

3. Correctly Interpret Your Data

Tip 8: Don't Rely on Numbers Alone

Don't Over-Interpret Numbers

- One number means one thing, don't put too much meaning into it.

Tip 8: Don't Rely on Numbers Alone

Don't Over-Interpret Numbers

- One number means one thing, don't put too much meaning into it.
  - body-mass index is not generic health measure
  - impact factor is not quality measure

Fraunhofer
MEVIS

Tip 8: Don't Rely on Numbers Alone

Don't Over-Interpret Numbers

- One number means one thing, don't put too much meaning into it.
  - body-mass index is not generic health measure
  - impact factor is not quality measure
- Don't calculate just because something is a number

Tip 8: Don't Rely on Numbers Alone

Don't Over-Interpret Numbers

- One number means one thing, don't put too much meaning into it.
  - body-mass index is not generic health measure
  - impact factor is not quality measure
- Don't calculate just because something is a number
  - temperature –20°C is not "twice as cold as" –10°C

3. Correctly Interpret Your Data

Tip 8: Don't Rely on Numbers Alone

Don't Over-Interpret Numbers

- One number means one thing, don't put too much meaning into it.
  - body-mass index is not generic health measure
  - impact factor is not quality measure
- Don't calculate just because something is a number
  - temperature –20°C is not "twice as cold as" –10°C
  - correlation coefficient 0.9 is not "50 % better than" 0.6

Tip 8: Don't Rely on Numbers Alone

Don't Over-Interpret Numbers

- One number means one thing, don't put too much meaning into it.
  - body-mass index is not generic health measure
  - impact factor is not quality measure
- Don't calculate just because something is a number
  - temperature −20°C is not "twice as cold as" −10°C
  - correlation coefficient 0.9 is not "50 % better than" 0.6
  - prostate carcinoma with Gleason score 4 is not "twice as large/dangerous/… as" score 2

Fraunhofer
MEVIS

Tip 8: Don't Rely on Numbers Alone

Don't Over-Interpret Numbers

- One number means one thing, don't put too much meaning into it.
  - body-mass index is not generic health measure
  - impact factor is not quality measure
- Don't calculate just because something is a number
  - temperature –20°C is not "twice as cold as" –10°C
  - correlation coefficient 0.9 is not "50 % better than" 0.6
  - prostate carcinoma with Gleason score 4 is not "twice as large/dangerous/… as" score 2
- Distinguish extensive and intensive quantities
  - 0.5 l water at 30°C and 0.5 l at 50°C is 1 l at 40°C

4. Present Your Data in a Useful Way
Contents

| Experiment | ➡ | Raw Data | ➡ | Processed Data | ➡ | Presentation | ➡ | Finding |
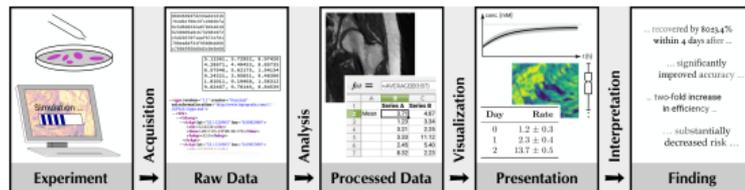
4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

Absolute/Relative Values/Differences

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

Absolute/Relative Values/Differences

- body temperature 37.6°C (absolute)
- overweight person (125 kg) has lost 10 kg (absolute difference)

4. Present Your Data in a Useful Way
Tip 9: Properly Present Data

Absolute/Relative Values/Differences

- body temperature 37.6°C (absolute)
- overweight person (125 kg) has lost 10 kg (absolute difference)
- wheat cultivation with 20 % higher yields (relative difference)

Tip 9: Properly Present Data

Absolute/Relative Values/Differences

- body temperature 37.6°C (absolute)
- overweight person (125 kg) has lost 10 kg (absolute difference)
- wheat cultivation with 20 % higher yields (relative difference)

Q: What is the baseline?

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

Absolute/Relative Values/Differences

- body temperature 37.6°C (absolute)
- overweight person (125 kg) has lost 10 kg (absolute difference)
- wheat cultivation with 20 % higher yields (relative difference)

Q: What is the baseline?

