Class overview today - March 21, 2016

- Part I - Common statistical methods in geoscience
  - The many challenges of geological data samples
  - Uncertainty in Earth science data
  - Basic mathematical representations of measurement uncertainty
- Part II - What do geochronological ages mean?
  - Geochronological ages and their meaning
  - Comparing predicted and measured ages
  - Quantifying the fit of predicted and measured ages

Introduction to Quantitative Geology
Lecture 3
Common statistical methods in geoscience

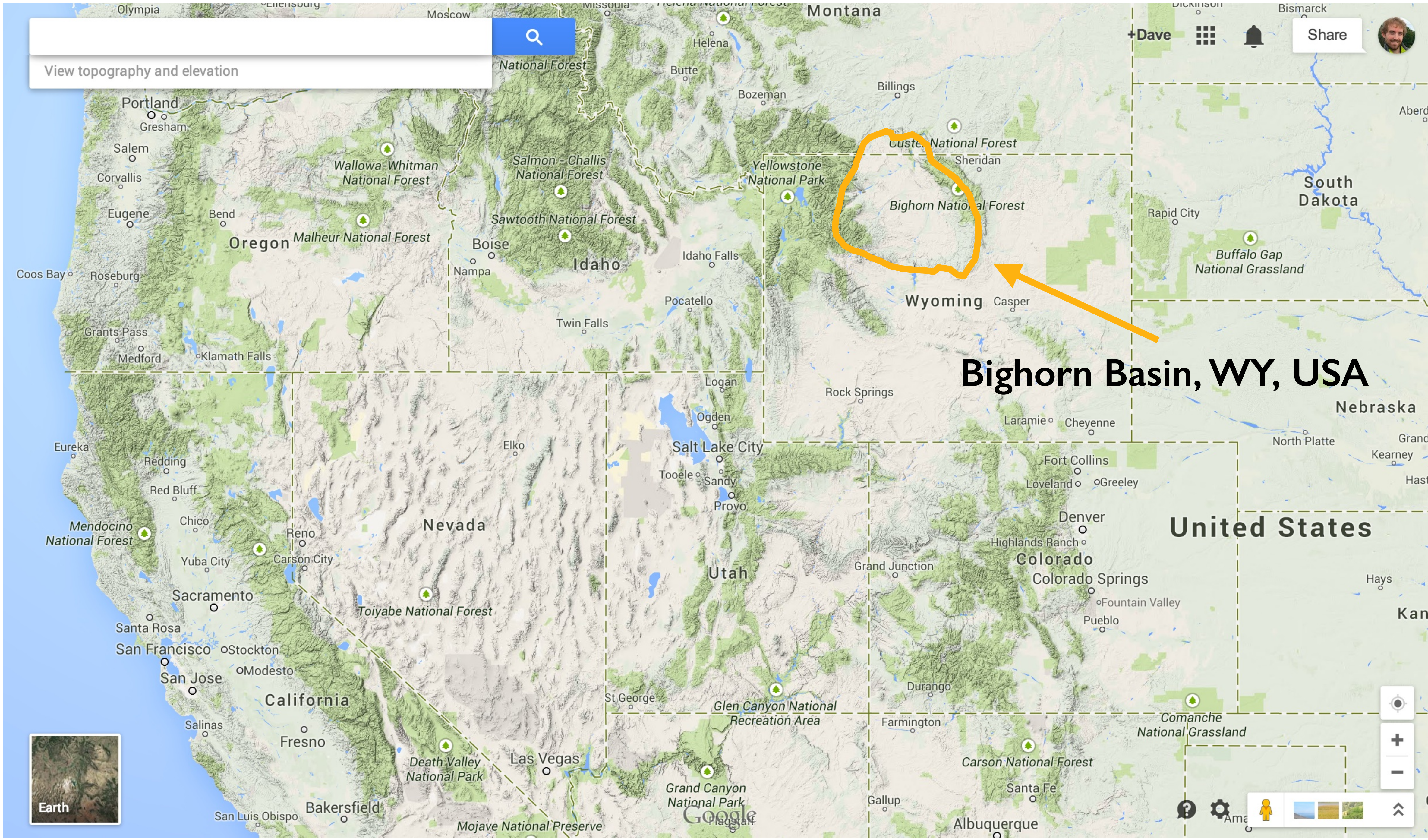
Lecturer: David Whipp
david.whipp@helsinki.fi

21.3.2016

Goals of this lecture

- Address the many challenges of **geological data samples**
- Discuss **uncertainty** in Earth science data
- Review the basic mathematical representations of **measurement uncertainty**

Why are statistics important for us?



