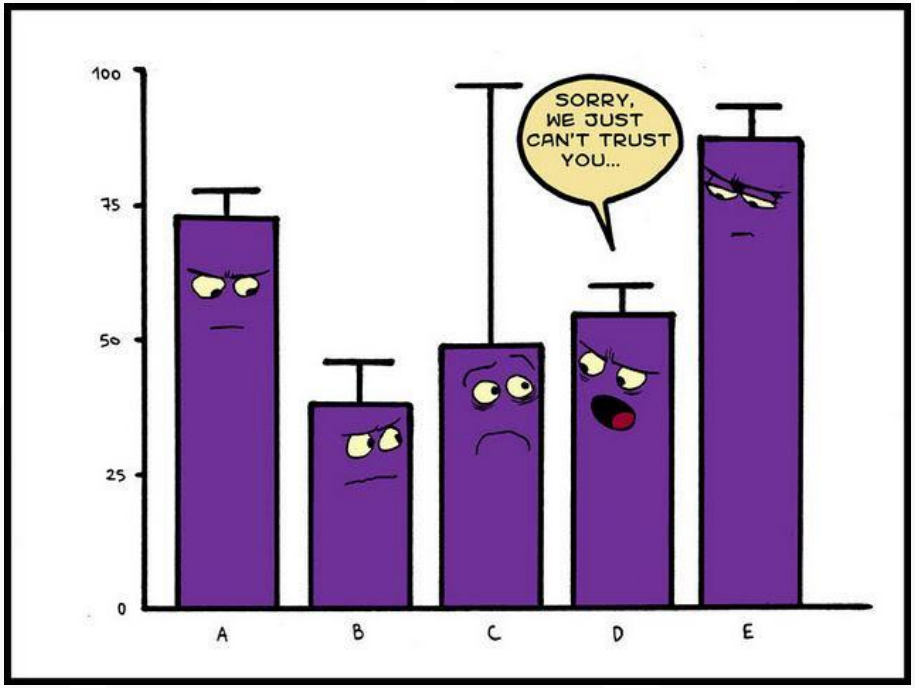
What can go wrong with statistics: Some typical errors & How to lie with statistics



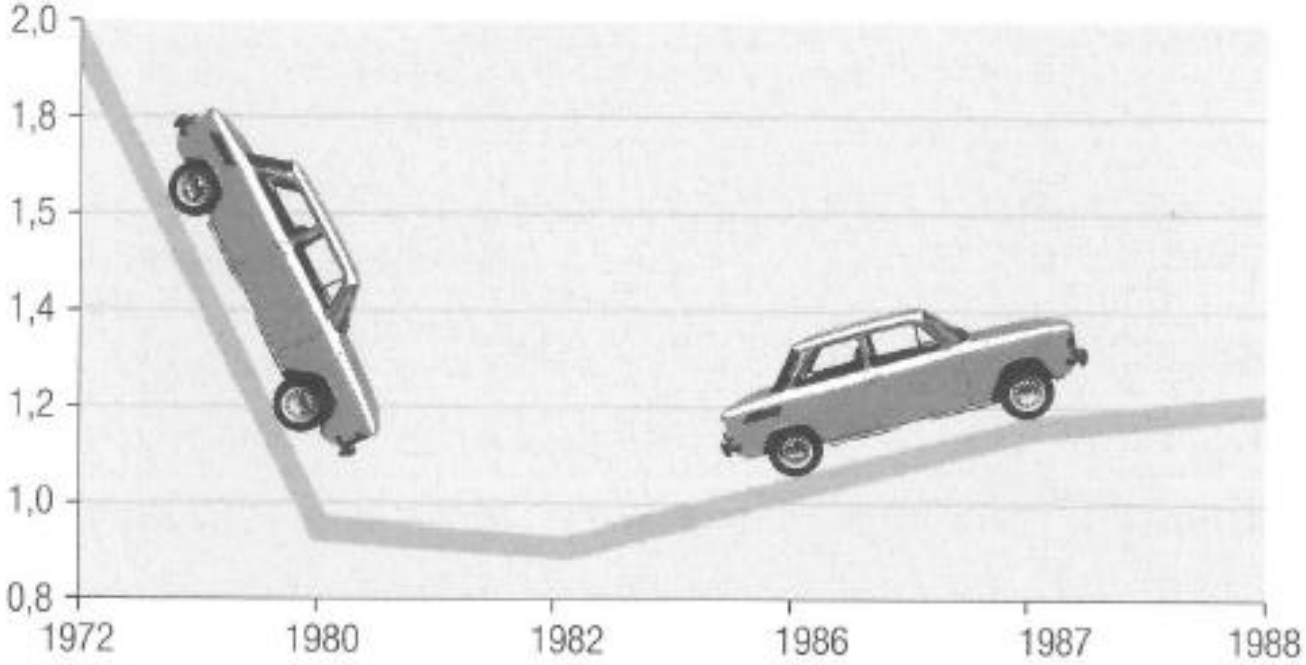
Content adopted partially from:

Lutz Prechelt
Daniel Huff
Jon Hasenbank
Gerd Bosbach /
Jens Jürgen Korff

Some Starters – Russian Election 2012

Some Starters – Car Production in UK

Produktion von Personenwagen in Großbritannien 1972 bis 1988 (Angaben in Mio.)



Die rasante Talfahrt der britischen Autoproduktion kam durch Manipulation der x-Achse zustande. Nach Walter Krämer: So lügt man mit Statistik.[2]

Motivation

"There are three kinds of lies:
Lies, Damned Lies, and Statistics."
– attributed to Benjamin Disraeli

- Statistics are commonly used to make a point or back-up one's position
  - 82.7% of all statistics are made up on the spot.

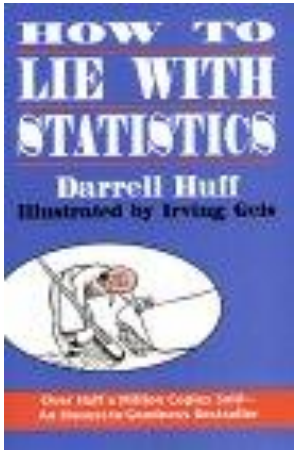
- Three sources of errors:
  - If done in manipulative way, statistics can be deceiving
  - If not done carefully, statistics can be deceiving
    - Inadvertent methodological errors and / or wrong assumptions also will fool the person who is doing the statistics!
  - If not read carefully, statistics can be deceiving

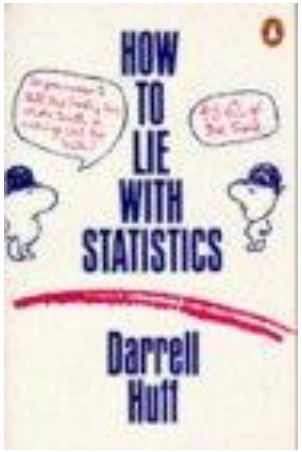
Purpose of this section

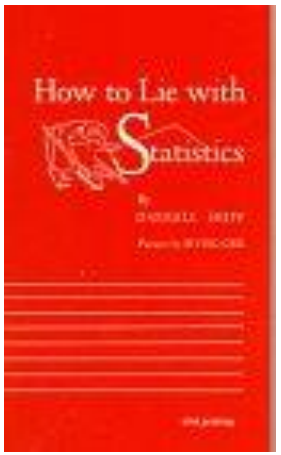
- Avoid common inadvertent errors
  - “Lessons for author”
- Be aware of the subtle tricks that others may play on you
  - (and that you should never play on others!)
  - “Lessons for reader”

Source #1

- Large parts of this slide set is based on ideas from Darrell Huff: *How to Lie With Statistics*,
  (Victor Gollancz 1954, Pelican Books 1973, Penguin Books 1991)
  - but the slides use different examples
  - Most slides made by Lutz Prechelt
  - The book is short (120 p.), entertaining, and insightful
  - Many different editions available
  - Other, similar books exist as well



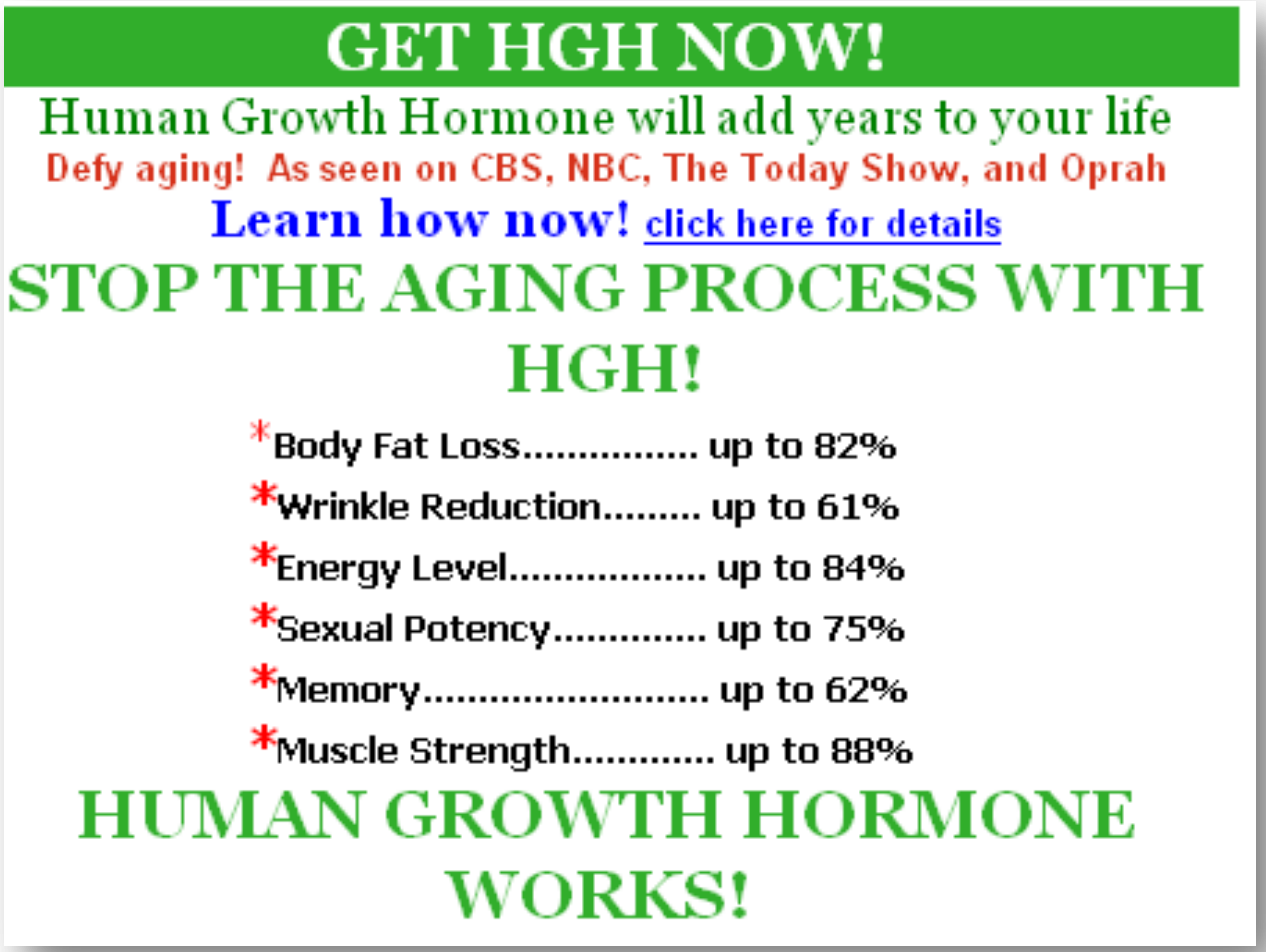




Source #2

- Other source of ideas:
  Gerd Bosbach, Jens Jürgen Korff: *Lügen mit Zahlen*
  (Heyne-Verlag, 2. Auflage, 2011)
  - The book is very readable and entertaining
  - You may notice strong political opinions – sometimes you might ask yourselves if the book does not itself use the power of numbers and graphs to manipulate the reader…

Example: Human Growth Hormone Spam (HGH)



Remark

- We use this real spam email as an arbitrary example
- and will make **unwarranted** assumptions about what is behind it
  - for illustrative purposes
  - I do not claim that HGH treatment is useful, useless, or harmful

Note:

- HGH is on the IOC doping list
  - http://www.dshs-koeln.de/biochemie/rubriken/01_doping/06.html
  - *"Für die therapeutische Anwendung von HGH kommen derzeit nur zwei wesentliche Krankheitsbilder in Frage: Zwergwuchs bei Kindern und HGH-Mangel beim Erwachsenen"*
  - *"Die Wirksamkeit von HGH bei Sportlern muss allerdings bisher stark in Frage gestellt werden, da bisher keine wissenschaftliche Studie zeigen konnte, dass eine zusätzliche HGH-Applikation bei Personen, die eine normale HGH-Produktion aufweisen, zu Leistungssteigerungen führen kann."*

Problem 1: What do they mean?

- "Body fat loss: up to 82%"
  - OK, can be measured
- "Wrinkle reduction: up to 61%"
  - Maybe they count the wrinkles and measure their depth?
- "Energy level: up to 84%"
  - What is this?
  - Also note they use language loosely:
    - Loss in percent: OK; reduction in percent: OK
    - Level in percent??? (should be 'increase')

Lesson for readers: What did they actually measure?