- about 37°C
- about 70 kg
- about 6 t/ha

Fraunhofer
MEVIS

Tip 9: Properly Present Data

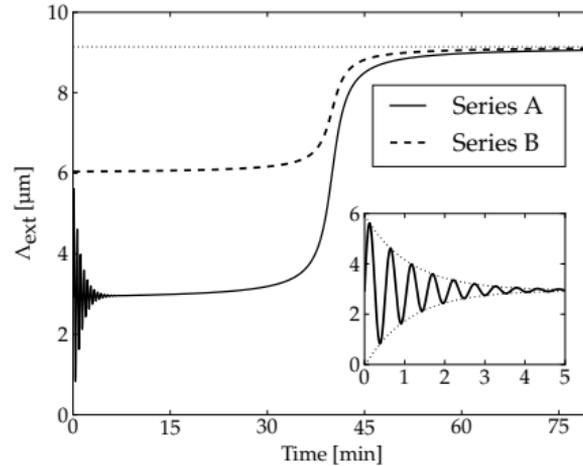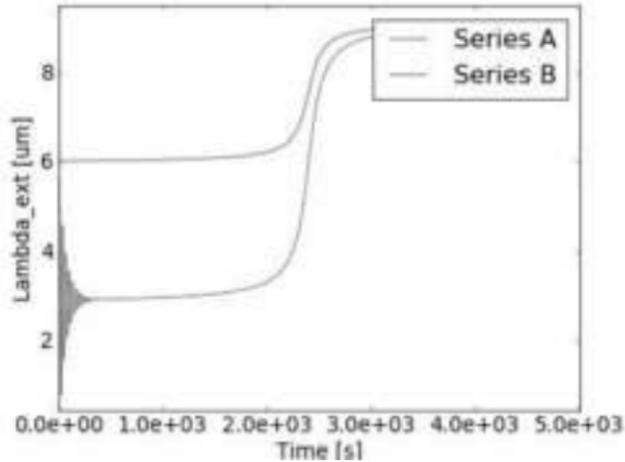**How to Present Data?**

- plot
  - linear vs. logarithmic
  - absolute vs. relative differences
  - interpretation of parallel lines

Tip 9: Properly Present Data

How to Present Data?

- plot—but that is not the only way
  - linear vs. logarithmic
  - absolute vs. relative differences
  - interpretation of parallel lines

Tip 9: Properly Present Data

How to Present Data?

- plot
  - linear vs. logarithmic
  - absolute vs. relative differences
  - interpretation of parallel lines
- tables

Tip 9: Properly Present Data

How to Present Data?

- plot
  - linear vs. logarithmic
  - absolute vs. relative differences
  - interpretation of parallel lines
- tables
- number in text

Tip 9: Properly Present Data

How to Present Data?

- plot
    - linear vs. logarithmic
    - absolute vs. relative differences
    - interpretation of parallel lines
- tables
- number in text
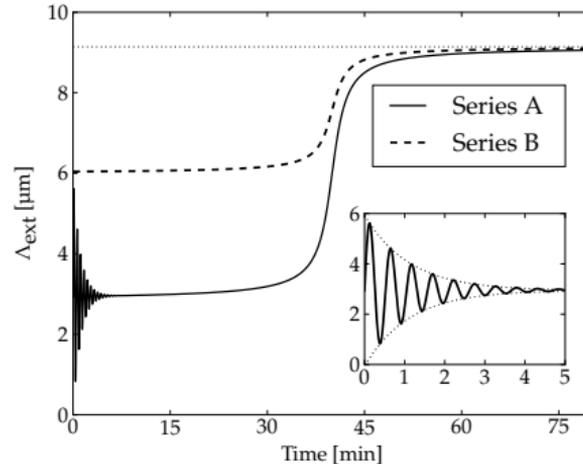- spatial or other visualization

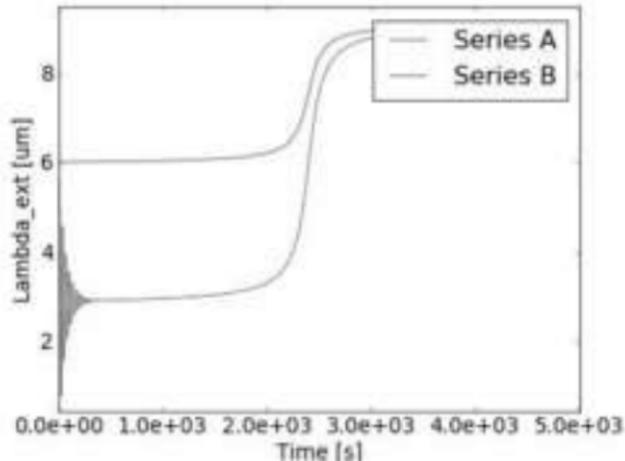# 4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

Ugly vs. Nice Plot: What is wrong on the left?

4. Present Your Data in a Useful Way

Tip 9: Properly Present Data

Ugly vs. Nice Plot: What is wrong on the left?