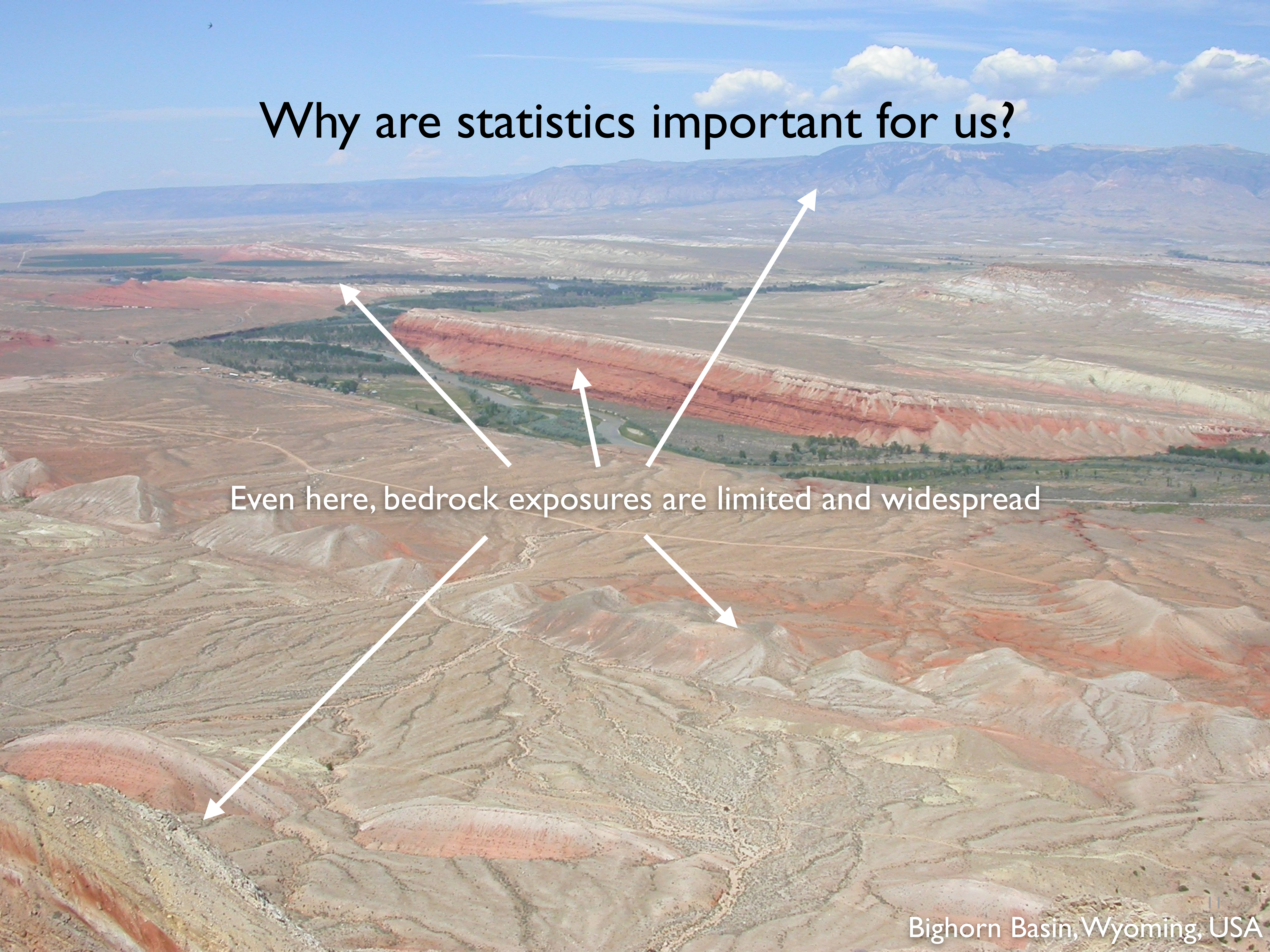
Bighorn Basin, Wyoming, USA


Bighorn Mtns, Wyoming, USA

Bighorn Basin, Wyoming, USA

Bighorn Basin, Wyoming, USA

Bighorn Basin, Wyoming, USA


Why are statistics important for us?
Even here, bedrock exposures are limited and widespread
Bighorn Basin, Wyoming, USA

Why are statistics important for us?

- We can directly observe only a tiny fraction of most geological features
  - Bedrock exposure is limited
  - Many rock units are very large and/or widely dispersed
  - Sometimes, even well-exposed units can be hard to access

Bighorn Basin, Wyoming, USA

A few statistical definitions

- **Population**: The total number of occurrences of a particular thing present in a defined area
    - We basically cannot access this geologically

A few statistical definitions

- **Population**: The total number of occurrences of a particular thing present in a defined area
  - We basically cannot access this geologically
- **Subset**: A collected portion of the population
- **Sampling units**: The collected material

A representative sample

- Our goals, as Earth scientists is to collect data that forms a **representative sample**
  - **Representative sample**: A sample that can be used to infer the characteristics of the population

- The best way to get a representative sample is to collect sampling units at **random**, without bias

A representative sample

- Our goals, as Earth scientists is to collect data that forms a **representative sample**
    - **Representative sample**: A sample that can be used to infer the characteristics of the population

- The best way to get a representative sample is to collect sampling units at **random**, without bias
    - **What challenges might we face in collecting a random geological sample?**

The trouble with "Random" sampling

Population

- Making our best effort to collect a "random" sample does not guarantee it is a representative sample