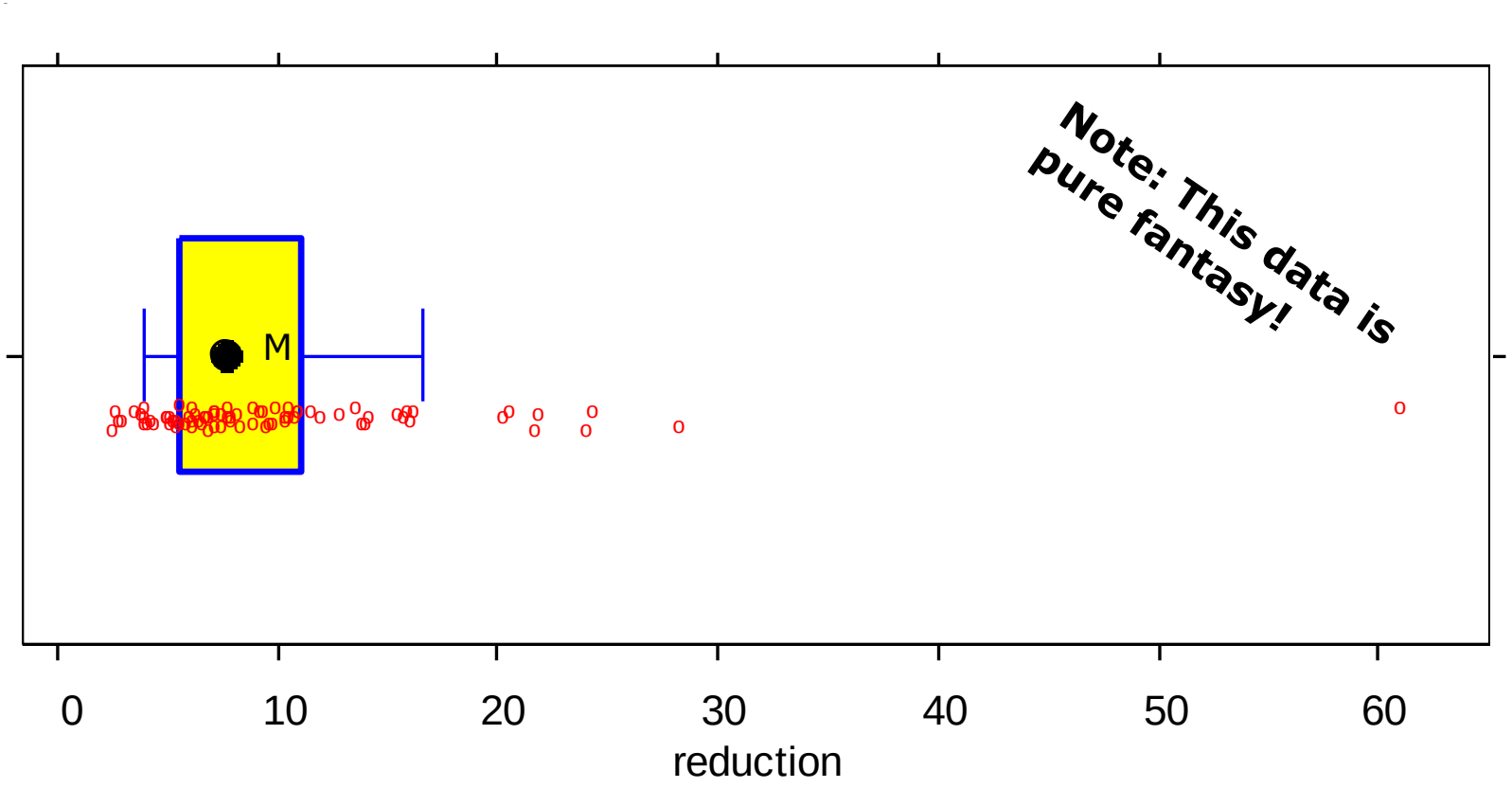
- Always question the definition of the measures for which somebody gives you statistics
  - Surprisingly often, there is no stringent definition at all
  - Or multiple different definitions are used
    - and incomparable data get mixed
  - Or the definition has dubious value
    - For example, "Energy level" may be a subjective estimate of patients who knew they were treated with a "wonder drug"

Lesson for authors: Be clear about what you measure

- Before you start:
  - What effect do you want to analyze?
  - What could be good metrics to measure it?
  - Try out different metrics and compare them
- When writing things up:
  - Define your metrics clearly and understandable.
  - Bad example: “We analyzed the delays in our simulated network”.
    - One-way or RTT?
    - Total delays? But what if wire length is constant?
  - Good example: “We analyzed the one-way delays in our simulated network. Since propagation delays are constant in a wired network, we analyzed only the queuing delays and transmission delays.”

Problem 2: A maximum does not say much

- Wrinkle reduction: up to 61%
- So that was the best value. What about the rest?
- Maybe the distribution was like this:



Lesson for readers: Dare ask for unbiased measures

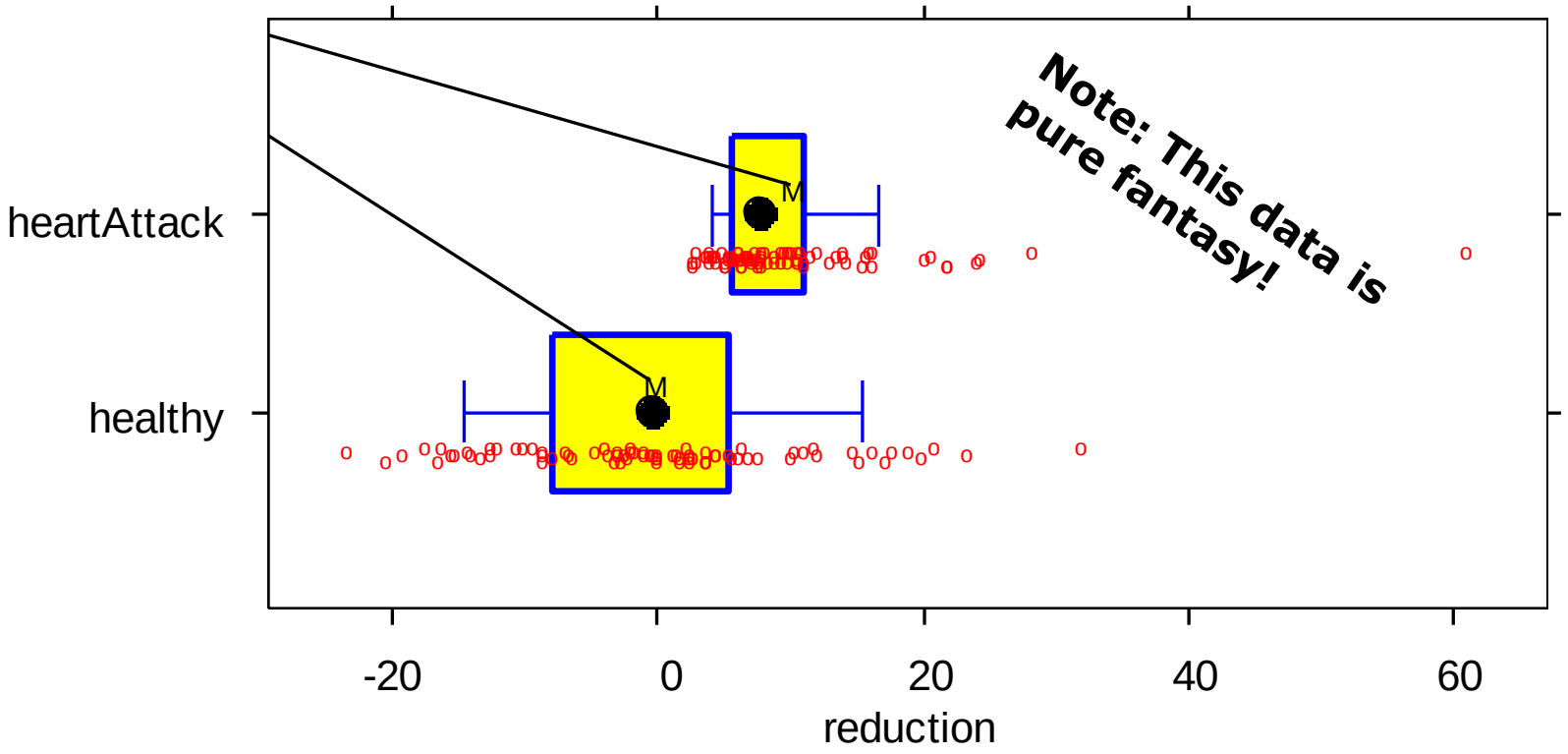
- Always ask for neutral, informative measures
  - in particular when talking to a party with vested interest
  - Extremes are rarely useful to show that someting is generally large (or small)
  - Averages are better
  - But even averages can be very misleading
    - see the following example later in this presentation
  - If the shape of the distribution is unknown, we need summary information about variability at the very least
    - e.g. the data from the plot in the previous slide has arithmetic mean 10 and standard deviation 8
  - Note: In different situations,
    rather different kinds of information
    might be required for judging something

Lesson for authors: Is it really significant?

- Are there many outliers?
- Do not use minimum or maximum values for comparison of, e.g., “before – after”
  - Compare the means
  - Think about what kind of mean to use:
    - Arithmetic mean?
    - Geometric mean?
  - Better: compare the medians
- Or even better: Use statistical tests (e.g., Student’s t test) to prove that the change (before – after) is statistically significant

Problem 3: Underlying population

- Wrinkle reduction: up to 61%
- Maybe they measured a very special set of people?



Lesson: Insist on unbiased samples

- How and where the data was collected can have a tremendous impact on the results
- It is important to understand whether there is a certain (possibly intended) tendency in this
- A fair statistic talks about possible *bias* it contains
- If it does not, ask.

Notes:

- A biased sample may be the best one can get
- Sometimes we can suspect that there is a bias,
  but cannot be sure, and we do not know the exact type of the bias

Lesson 4: ‘Cum hoc ergo propter hoc’ is wrong!

- Translation: “With this, therefore because of this”
- Meaning: Correlation does not mean causation
- Correlation may suggest causation (effect A causes effect B), but there also can be other reasons for a correlation between A and B

- Nitpicking: ‘Post hoc ergo propter hoc’ is almost the same thing:
  - After this, therefore because of this
  - Implies a temporal relation between A and B,
  - whereas ‘cum hoc…’ only implies some correlation

Correlation does not mean causation (1)

- "If A is correlated with B, then A causes B"
  - Perhaps neither of these things has produced the other, but both are a product of some third factor C
  - It may be the other way round: B causes A
  - Correlation can actually be of any of several types and can be limited to a range
  - The correlation may be pure coincidence,
    e.g. #pirates vs. global temperature
  - Given a small sample, you are likely to find some substantial correlation between any pair of characters or events

Correlation does not mean causation (2)

- Example 1: “Queueing delays increased; therefore throughput for individual TCP connections decreased”
  - Could be true
  - Could be due to an increased # of total TCP connections
  - Could be actually unrelated

- Example 2: “Chance for recovery decreases with an increasing period of cancer treatment by radiation; this shows that longer exposure to radiation is dangerous”. Well, maybe, but…
  - …usually, longer therapies are required for more severe/bigger types of cancer – and you are less likely to surveve these

Correlation does not mean causation (3)

- Example 3:
  “Birth rates have been decreasing for decades.
  So has the number of storks.
  This proves that babies are delivered by the stork!”

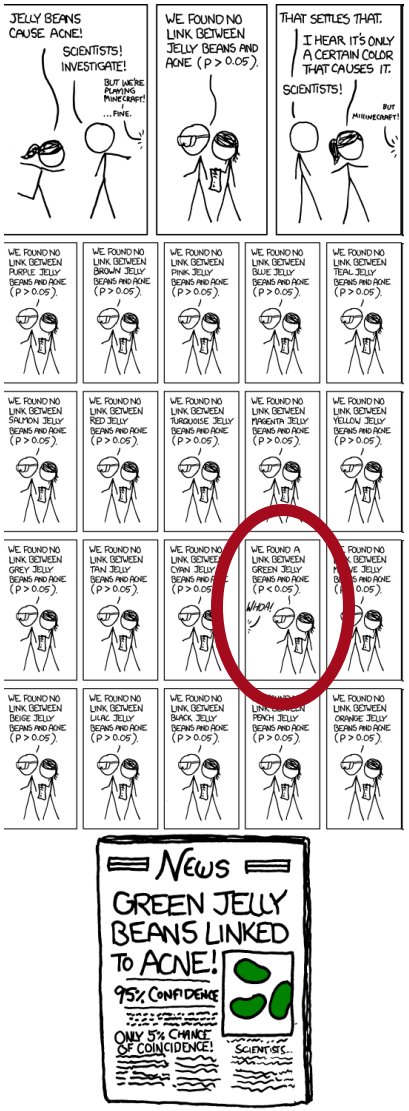
- Example 4:
  “The number of TV stations has increased, as well as the amount of money that people spend on travelling.
  This proves the efficiency of travel ads on TV.”

Correlation does not mean causation: Lessons

- Often, there is a hidden background variable (e.g., size of the tumor)
- Time is a good candidate for a background variable (e.g., storks vs babies, TV stations vs. travel expenses)

Fishing for correlations

- Correlation can be a purely random effect!
- Statisticians assume that in ~5% of all cases, two arbitrarily chosen variables appear to be correlated
- Example:
  - Determine 20 parameters (=rnd variables) in some simulation experiment
  - Can create ½ · 20 · 19 = 190 pairs of random variables
  - 5% of 190 = about 9 – 10 “correlations” that are in fact purely random!



http://www.xkcd.com/882/

Lesson: Question causality

- Sometimes the data is not just biased,
  it contains hardly anything other than bias

- If you see a presumably (=author) or assertedly (=reader) causal relationship ("A causes B"), ask yourself:
  - Does it really make sense?
  - Would A really have this much influence on B?
  - Couldn‘t it be just the other way round?
  - What other influences besides A may be important?
  - What is the relative weight of A compared to these?