A lot: text ugly, numbers hard to read and not intuitive, curves not distinguishable, legend occludes data, oscillation not visible, value range tries but fails to start at zero, image is blurred and low-quality jpeg, relevant information missing

Color Scales

Color Scales

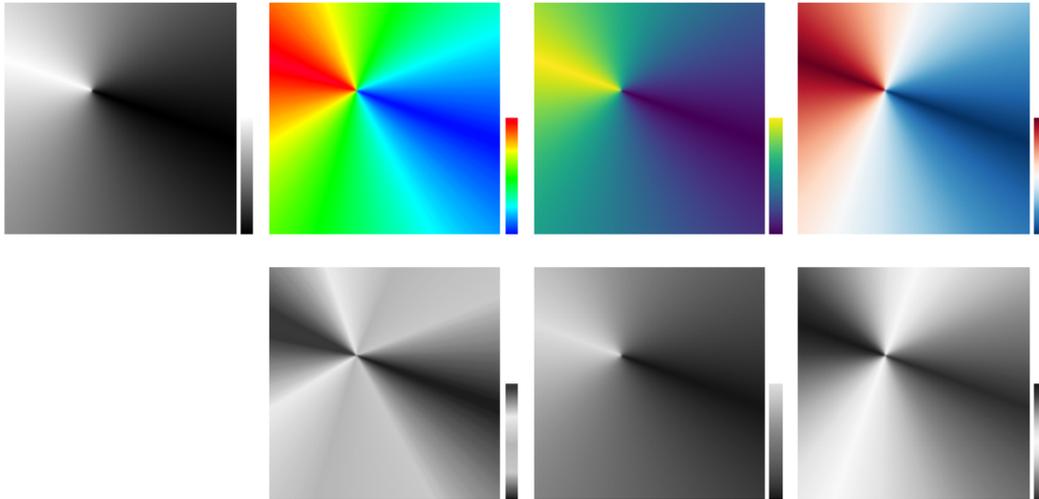

- distinguish sequential and divergent (and categorical) color scales
- consider color-blind people and crappy b/w printers
- avoid rainbow colormap

Fraunhofer
MEVIS

Tip 10: Share Your Data

Tip 10: Share Your Data

Why?

Tip 10: Share Your Data

Why?

- increase credibility
- funding agencies like "open data"
- ensures long-term availability (good scientific practice)
- some journals require sharing data

Tip 10: Share Your Data

Why?

- increase credibility
- funding agencies like "open data"
- ensures long-term availability (good scientific practice)
- some journals require sharing data
- preparing (cleaning up, standardizing, …) and documenting data may reveal mistakes

4. Present Your Data in a Useful Way

Tip 10: Share Your Data

Why?

- increase credibility
- funding agencies like "open data"
- ensures long-term availability (good scientific practice)
- some journals require sharing data
- preparing (cleaning up, standardizing, …) and documenting data may reveal mistakes
- people might use the data
  - yourself at different lab
  - collaborations with others might evolve
  - others might find gems you will never go looking for

How?

4. Present Your Data in a Useful Way

Tip 10: Share Your Data

How?

- own, institutional, project website (not optimal)
- domain-specialized or general purpose data repositories
- supporting information

4. Present Your Data in a Useful Way

Tip 10: Share Your Data

How?

- own, institutional, project website (not optimal)
- domain-specialized or general purpose data repositories
- supporting information

- all data and python code for this paper available as supplement

4. Present Your Data in a Useful Way

Tip 10: Share Your Data

How?

- own, institutional, project website (not optimal)
- domain-specialized or general purpose data repositories
- supporting information

- all data and python code for this paper available as supplement
- We don't believe anyone can make use of it, but …

How?

- own, institutional, project website (not optimal)
- domain-specialized or general purpose data repositories
- supporting information

- all data and python code for this paper available as supplement
- We don't believe anyone can make use of it, but …

    "… we are looking forward to being proven wrong." [Schwen and Rueschenbaum 2018]

5. Summary

Tips & Tricks for Getting Most Value out of Numerical Data

1. Keep Your Data Correct
2. Track Your Sources
3. Beware of Computational Pitfalls
4. Treat Units With Respect
5. Verify Your Formulas

6. Know Thy Statistical Methods
7. Keep Track of Accuracy
8. Don't Rely on Numbers Alone
9. Properly Present Data
10. Share Your Data

Experiment → Raw Data → Processed Data → Presentation → Finding

Acquisition → Analysis → Visualization → Interpretation

Fraunhofer MEVIS