Figure 1.1, McKillup and Dyar, 2010

The trouble with "Random" sampling

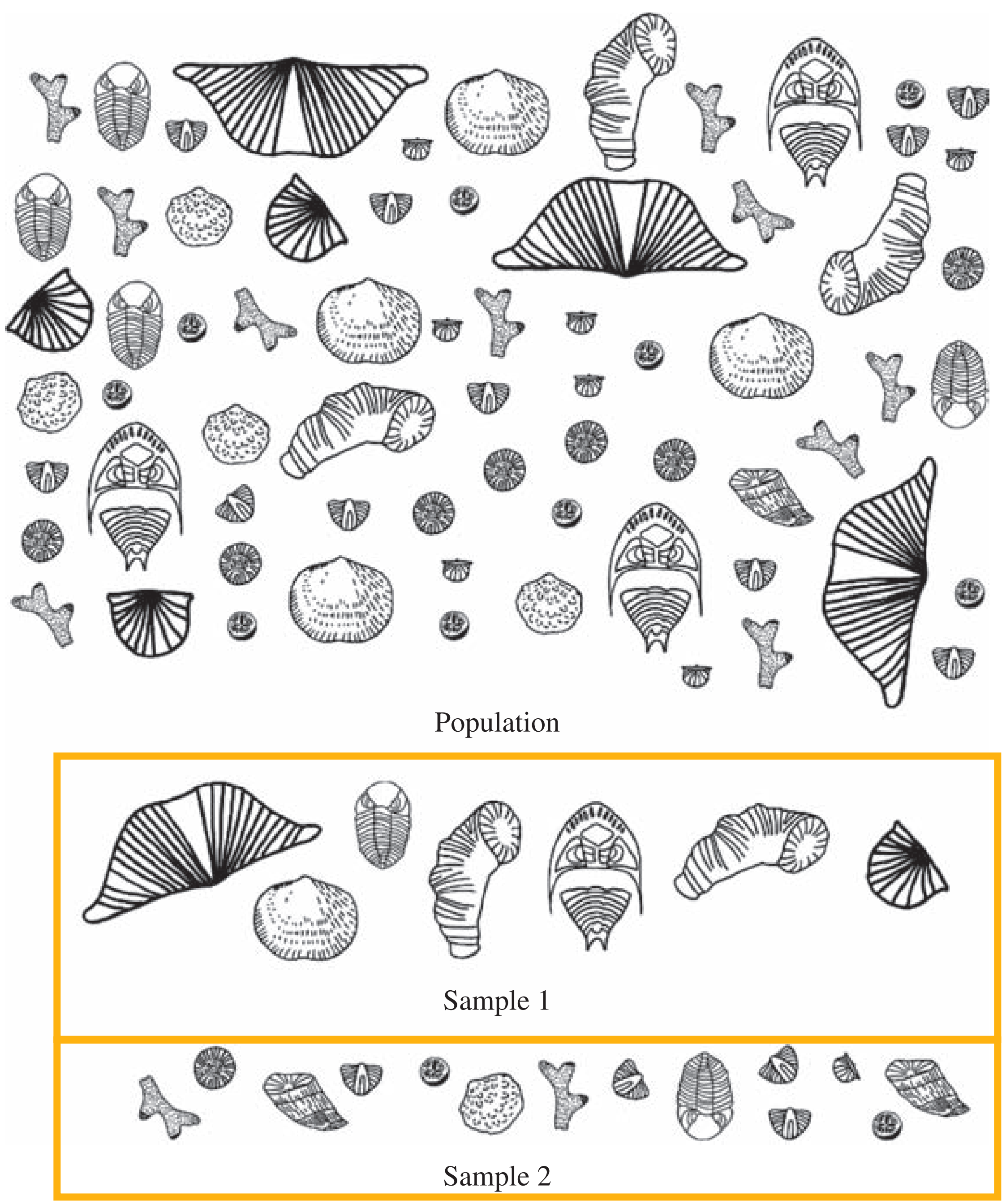


Figure 1.1, McKillup and Dyar, 2010

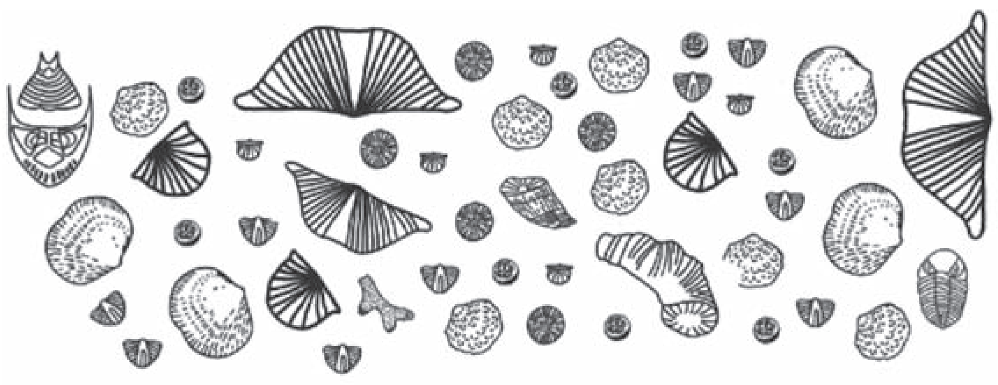
- Making our best effort to collect a "random" sample does not guarantee it is a representative sample

1. Samples taken from the same population may be very different from one another

Large fossils

Small fossils

The trouble with “Random” sampling



- Making our best effort to collect a “random” sample does not guarantee it is a representative sample