Percentages

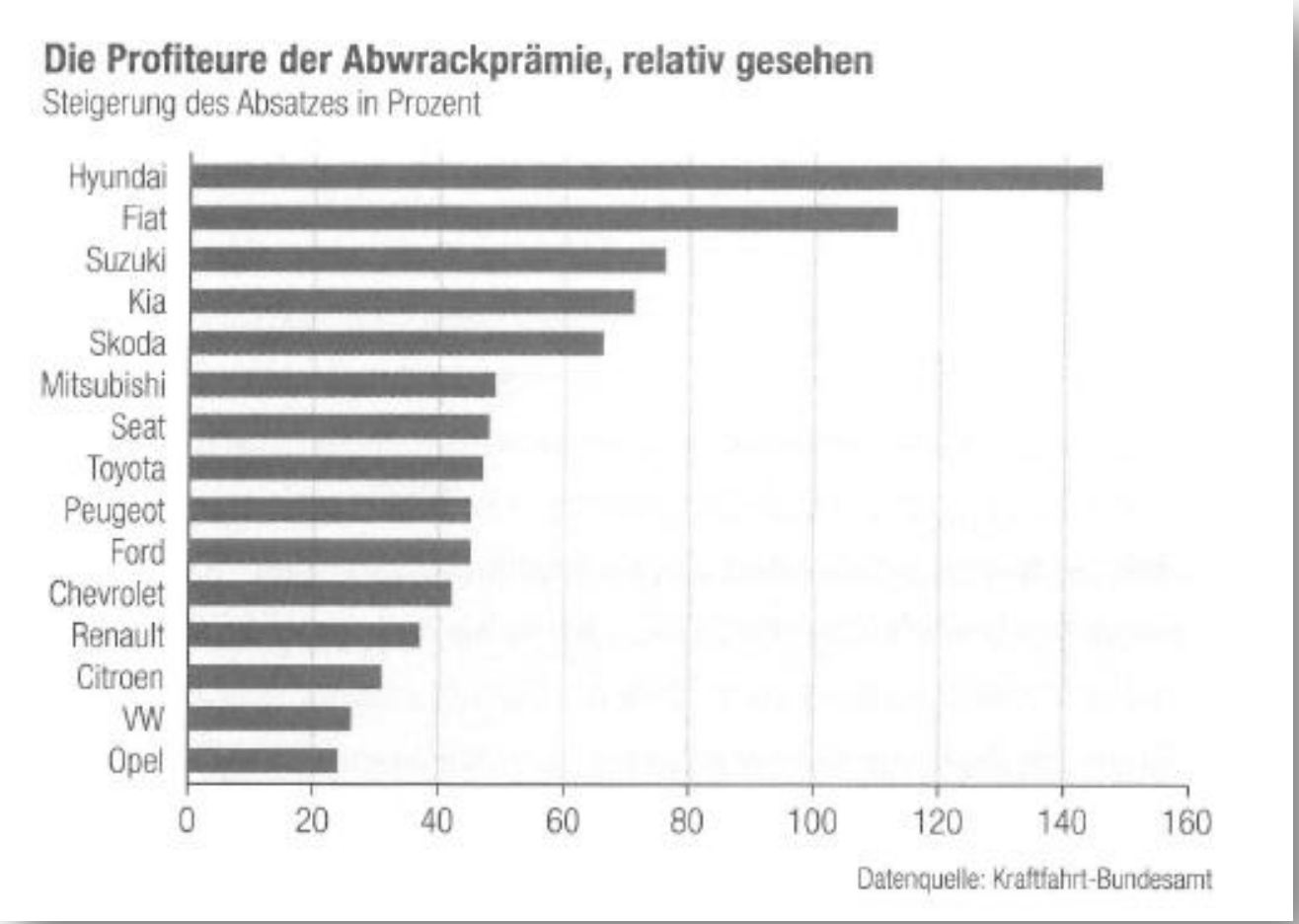
- “Wohl- und übelwollende Benutzer gleichermaßen schätzen es [das Prozent] wegen seiner Aura von mathematischer Neutralität und Sachlichkeit. ‘Prozent’ […] riecht man Kaufmannskontor und doppelter Buchführung; die Seriosität quillt nur so aus den Knopflöchern. Prozente stehen für Glaubwürdigkeit und Autorität, Prozente strahlen Gewissheit aus, Prozente zeigen, dass man rechnen kann, sie verleihen Autorität und Überlegenheit, umso mehr, und wahrscheinlich noch dadurch verstärkt, als so mancher Adressat einer modernen Prozentpredigt überhaupt nicht weiß, was eigentlich Prozente sind.”
  – Walter Krämer

Percentages and absolute numbers (1)

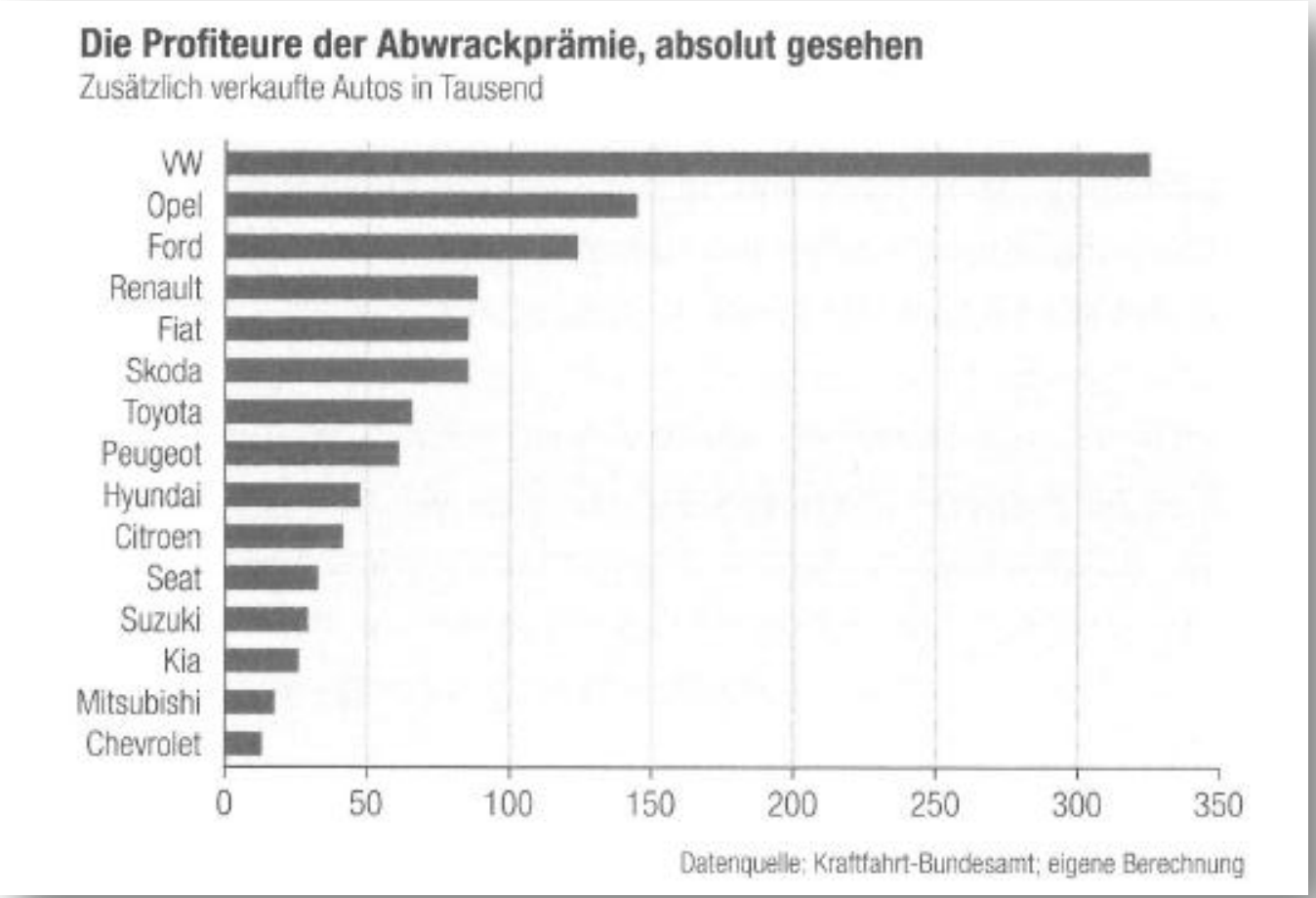
You're in hospital, and the doctor tells you…:

- "Medication A has a 10% higher chance to cure your disease, but the thrombosis risk is increased by 100% in comparison to medication B."
  - Which one would you pick?
- "With medication B, about 1 in 7,000 patients suffers from thrombosis. With medication A, about 2 in 7,000 patients suffers from thrombosis, but it has a 10% higher chance to cure your disease."
  - Which one would you pick?
  - Mathematically, the two descriptions are equivalent!
- Your decision probably depends on the graveness of your disease (e.g., headache vs. liver cancer)
- Lesson: Percentages can be misleading!

Example - Percentages and absolute numbers



Example - Percentages and absolute numbers



Percentages and absolute numbers (2)

- “In the past year, we have employed an additional 1,000 teachers in North Rhine Westphalia. This shows our great commitment and financial efforts to improve our school system.” – Sounds good, doesn’t it?
- How many schools are there in NRW?
  - About 7,000
  - Only one in seven schools (about 14%) gets an additional teacher!
- How many teachers are there in NRW in total?
  - About 130,000
  - Result: Less than 1% increase…

- Lesson: Absolute numbers can be misleading, too!

Percentages of what? – Two examples

- In 2008, President Bush asserted that the USA would reduce their emissions of greenhouse gases by the year 2050 by at least 50%.
- 50% – but as compared to what?
  - In relation to the year 1990? – International standard
  - In relation to the year with the highest emissions?
    - …which might yet be to come!?

- The share of nuclear energy in Germany is about 25%
  - True for electrical energy
- The share of nuclear energy in Germany is about 13%
  - True for total primary energy consumption

Percentages (4)

- "In the past year, we could boost our company's rate of return by 400%!"
  - Wow, 400%. Impressive!
- "That is because we increased our rate of return from 0.1% to 0.5%."
  - Just 0.5%. How inefficient!

- Lessons
  - Always ask (or write out): "percentage of what?"
  - Always ask for (or write out)
    - The percentages
    - *And* the absolute numbers
  - Percentages of percentages often don't make sense and can be an indication of foul play (cf. next slide)

Prozentzahlen und Prozentpunkte

- Wahl 2010:
  - Partei A: 40%
  - Partei B: 10%
- Wahl 2014:
  - Partei A: 30%
  - Partei B: 20%
- „Partei A hat 10% verloren, Partei B hat 10% gewonnen“
  - Falsch: Partei A hat
    - 10 Prozent**punkte** verloren
    - 25% verloren (denn 40/30 = 0,75)
      - …aber auch nicht der absoluten Stimmen, da vermutlich unterschiedliche Wahlbeteiligung, unterschiedliche Anzahl Wahlberechtigte, etc. etc.
- Lektion: Es gibt einen wichtigen Unterschied zwischen Prozent und Prozentpunkten!

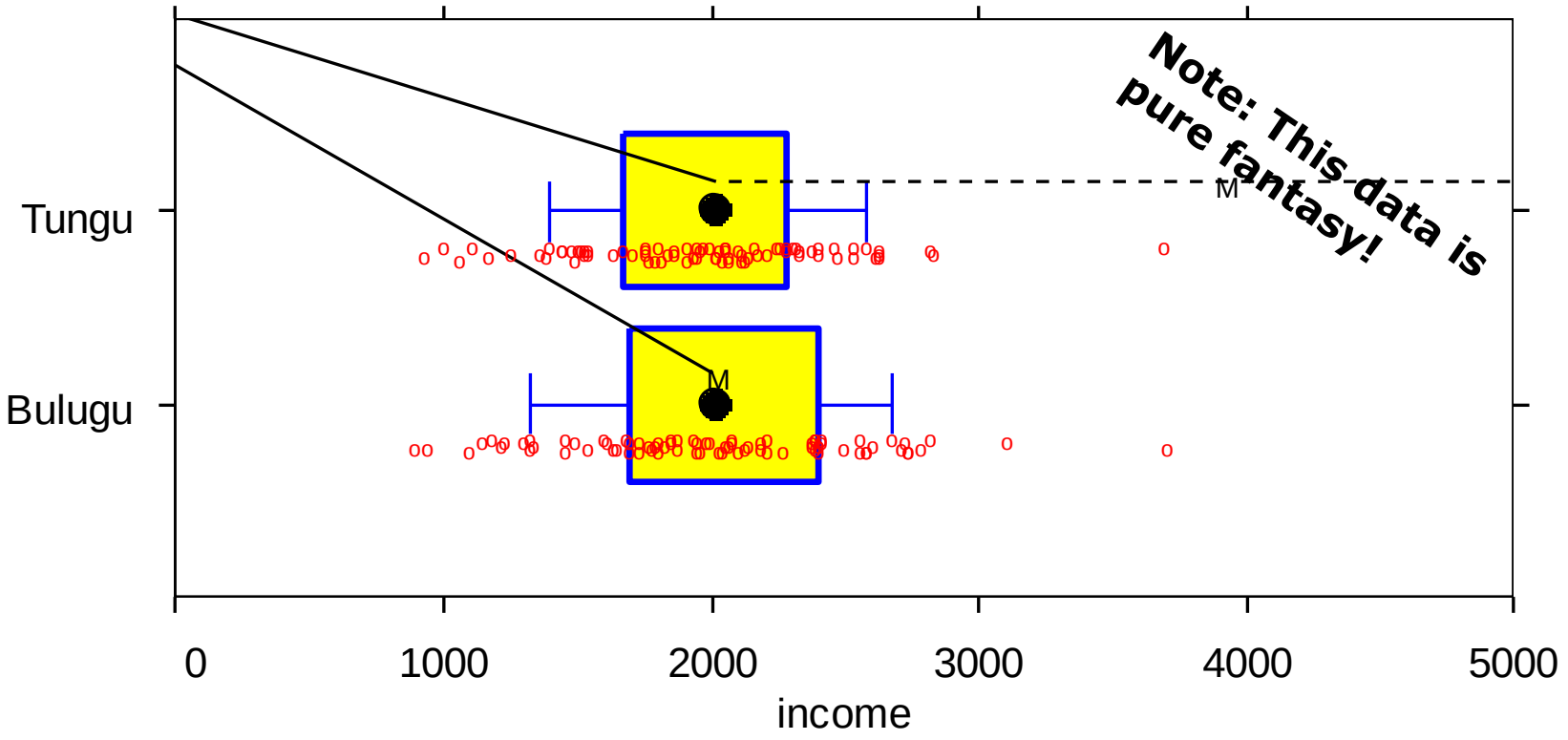
Example 2: Tungu and Bulugu

- We look at the yearly per-capita income in two small hypothetic island states:
  Tungu and Bulugu

- Statement:
  "The average yearly income in Tungu is 94.3% higher than in Bulugu."

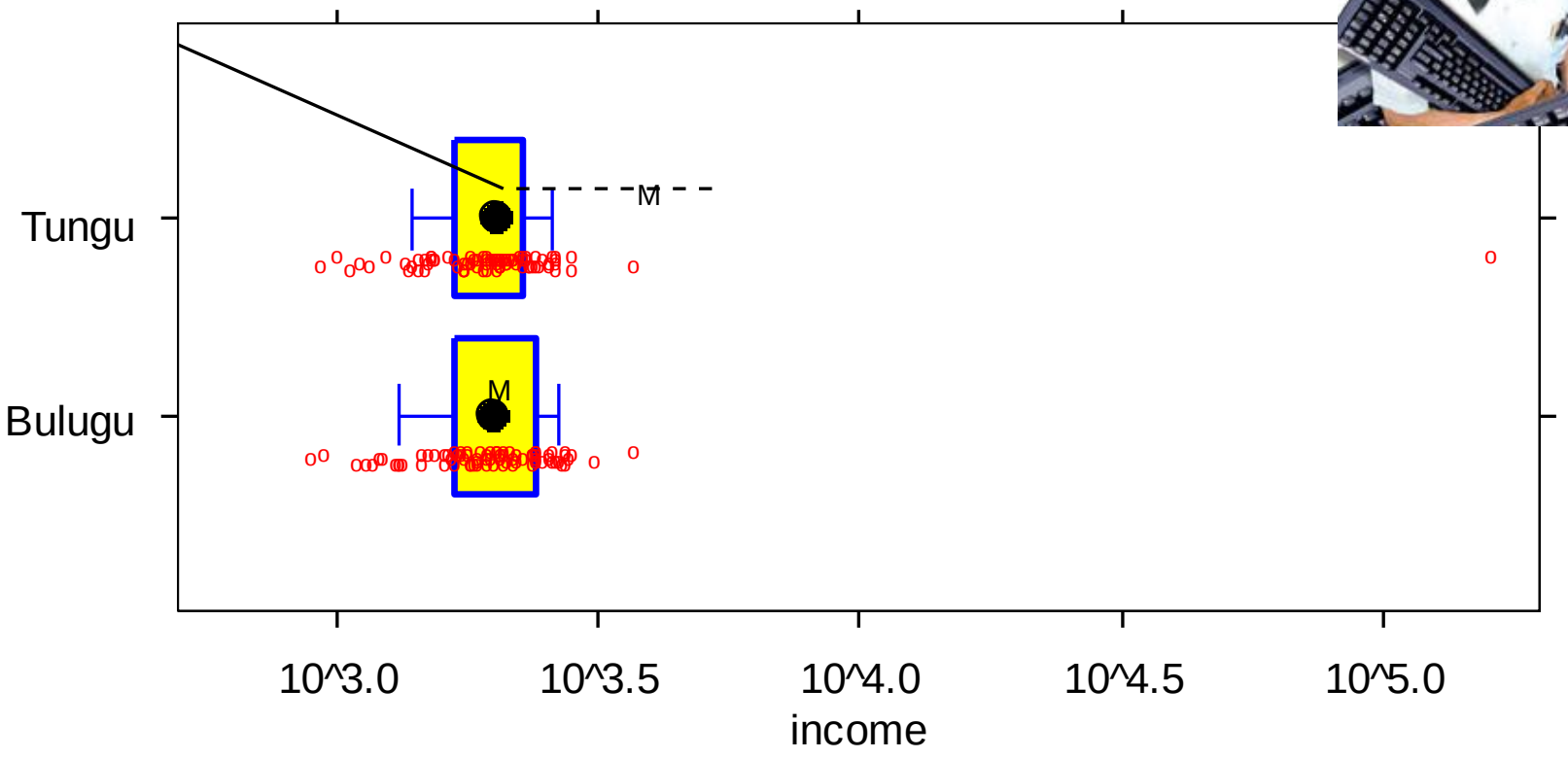
Problem 1: Misleading averages

- The island states are rather small:
  81 people in Tungu and **80** in Bulugu
- And the income distribution is not as even in Tungu:



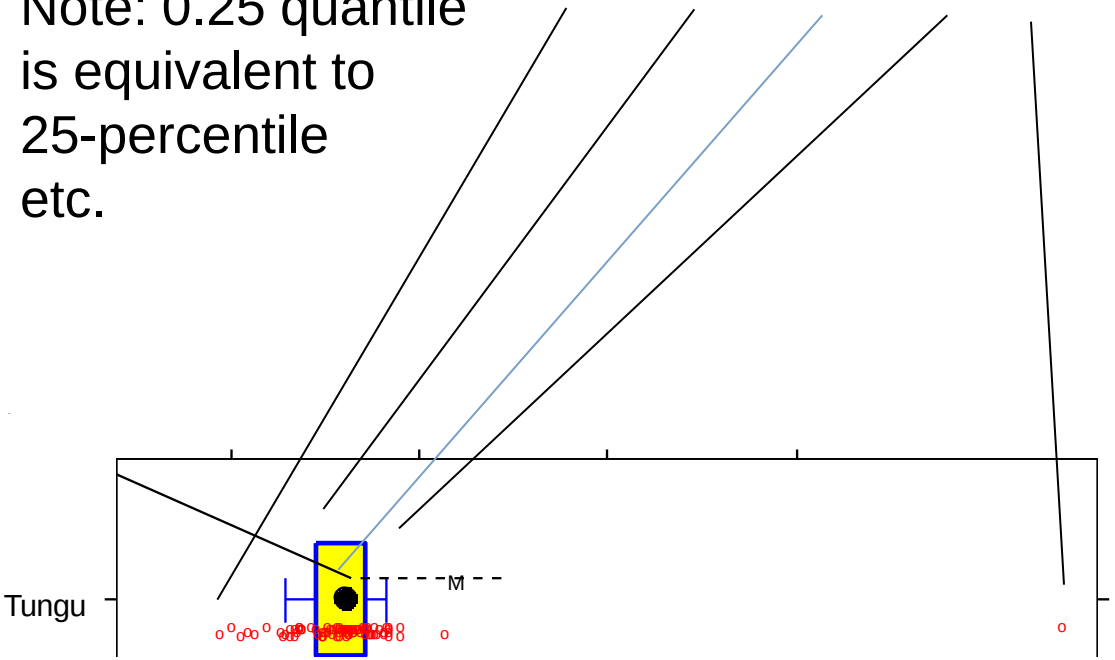
Misleading averages and outliers

- The only reason is Dr. Waldner, owner of a software company, who has been enjoying his retirement in Tungu for a year



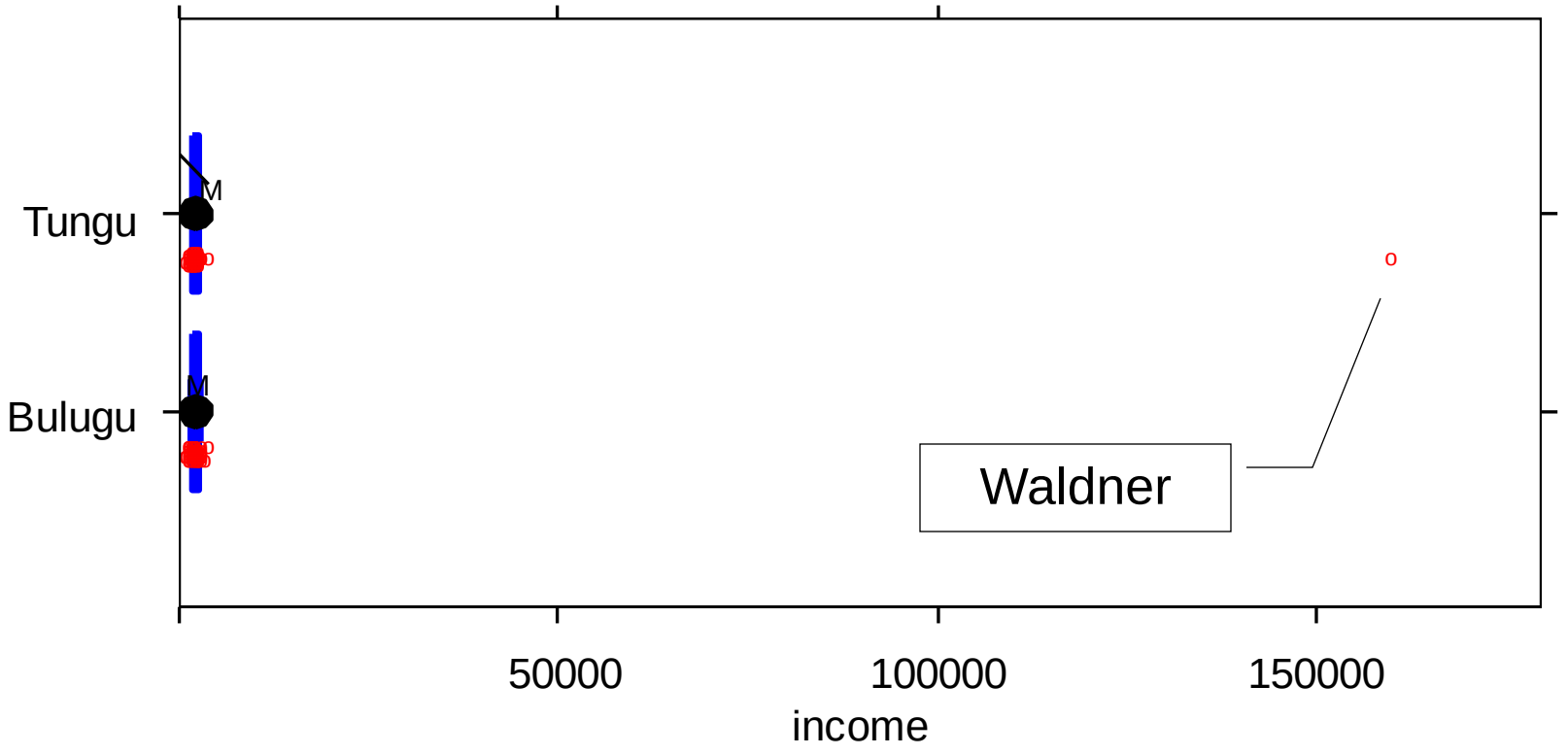
Lesson: Question appropriateness

- A certain statistic (very often the arithmetic average) may be inappropriate for characterizing a sample
- If there is any doubt, ask that additional information be provided
  - such as standard deviation
  - or some quantiles, e.g.: 0, 0.25, 0.5, 0.75, 1
    Note: 0.25 quantile
    is equivalent to
    25-percentile
    etc.



Logarithmic axes

- Waldner earns 160.000 per year.
  How much more that is than the other Tunguans have, is impossible to see on the logarithmic axis we just used



Lesson: Beware of inappropriate visualizations (#1)

- Lesson for reader: Always look at the axes. Are they linear or logarithmic?

- Lesson for author:
  - Logarithmic axes are very useful for reading hugely different values from a graph with some precision
  - But they totally defeat the imagination!
  - If you decide to use logarithmic axes, always state this fact in your text!

- There are many more kinds of inappropriate visualizations
  - see later in this presentation

Problem 4: Misleading precision

- "The average yearly income in Tungu is **94.3%** higher than in Bulugu"

- Assume that tomorrow Mrs. Alulu Nirudu from Tungu gives birth to her twins

- There are now 83 rather than 81 people on Tungu
- The average income drops from 3922 to 3827
- The difference to Bulugu drops from 94.3% to 89.7%

Lesson for reader: Do not be easily impressed

- The usual reason for presenting very precise numbers is the wish to impress people
  - *„Round numbers are always false“*
  - But round numbers are much easier to remember and compare
- Clearly tell people you will not be impressed by precision
  - in particular if the precision is purely imaginary

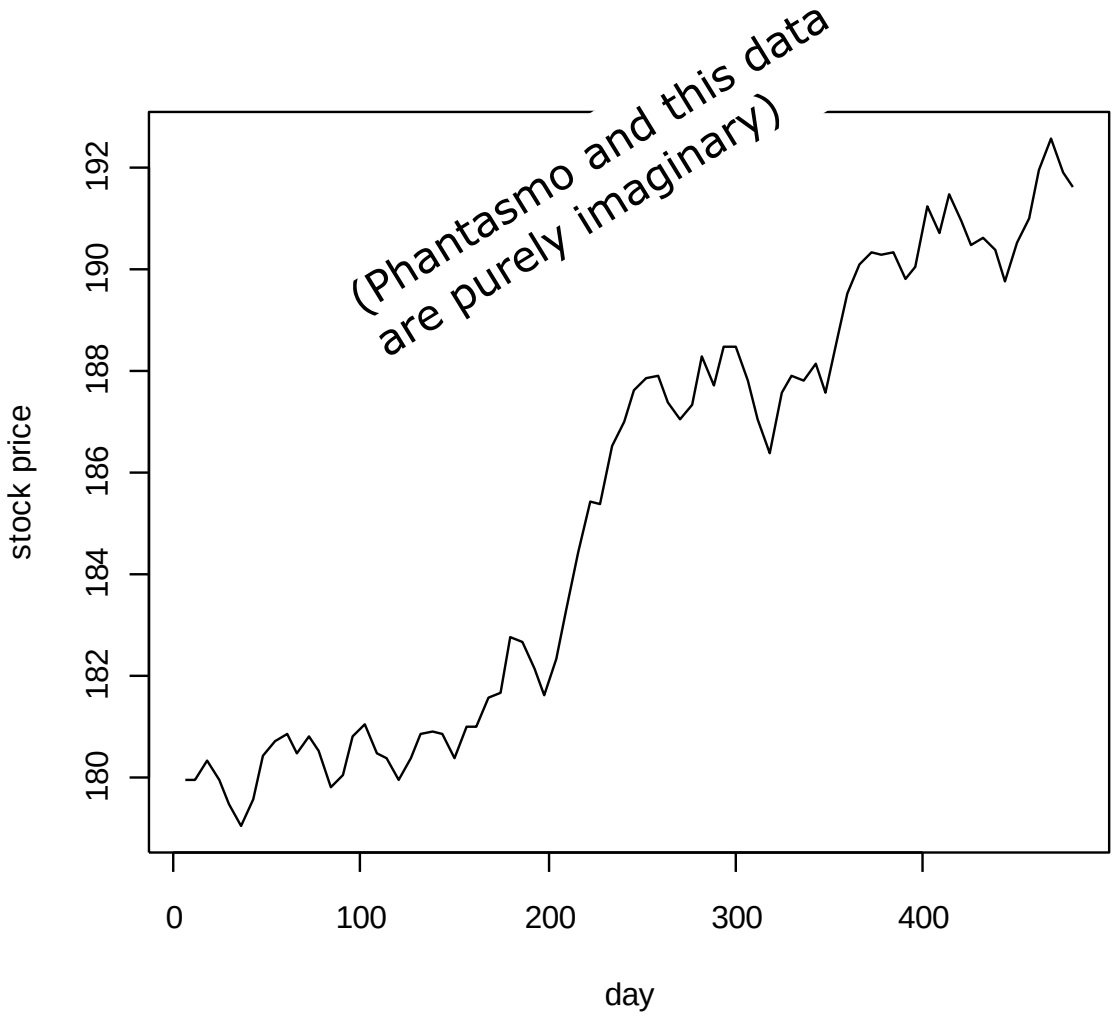
Lesson for author: Think about precision

- Do you really have enough data that would make sense to give out precise numbers?