Figure 1.2, McKillup and Dyar, 2010

The trouble with "Random" sampling

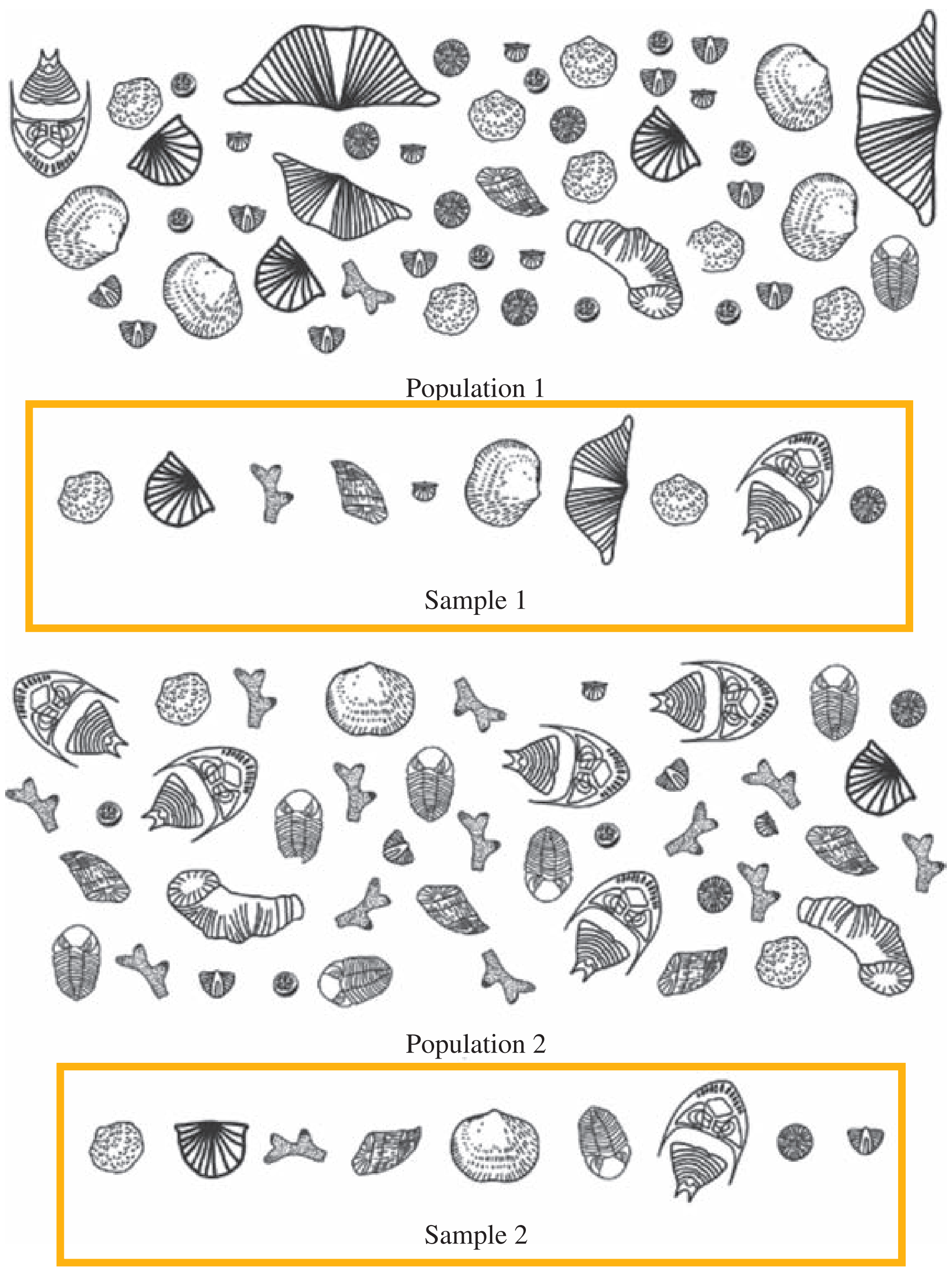


Figure 1.2, McKillup and Dyar, 2010

- Making our best effort to collect a "random" sample does not guarantee it is a representative sample

2. Samples taken from very different populations may be quite similar

Using a sample to infer something about a population

- Collecting a representative sample is a clear challenge and needed to be able to make inferences about a population
  - Every precaution must be taken to collect a representative sample
    - Collecting a sufficient number of measurements
    - Making measurements or collecting samples in the most logical locations
    - Carefully preparing samples for analysis or measurement

Uncertainty

Uncertainty

Deep water oil/gas exploration is expensive
Drill ship: ~350,000 € per day
Oil platform: ~300,000 € per day

Uncertainty

Deep water oil/gas exploration is expensive
Drill ship: ~350,000 € per day
Oil platform: ~300,000 € per day

Know your uncertainties!

Uncertainty in Earth science

- In addition to challenges collecting a representative sample, all geological measurements have inherent **uncertainty**
    - It is not possible to make an exact measurement
    - We may use tools with very high precision, but measurements are still uncertain

- A scientist's goal is to reduce uncertainty when taking measurements by making them as **accurate** and **precise** as possible
- Our goal should be to always include that uncertainty when comparing measured values to predictions