- Compromise: Give exact number in tables/figures, but round them in text.
- Do not exaggerate: If you find your systems yields a 52,91% increase in throughput
  - Don’t say: “Our system increases throughput by more than 50%”
  - Do say: “Our experiments suggest that our system can achieve throughput increases of around 50%”

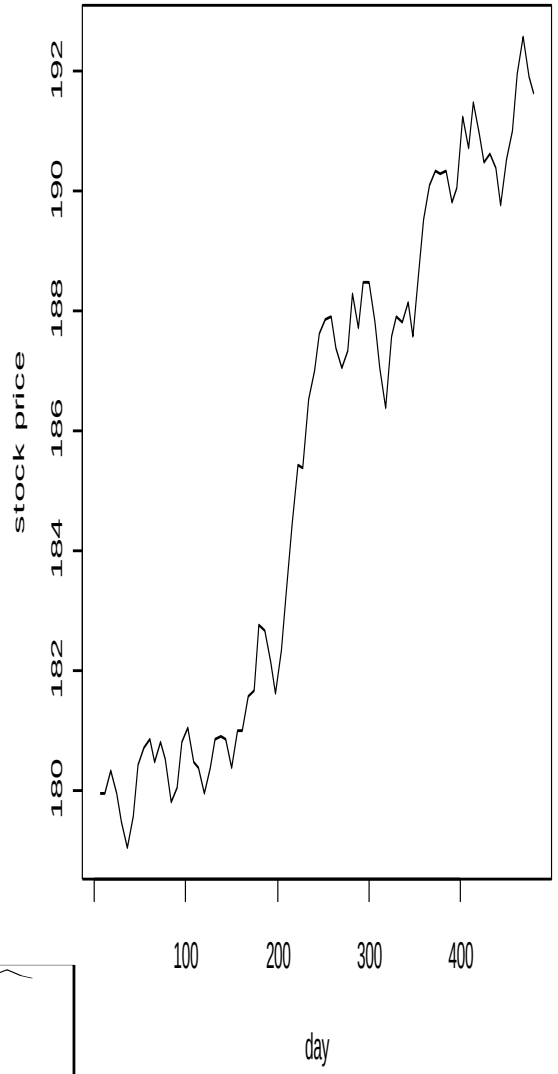
Example 3: Phantasmo Corporation stock price

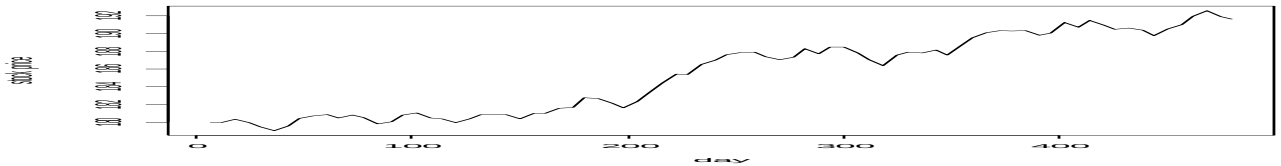
- We look at the recent development of the price of shares for Phantasmo Corporation
- *"Phantasmo shows a remarkably strong and consistent value growth and continues to be a top recommendation"*



Problem: Looks can be misleading

- The following two plots show exactly the same data!
  - and the same as the plot on the previous slide!

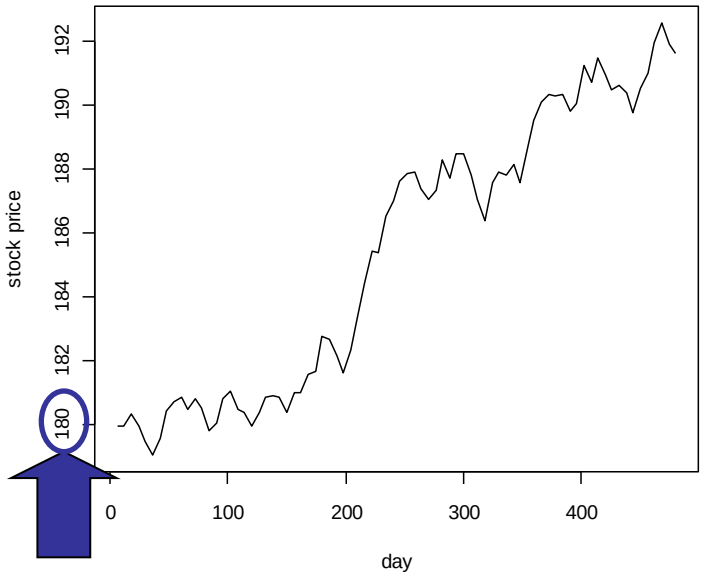


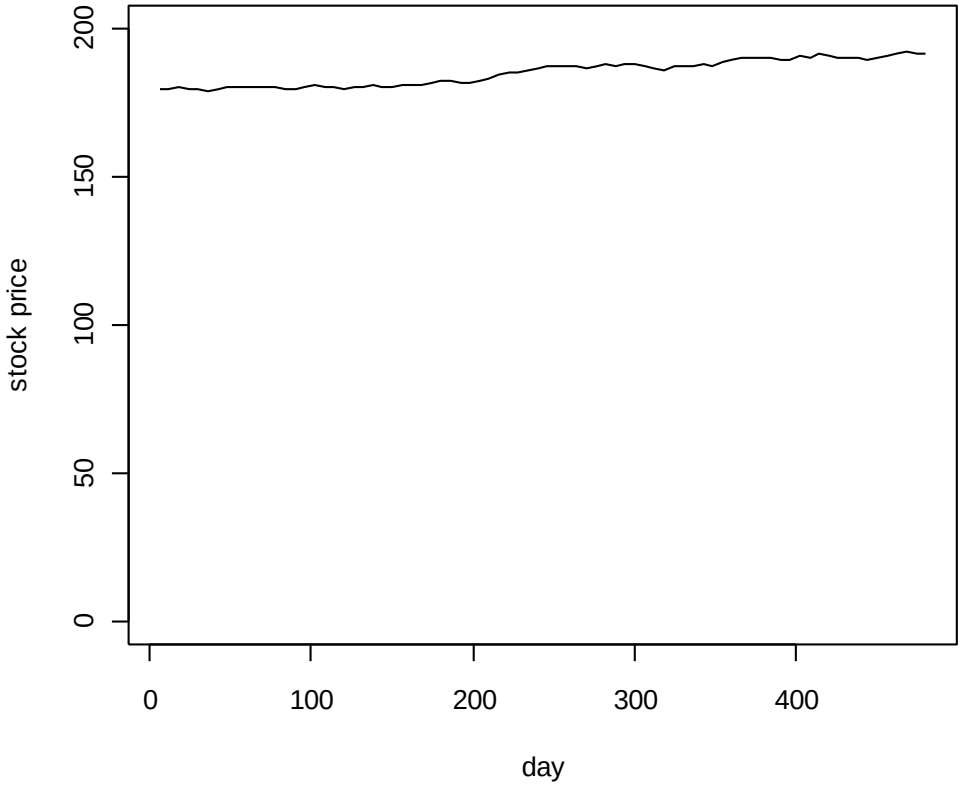


Problem: Scales can be misleading

- What really happened is shown here:

  We intuitively interpret a trend plot on a ratio scale



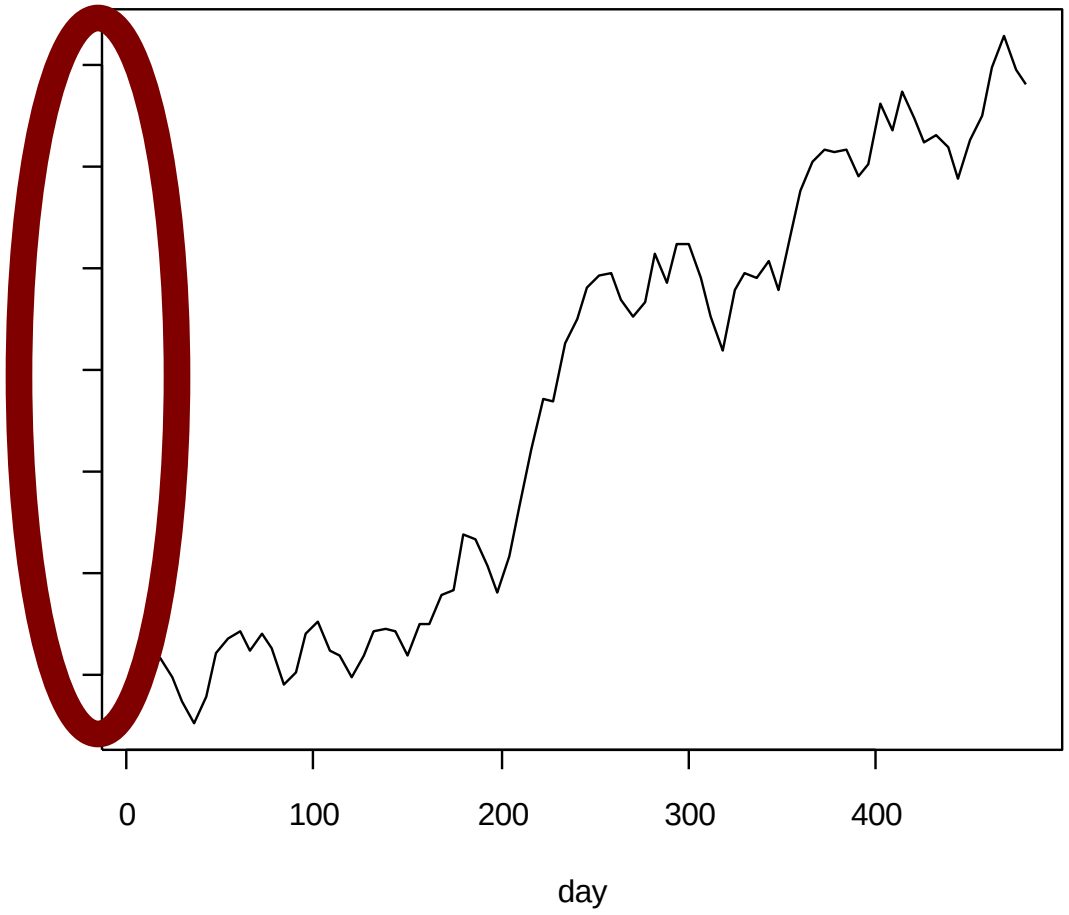


Problem: Scales can be missing

- The most insolent persuaders may even leave the scale out altogether!

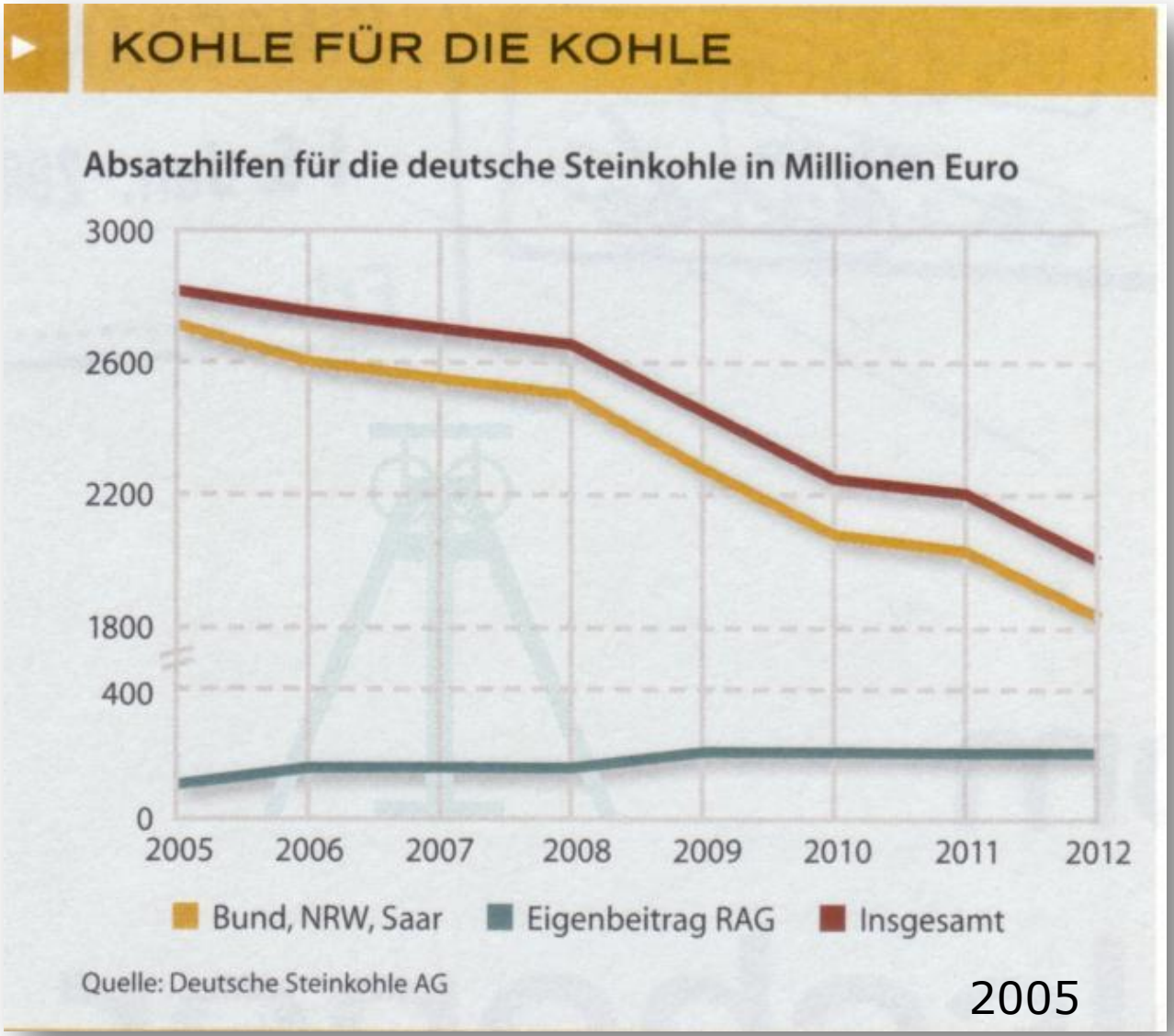
•Never forget to label your axes!

•Never forget to put a scale on your axes!