Precision versus accuracy

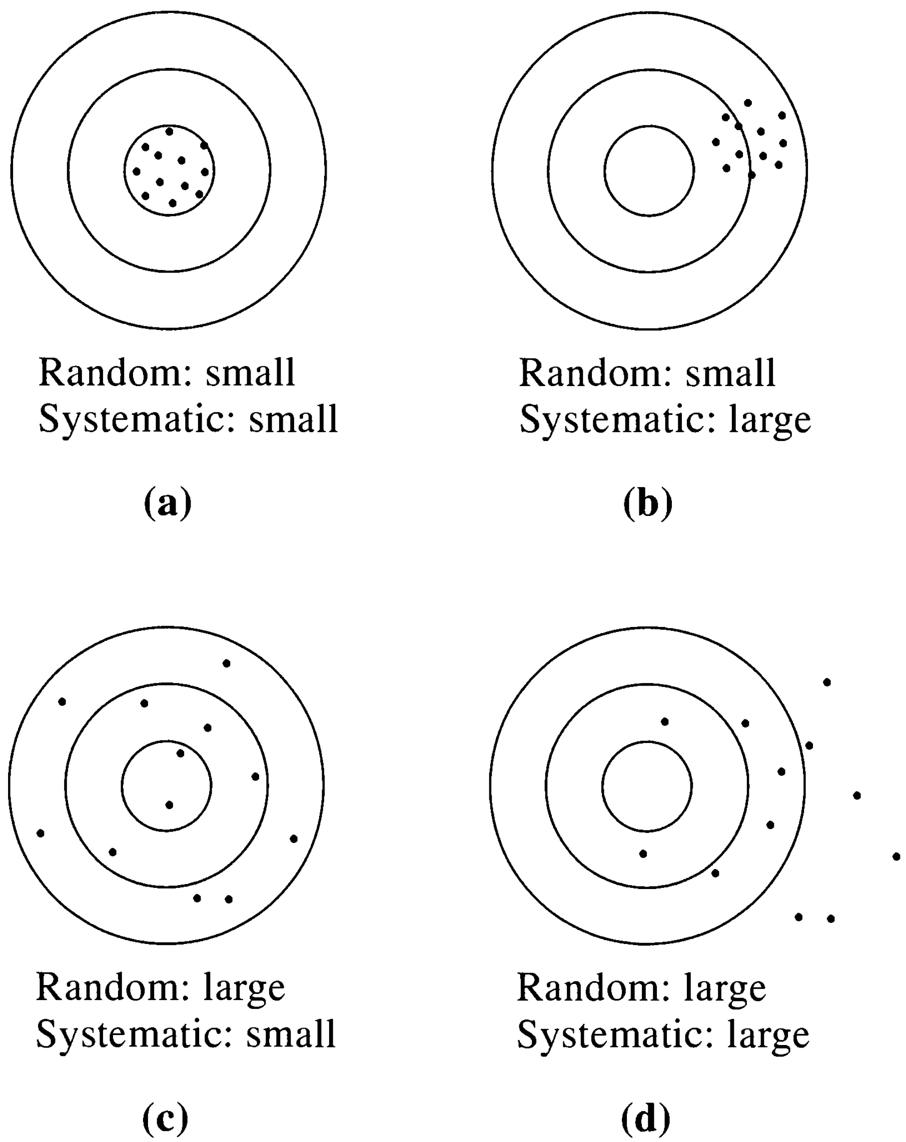


Fig. 4.1, Taylor, 1997

- **Random error**: Experimental uncertainty revealed by repeated measurements
  - Small random error = **Precise**

Precision versus accuracy

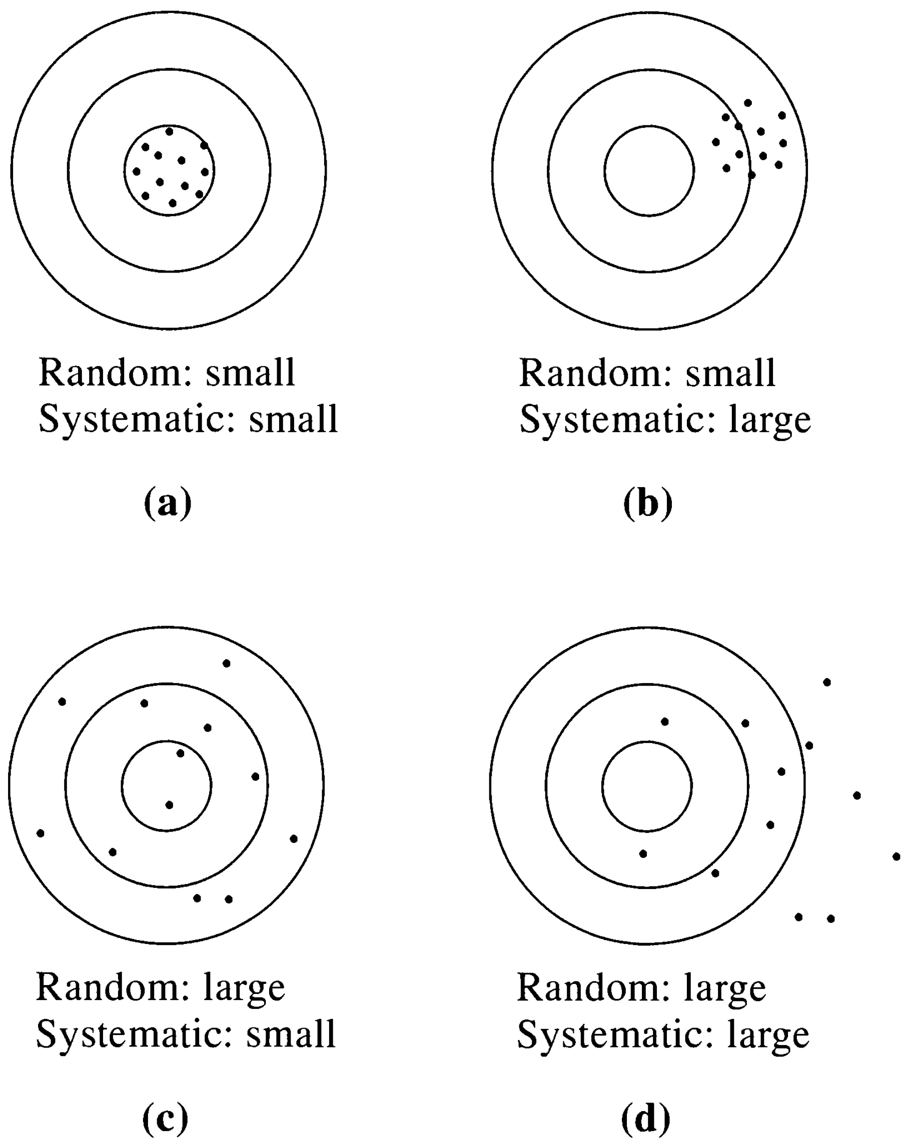


Fig. 4.1, Taylor, 1997

- Random error: Experimental uncertainty revealed by repeated measurements
  - Small random error = **Precise**
- Systematic error: Error that cannot be revealed by repeated measurement
  - Small systematic error = **Accurate**

Precision versus accuracy

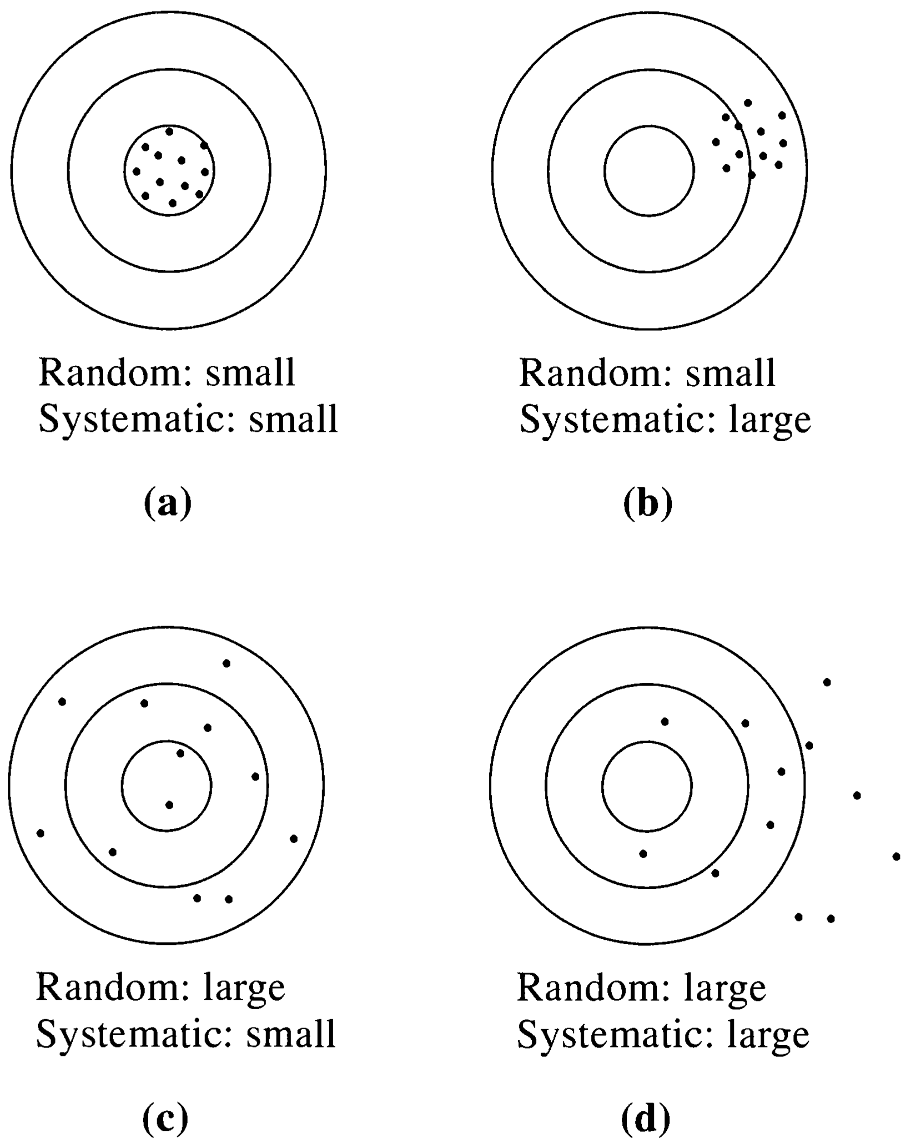


Fig. 4.1, Taylor, 1997

- **Random error**: Experimental uncertainty revealed by repeated measurements
  - Small random error = **Precise**
- **Systematic error**: Error that cannot be revealed by repeated measurement
  - Small systematic error = **Accurate**
- **How might we detect systematic error?**