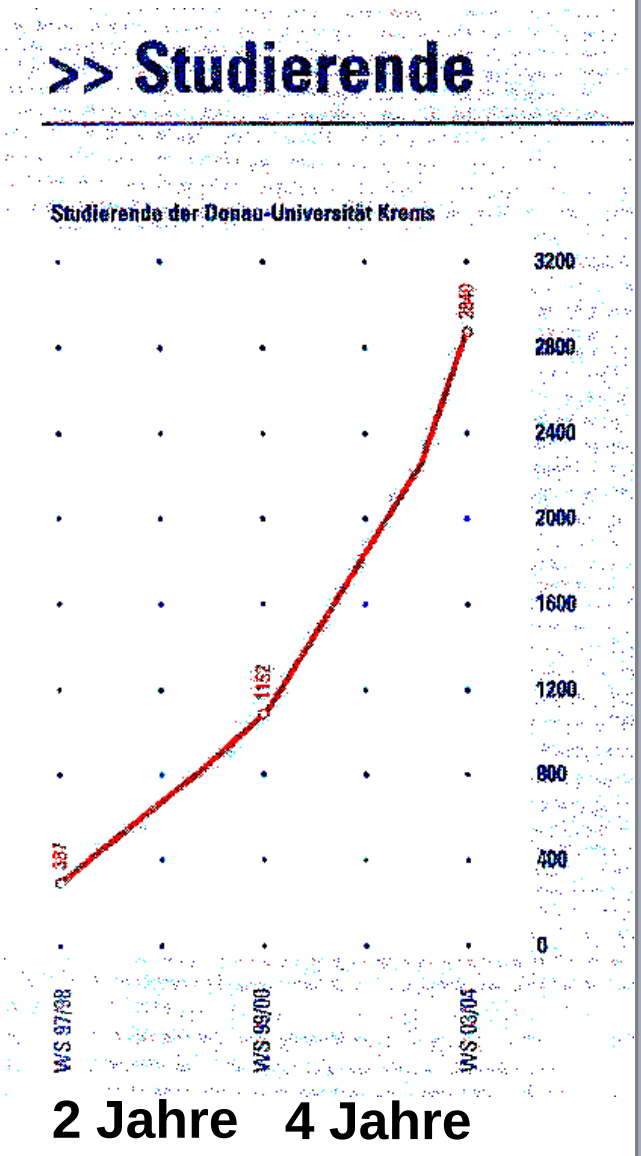
Problem: Scales can be abused

- Observe the global impression first

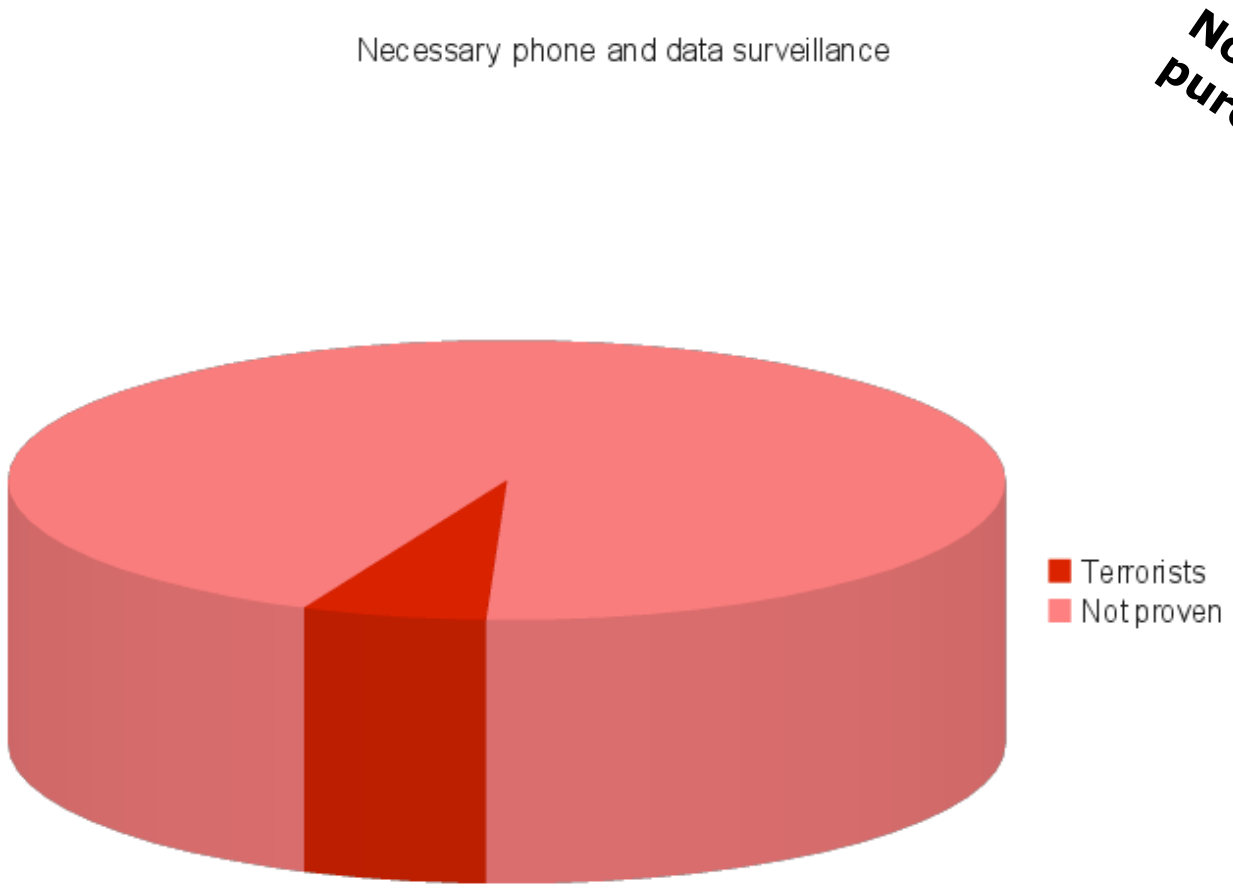


Problem: People may invent unexpected things

- Quelle: Werbeanzeige der Donau-Universität Krems
  - DIE ZEIT, 07.10.2004
  - What‘s wrong?

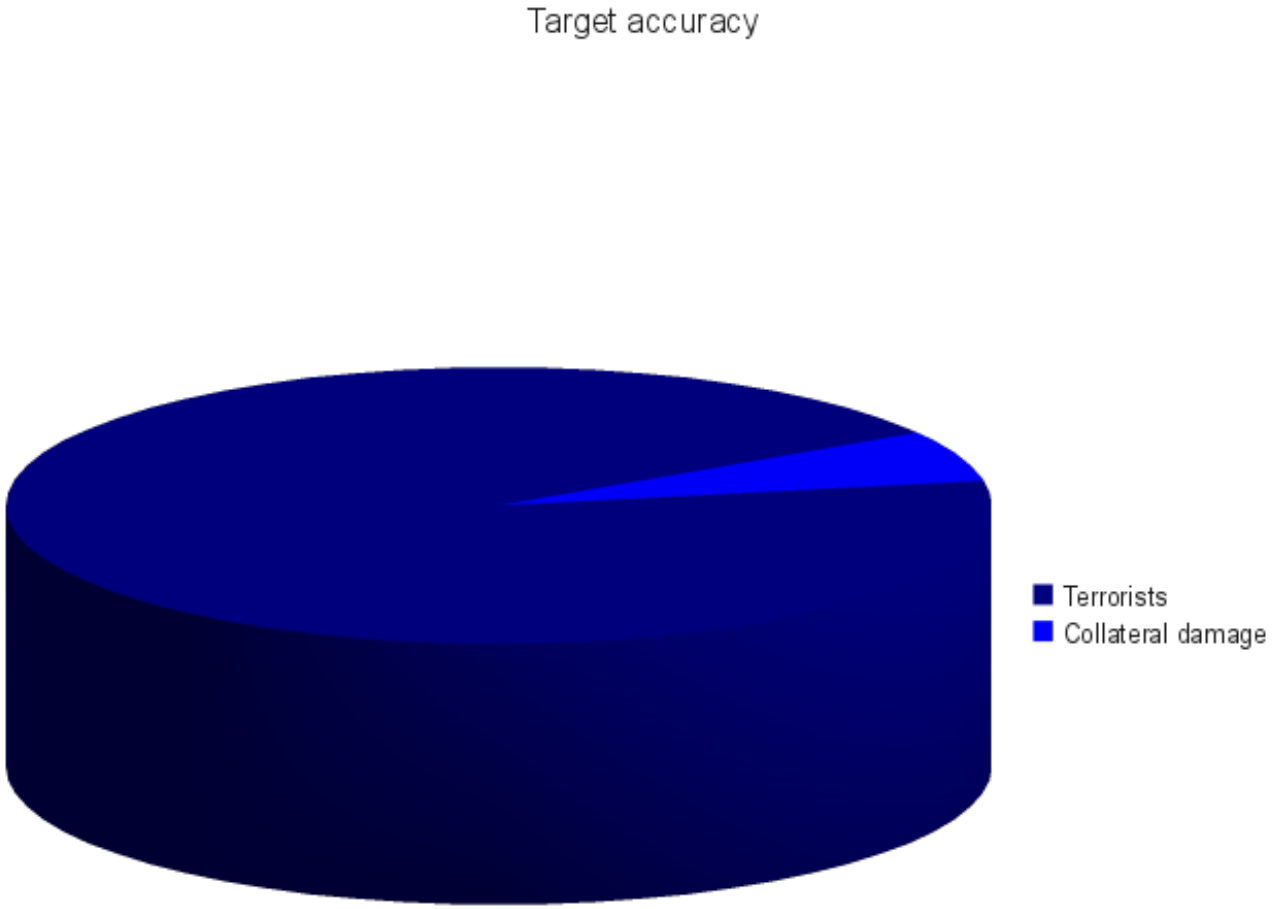


Pie charts (1/3)



Note: This data is pure fantasy!

Pie charts (2/3)



te: This data is
fantasy!

Pie charts (3/3)

- What percentages do the two graphs show?
  Guess!

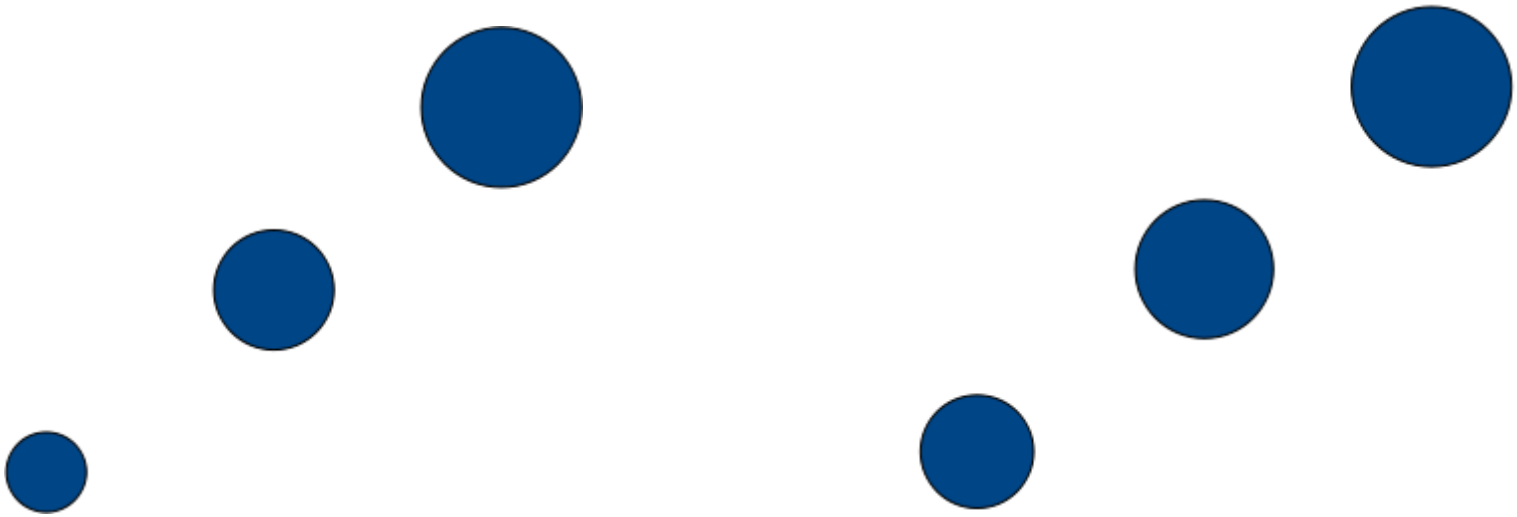
- Answer:
  - **Both** show the same data: A 94% : 6% ratio!
  - The difference only lies in the angle of the pies.

Lesson: Distrust pie charts!

- Pie charts should **never** be used
  - Perception dependent on the angle
  - Even worse with 3D pie charts:
    Parts at the front are artificially increased due to the pie's 3D height; they thus seem to be bigger
  - A very subtle way to visually tune your data
  - Unfortunately, still very common

- Distrust pie charts that do not give numbers as well
  - Think about the numbers, compare them
  - Think about the presentation: are they trying to beautify the impression?

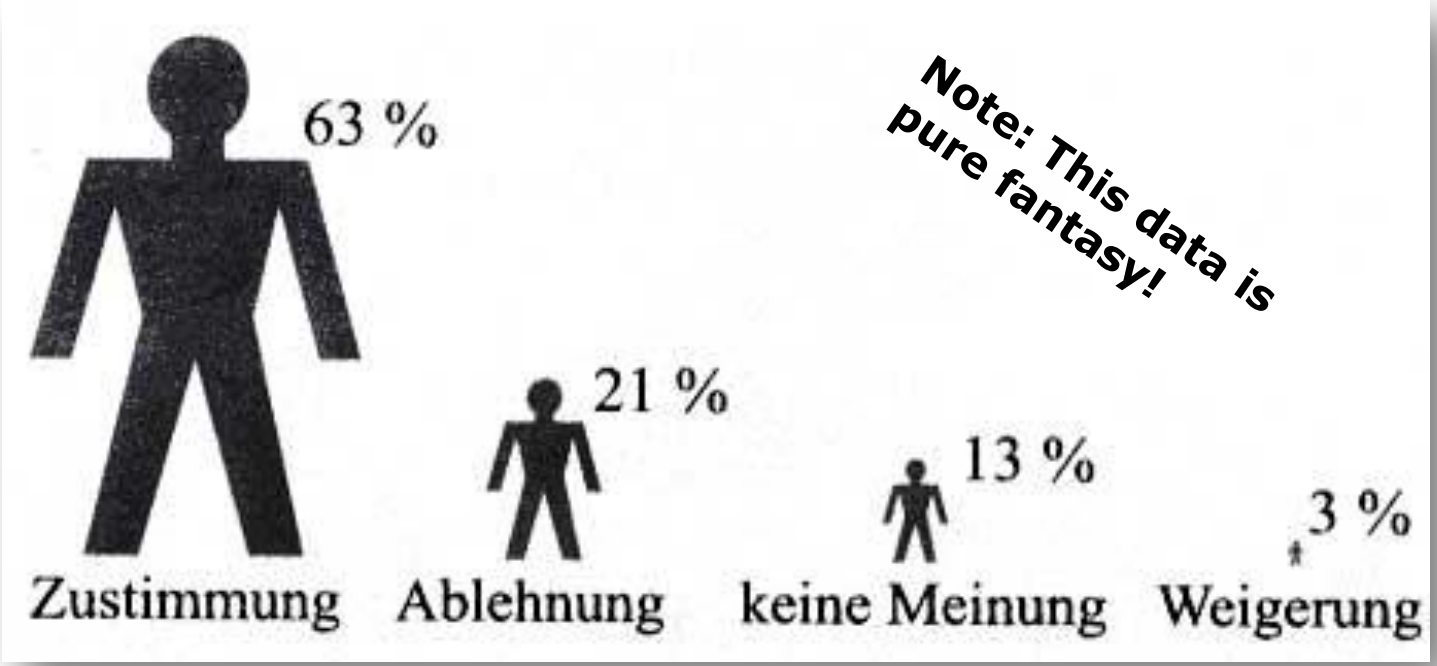
Bubble charts



Note: This data is pure fantasy!