Precision versus accuracy

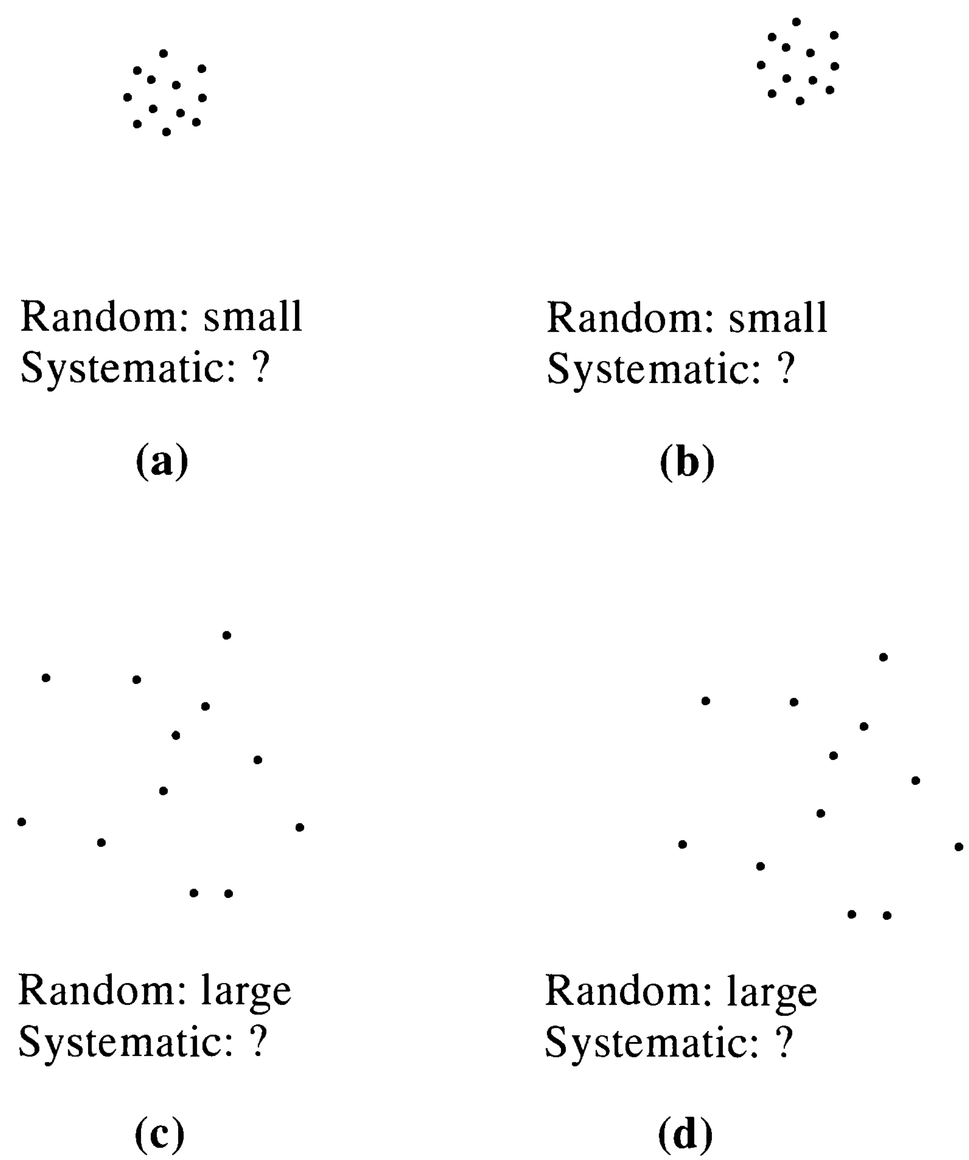
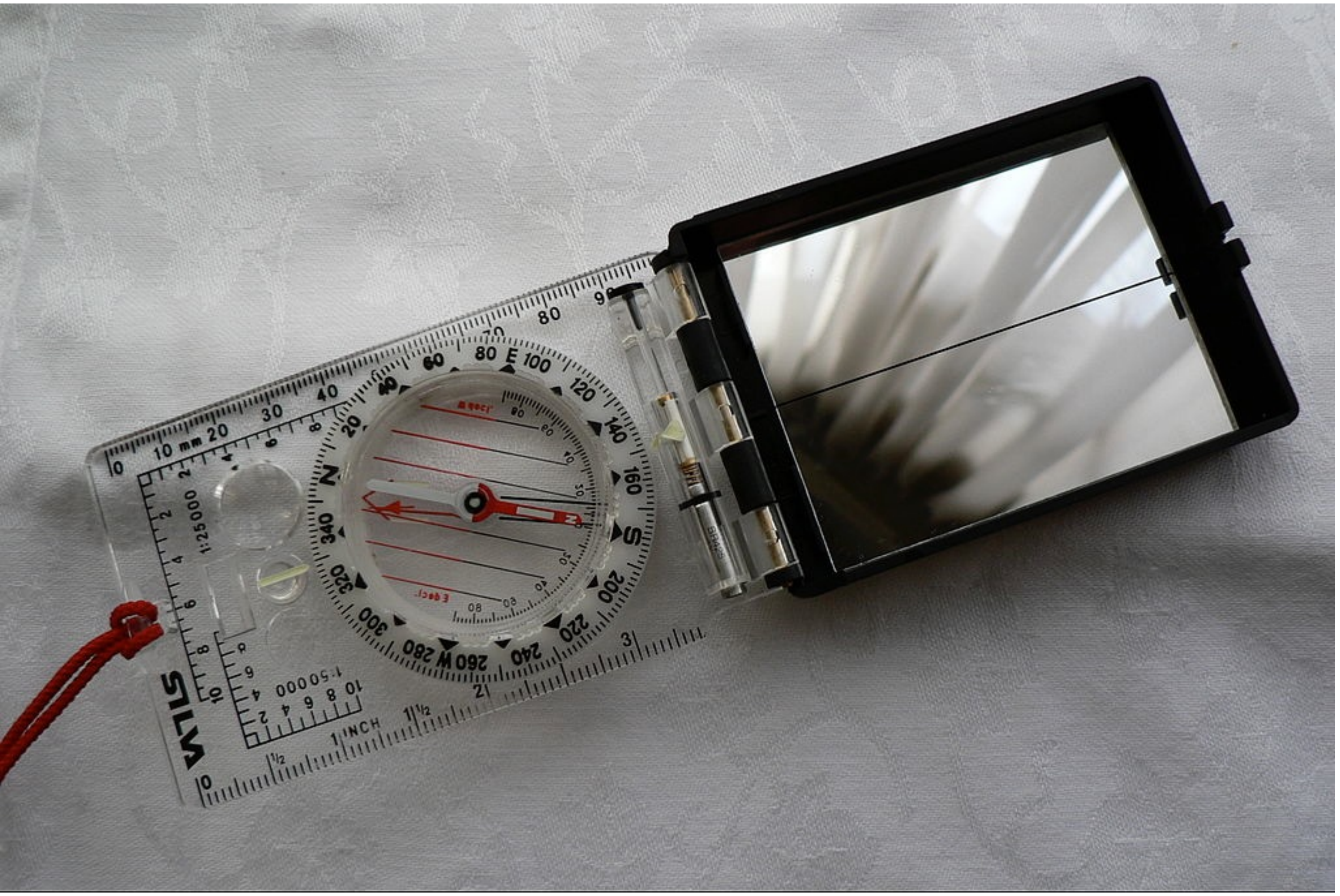