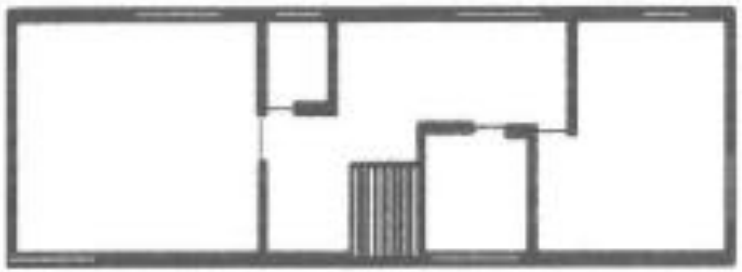
- Which diagram shows the values 2, 3, 4?
- Both do!
- Left one: Radius is proportional to measurements
  - Exaggerates differences: 4 looks much larger than 2
- Right one: Area is proportional to measurements
  - Underestimates differences: 4 looks only slightly larger than 2

Pictograms

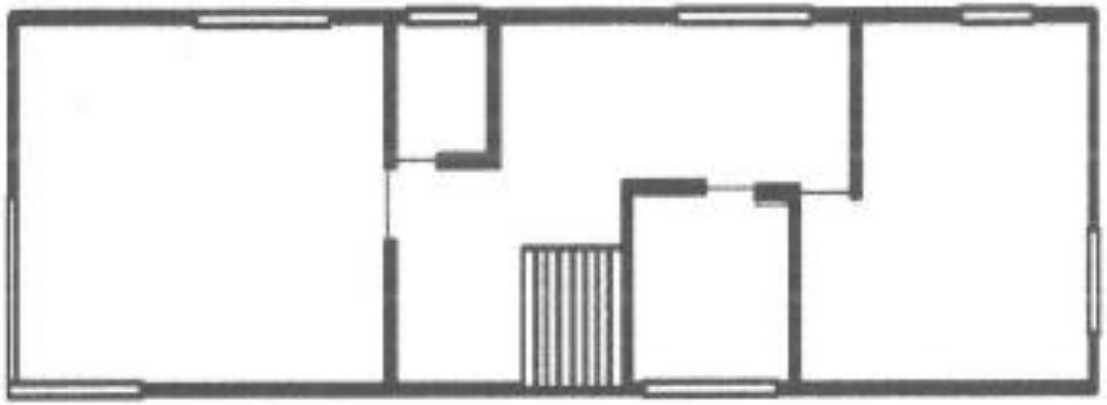


http://sciencev1.orf.at/static2.orf.at/science/storyimg/storypart_155543.jpg

Pictogram – Comparison Apartment size



ca. 58 m²



ca. 82 m²

Vorlage: »Zahlenspiegel« Bundesrepublik Deutschland – DDR

Lesson: Bubble charts and pictograms

- This lesson is more or less similar to pie charts:
- Bubble charts usually should not be used
  - Radius proportionality exaggerates differences, area proportionality lets underestimate differences
  - A very subtle way to visually tune your data
  - Of course, a bubble chart + pie chart may convey more information, but *please* try to visualize it differently…
  - If you really, really want to use a bubble chart, then use the area proportionality variant, and clearly explain this in your text
- Distrust bubble charts that do not give the numbers as well
  - Think about the numbers, compare them
  - Think about the presentation: Did they really need to use bubble charts? Or are they trying to beautify the impression?

Sometimes size really matters.

Summary lesson for the reader: Seeing is believing

- …but often, it shouldn't be!

- Always consider what it really is that you are seeing
- Do not believe anything purely intuitively
- Do not believe anything that does not have a well-defined meaning

Example 4: blend-a-med Night Effects

- What do they not say? Think about it…

blend-a-med Night Effects

Sichtbar hellere Zähne nach 14 Nächten – für mindestens 6 Monate.

- Zahnaufhellungsgel für die Nacht
- Klinisch getestet
- Einfach aufpinseln
- Mit patentierter LiquidStrip Technologie

- What exactly does "sichtbar" mean?
  What exactly does „hell“ or „heller“ mean?
- What was the scope, what were the results of the clinical trials?
- What other effects does Night Effects have?

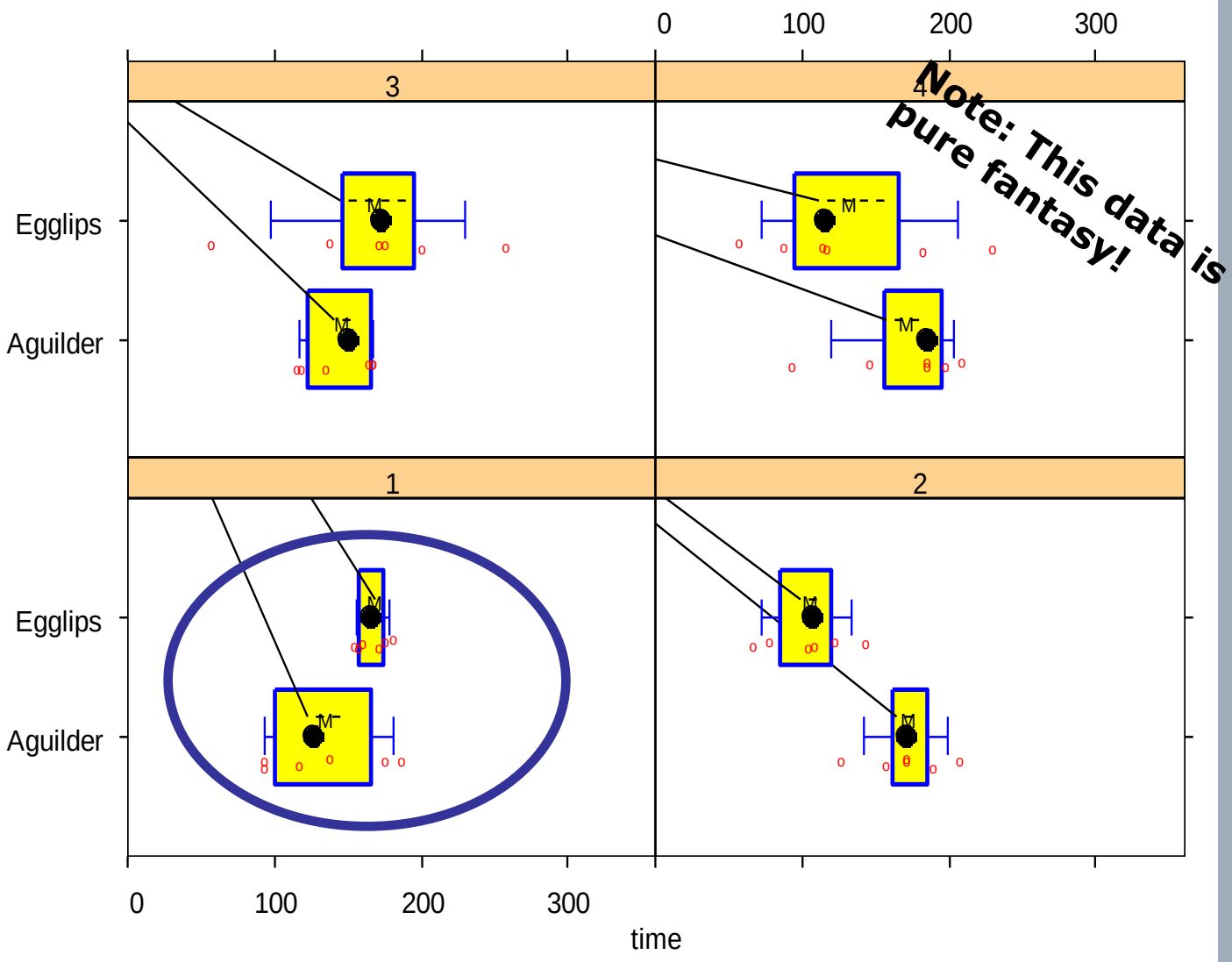
Example 5: The better tool?

- We consider the time it takes programmers to write a certain program using different IDEs:
  - *Aguilder* or
  - *Egglips*
- Statement (by the maker of Aguilder):
  *"In an experiment with 12 persons, the ones using Egglips required on average **24.6% more time** to finish the same task than those using Aguilder.*
  Both groups consisted of equally capable people and received the same amount and quality of training."

- Assume Egglips and Aguilder are in fact just as good.
  What may have gone wrong here?

Problem: Has anybody ignored any data?

- Solution: Just repeat the experiment a few times and pick the outcome you like best



Lesson for the reader: Demand complete information

- If somebody presents conclusions
  - based on only a subset of the available data
  - and has selected which subset to use
  - then everything is possible

- There is no direct way to detect such repetitions,

BUT for any one single execution . . .

Digression: Hypothesis testing

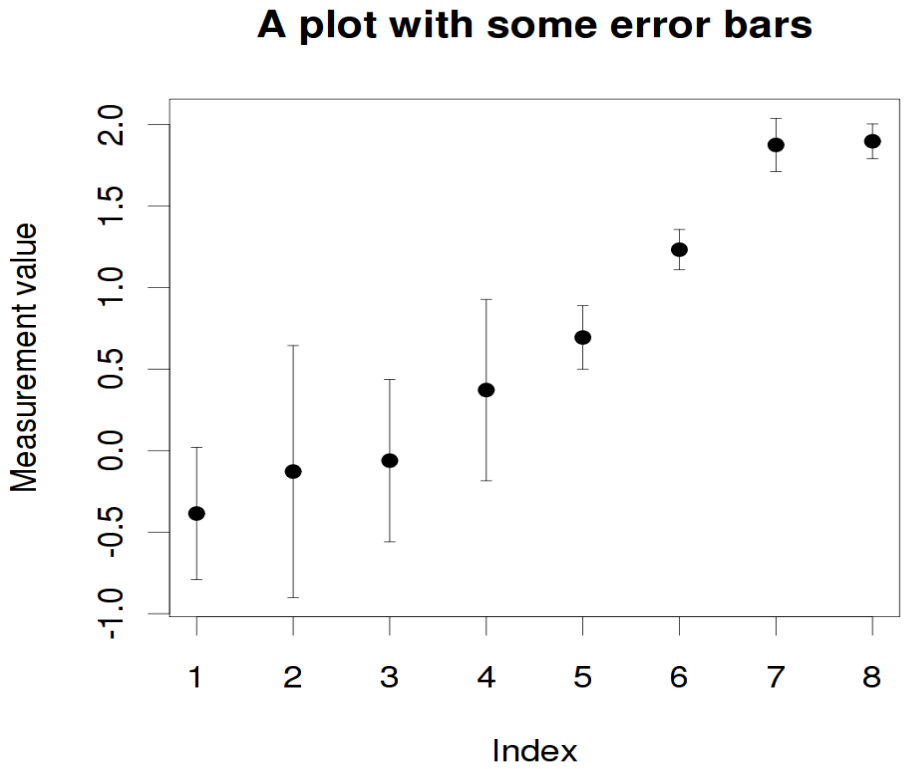
- …a *significance test* (or confidence intervals) can determine how likely it was to obtain this result if the conclusion is wrong:
  - Null hypothesis: Assume both tools produce equal work times overall
  - Then how often will we get a difference this large when we use samples of size 6 persons?
    - If the probability is small,
      the result is plausibly real
    - If the probability is large,
      the result is plausibly incidental

Statistical significance test: Example

- Our data:
  - Aguilder: 175, 186, 137, 117, 92.8, 93.7 (mean 133)
  - Egglips: 171, 155, 157, 181, 175, 160 (mean 166)
- Null hypothesis:
  - We assume the distributions underlying these data are both normal distributions with the same variance and
  - the means of the actual distributions are in fact equal
- Then we can compute the probability for seeing this difference of 33 from two samples of size 6
- The procedure for doing this is called the *t-test*
  (recall the confidence intervals? – It‘s a very similar calculation)

Example: Error bars

- “Although a high variability in our measurements results in rather large error bars, our simulation results show a clear increase in [whatever].”
- What’s wrong here?



Lesson: Error bars

- What are the error bars? How are they defined?
  - Minimum and maximum values?
  - Confidence intervals?
    - If so, at which level? 95%? 99%?
  - Mean ± two standard deviations?
  - First and third quartile? 10% and 90% quantile?
  - Chebyshov* or Chernov bounds?
    *also: Tschebyscheff, Tschebyschow, Chebyshev, … Same with Tschernoff, …

- Reader: Distrust error bars that are not explained
- Author:
  - Clearly state what kind of error bars you're using
  - Usually, the best choice is to use confidence intervals, but stddev is also quite common

Lesson for the author: Common errors for t tests and confidence intervals

- Recall: “But unless the distribution of your samples is very strange or very different, using the t-test is usually OK.”
- If you do not have many samples (less than ~30), then you must check that your input data looks more or less normally distributed
  - At least check that the distribution does not look terribly skewed
  - Better: do a QQ plot
  - Even better: use a normality test
- You might make many runs, group them together and exploit the Central Limit Theorem to get normally distributed data, but…:
  - Warning: Only defined if the variance of your samples is finite!
  - Therefore won’t work with, e.g., Pareto-distributed samples ($\alpha<2$)
- You must ensure that the samples are not correlated!
  - For example, a time series is often autocorrelated
  - Group samples and calculate their average (Central Limit Theorem); make groups large enough to let autocorrelation vanish
  - Check with ACF plot
    or autocorrelation test
    or stationarity test

Lesson for the author: Check your prerequisites and assumptions!

- Similar errors can be committed with other statistical methods
- Usual suspects:
  - Input has to be normally distributed, or follow some other distribution
  - Input must not be correlated
  - Input has to come from a stationary process
  - Input must be at least 30 samples (10; 50; 100; …)
  - The two inputs must have the same variances
  - The variance must be finite
  - The two inputs must have the same distribution types
  - …
  - of course, all this depends on the chosen method!

Will Rogers phenomenon (1)

- Revenues per salesman of company HuiSoft for two consecutive years, in k€:

2010		2011	
Bielefeld	München	Bielefeld	München
5000	5000	5000	5000
6000	10000	6000	
7000	15000	7000	15000
	20000	10000	20000
μ=6000	μ=12500	μ=7000	μ=13333
		+16.7%	+6.7%

- No increase in total numbers
- Just one employee moved from München to Bielefeld
- Yet an increase in revenue per salesman at both POPs!

Will Rogers phenomenon (2)

- Will Rogers (1879–1935),
  American comedian and philosopher
- Named after one of his jokes:

  Frage: Wenn die 10% dümmsten Saarländer nach Rheinland-Pfalz ziehen, was passiert dann?

  Antwort: In beiden Bundesländern steigt der IQ an.

- (originally with Oklahomans and Californians…)

- Lesson:
  - Will Rogers phenomena are ubiquitous,
  - yet can be difficult to spot
  - …even for the authors themselves!
  - **Warning – it's a sword that cuts both ways:**
    Sometimes looking at the details is better, sometimes looking at the aggregated numbers makes more sense (as in the sales example)

Simpson Paradox (1)

- Universität Eschweilerhof discriminates against female students!
- Let's see what faculties are the most sexist ones:

	Applications				Acceptance rate	
Faculty	female	acc.	male	acc.	female	male
Engineering	10	8	80	50	80%	63%
CS	5	4	60	40	80%	67%
Philosophy	80	20	40	10	25%	25%
Law	30	15	40	10	50%	25%
Total	125	47	220	110	(←significant numbers)	
Acc. rate		37.6%		50.0%		

- None of them!? How can that be?
  - Women applied at faculties with more competition

Simpson Paradox (2)

- So who is right? Should the university be punished?
  - The women's rights activists? After all, 37.6% vs. 50% is significant – and dividing the total number into faculties simply introduces a bias into the picture.
  - The university? After all, not a single faculty does actually discriminate against women (in fact, most discriminate against men).
- Answer: In *this* case, the university is right
  - A student applies at a specific faculty that he or she chooses herself
  - A student does not apply at university and lets the university choose the faculty
- **Lesson:**
  - Simpson Paradox is more ubiquitous than you would think,
    yet can be difficult to spot …even for the authors themselves!
  - **Warning – it's a sword that cuts both ways:**
    Sometimes looking at the details makes more sense (as in this case), sometimes looking at the aggregated numbers is better.

Simpson Paradox (3)

Ärzte in Hansistan: Jahreseinkommen vor Steuern

	Schnibbler		Tröster		Knochenflicker	
	Flach-land	Berg-land	Flach-land	Berg-land	Flach-land	Berg-land
Anzahl Ärzte	200	40	20	20	30	100
davon über 200 000 Piepen	60	8	16	2	24	70
Anteil Best-verdiener in %	30	20	80	10	80	70

Wo verdienen Ärzte besser – im Flachland oder im Bergland? Im Detail deutet alles aufs Flachland hin.

Ärzte in Hansistan: Jahreseinkommen vor Steuern

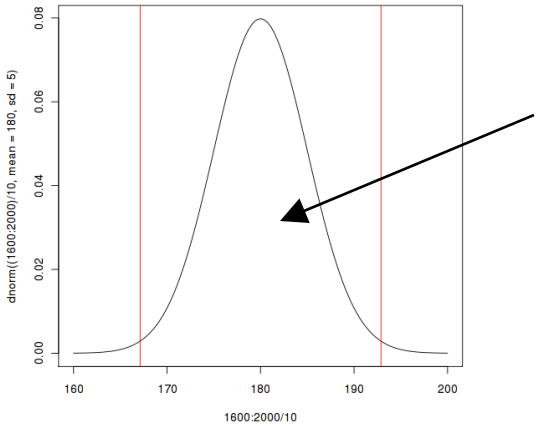
	Alle Ärzte	
	Flachland	Bergland
Anzahl	250	160
davon über 200 000 Piepen	100	80
Anteil Best-verdiener in %	40	50

Im Gesamtüberblick scheint das Bergland jedoch vorne zu liegen.

Philosophical / meta-aspects

Problem: Skew/leptokurtic distributions are not made for man(1)

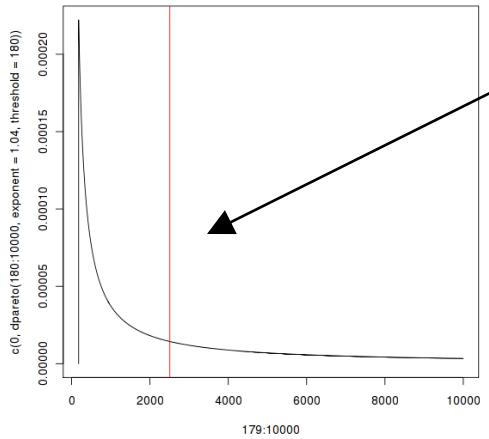
- In the stone age, man was surrounded mainly by more or less normally distributed (i.e., symmetrically distributed) random variables: Sizes of people, pregnancy durations, food consumption, etc.
  - Once you've seen a few samples, you get the picture
  - Outliers are rare
  - Outliers do not affect the mean (e.g., avg weight is 80kg, fattest man on earth weighs 400kg)



99% of all values between the red bars

Problem: Skew/leptokurtic distributions are not made for man(2)

- Today, man is surrounded by skew distributions with high kurtosis (leptokurtic), e.g., income (log-normal/ Pareto), earth quakes (Pareto), popularities (Zipf),…
  - Outliers like Dr. Waldner are comparably common – but you need more than just “a few” samples to see them
  - Outliers like Dr. Waldner do strongly affect the mean!



90% of all values right of red bar;
Median way more to the right;
Mean even waaaaaay more to the right

- Lesson: Ask: Is it a skew, leptokurtic distribution?

Catastrophe probabilities

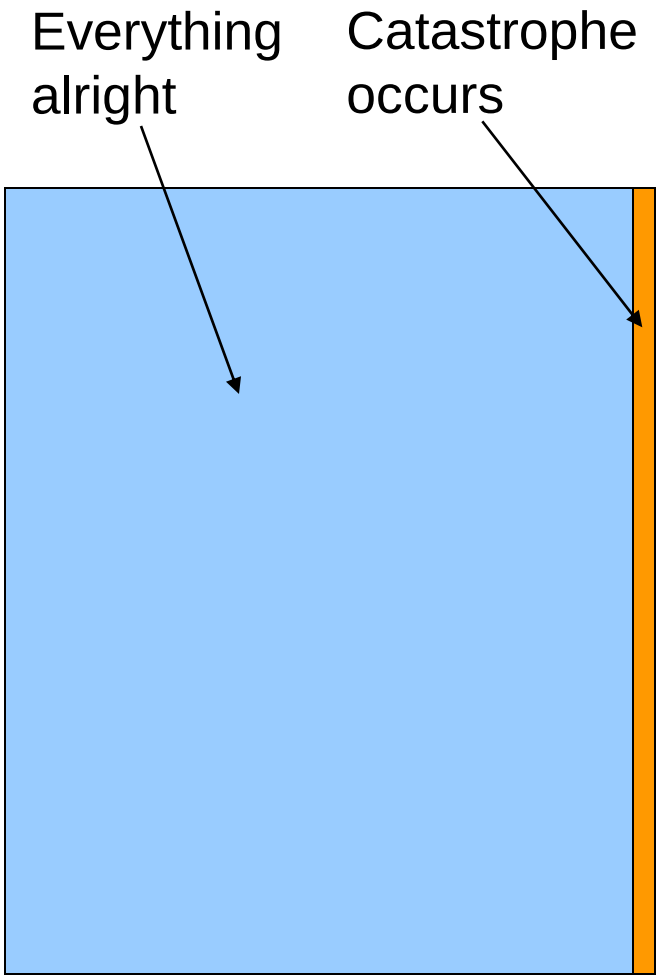
- Some (fictitious!) statements:
  - The probability that nuclear power plant X suffers a catastrophic accident is less than 10^{-10} per year
  - The probability that the AFDX avionics network in an aircraft fails is less than 10^{-11} per hour of operation
  - The probability that Rigel will burst into a supernova is less than 10^{-7} during the next thousand years
  - The probability for an eruption of the Laacher See volcano in the Eifel region is less than 10^{-8} during the next hundred years
- What do they have in common? (apart from being made up)
  - A [catastophic] high-impact event…
  - …with an extremely low probability

Low probabilities, high stakes

- On what grounds do these probabilities hold?
  - The underlying theory is correct
  - The underlying theory is applicable for the case being considered
  - The case being considered is really the general case, not a hidden special case
  - The confidence level for the result (if applicable) also shows a very high probability that the result is correct
  - The system under consideration has been correctly transformed into a correct theoretical model
  - The measurement data used to parameterize/calibrate the theoretical model has been measured correctly
  - The software that analyses the theoretical model (e.g., simulation, numerical analysis,…) has been correctly implemented
  - The hardware that executes the model software does not introduce errors (FDIV bug; RAM contents altered due to α particle decay; …)
- If just one condition fails, the entire probability calculation is flawed!

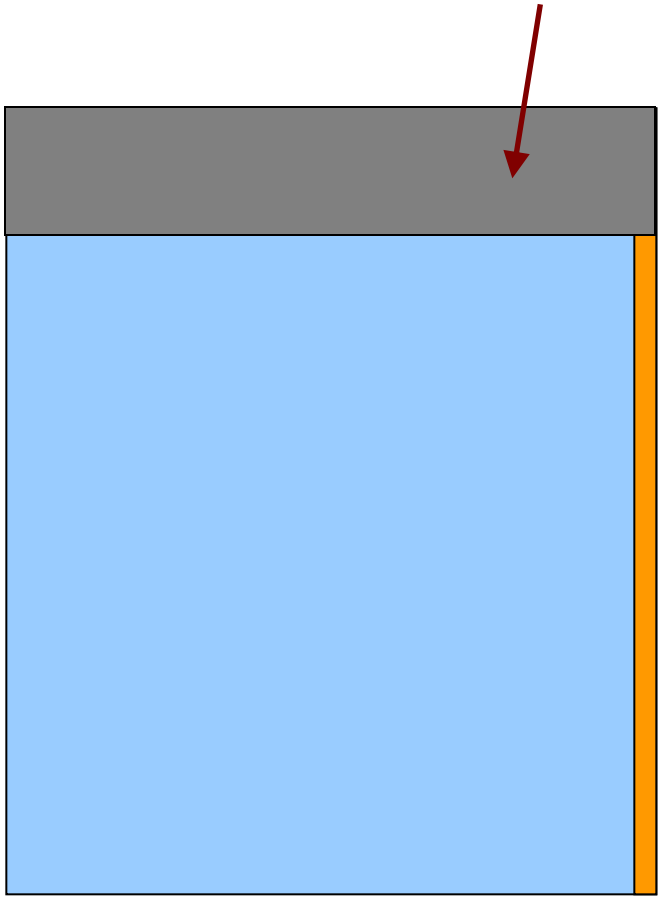
Low probabilities, high stakes

- Claim



- Reality

Don’t know, because the calculations are flawed



Low probabilities, high stakes

- Estimated probability that a scientific claim is flawed?
  - About 10^{-4}, according to the paper below
  - Mileage will vary – some more rigid, some less
- Consequences
  - Let's not take any risks!? No LHC, no SETI, no biotech, no ITER, no-nothing? Should we live in caves!?
  - Have we become too risk-averse?

- More information in this very readable paper:
  Ord, Toby, Hillerbrand:
  Probing the improbable: Methodological challenges for risks with low probabilities and high stakes.
  Journal of Risk Research, 2010

Lessons

- For authors:
  - Know your boundaries
  - Clearly state your assumptions
  - Clearly warn about possibilities that assumptions may not hold in reality

- For readers:
  - Double-check the assumptions
  - Ask for seconds, third, … opinions, preferrably using completely different methods

Risk aversion: How we lie to ourselves

- Do mobile phones cause cancer?
  - Very little evidence, long-term studies were needed
  - Result:
    - Possibly causes cancer
    - Only for people who use them for many hours per week
    - Still a very low incidence rate
- But many people try to get rid of base stations in their neighbourhood
  - “Well, it is just in case – you never know if there is something about those allegations”
- How often is calling an ambulance/the firemen via a mobile phone significantly faster than running to the nearest land-line phone?
  - How many “non-casualties” this way per year?

Risk aversion: How we lie to ourselves

- Do cars and motorcycles cause deaths?
  Yes, and very much so:
  - About 4,000 casualties in Germany per year (p.a.) due to traffic accidents
  - About 80,000,000 inhabitants in Germany
  - Roughly 800,000 people die in Germany p.a.
- Incidence:
  About 0.5% of all deaths are traffic accidents!
  - That's just the deaths. We are ignoring other serious consequences such as mutilations, month-long recovery treatments, psychological traumata, financial losses, etc.
- Compare: How many % of all deaths in Germany are directly or indirectly linked to mobile phones p.a.?

Risk aversion: How we lie to ourselves

- `Reproduction` is fun! (if done on purpose…)

- But what about the risks?
  - Mortality among mothers in labour: 80 ppm => 0.008%
  - Risk that the child suffers from a chromosome aberration (trisomy 21/Down syndrome, Cri du Chat, trisomy 18, trisomy 13, etc.): about 1/160 = 0.63%

- Would you enter a car if the risk of having a serious accident (fatal or heavy injuries) were 0.63% per…
  - Per journey?
  - Per 100km?
  - Per 10,000km?
  - Per car lifetime?

Risk aversion: How we lie to ourselves

- Lessons:
  1. Often, we take risks without noticing their true extent (even though we actually know it)
  2. Often, we refuse taking risks that are magnitudes smaller than those from point 1.
  3. Most occurrences of point 2 do not make any sense, but we just do not notice.
  4. On the other hand: If we are aware of these phenomena, if we counter them by acting “rationally” against our intuition/common standards, and *then* the unlikely accident happens, we will feel very guilty, and everybody will say “I told you so”…
  5. Also note that we mostly are talking about very low probabilities again…

Summary

- When confronted with data or conclusions from data one should always ask:
  - Can they possibly know this? How?
  - What do they really mean?
  - Is the purported reason the real reason?
  - Are the samples and measures unbiased and appropriate?
  - Are the measures well-defined and valid?
  - Are measures or visualizations misleading?
  - Has something important been left out?
  - Are there any inconsistencies (contradictions)?

- When we collect and prepare data, we should
  - work thoroughly and carefully
  - check our assumptions and prerequisites
  - avoid distortions of any kind