Fig. 4.2, Taylor, 1997

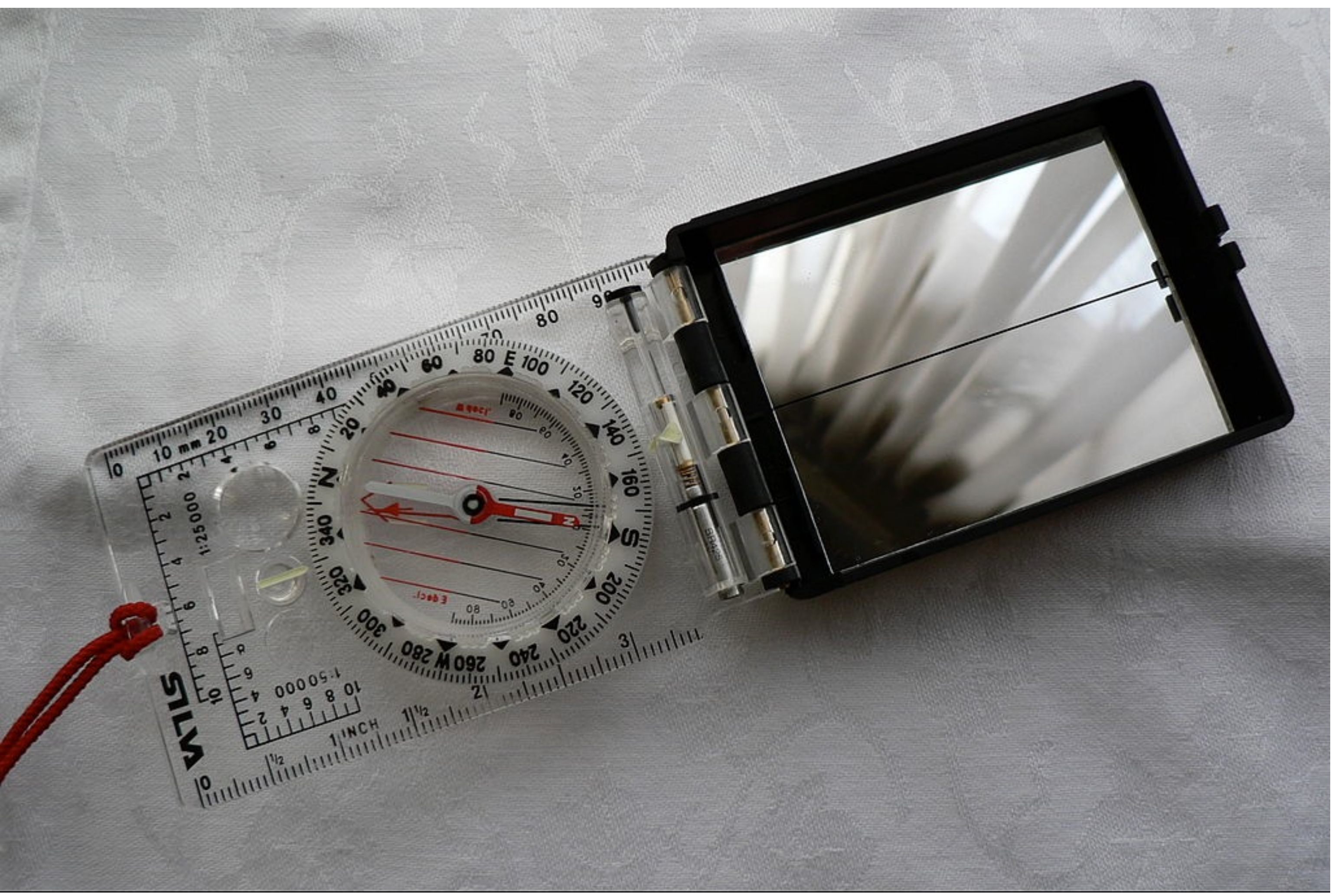
- As it turns out, detecting systematic error is difficult
- Scatter might reveal large random errors, but without a reference for the true value, the systematic error is unclear
    - This is one reason that blanks and standards are used for measuring isotope concentrations when measuring geochronometer ages, for example

Sources of uncertainty

- If we think about it, uncertainty is everywhere



Sources of uncertainty



- If we think about it, uncertainty is everywhere
  - A typical field compass is graduated in 1°, so we can't expect any higher precision
  - Magnetic declination can also only be set to within 1°, and the uncertainty larger if not set correctly
  - Outcrops generally don't provide perfect exposure, and repeated measurements vary often by 3-5° (at least)

Reported measurements

- Ignoring the declination, the most precise measurement we can make with our compass is

$$x_{\text{best}} \pm 0.5°$$

- Imagine we take 5 measurements of the strike of a rock unit and find: 33°, 36°, 32°, 35°, 34°
  - Based on these numbers alone, we could state the strike is 34±2°
  - Including the uncertainty in reading the compass, we would say the strike is 34±2.5°
- Generally, any measurement should be reported in the form

$$x_{\text{best}} \pm \delta x$$

Mean and standard deviation

- For the compass measurements, it is quite natural to take the average value, or **mean**, for the best estimate of the strike
- The **mean**, $\bar{x}$, is simply the sum of the measured values divided by the number of measurements

$$\bar{x} = \frac{x_1 + x_2 + ... + x_N}{N}$$
$$= \frac{\Sigma x_i}{N}$$

where x_i is the ith measured value, N is the total number of measurements and Σ represents sigma notation of a sum

$$\sum_i x_i = \sum_{i=1}^{N} x_i = x_1 + x_2 + ... + x_N$$

Mean and standard deviation

- For our example, calculating the mean is easy

$$\bar{x} = \frac{33^\circ + 36^\circ + 32^\circ + 35^\circ + 34^\circ}{5}$$
$$= 34^\circ$$

- If we consider the **mean** as the best estimate of the quantity, a natural value to consider as well is the **deviation** (or **residual**) for each measurement

$$d_i = x_i - \bar{x}$$

- If the **deviation** is small, the measurements are **precise**; if it is large, the measurements are **not so precise**

Mean and standard deviation

- It might be tempting to average the **deviation** to determine how much measurements deviate from the mean on average
  - **What is the average deviation in our dataset?**

Measured value	Mean	Deviation
33	34	-1
36	34	2
32	34	-2
35	34	1
34	34	0

Mean and standard deviation

- It might be tempting to average the **deviation** to determine how much measurements deviate from the mean on average
  - **What is the average deviation in our dataset?**
- If we square the **deviation**, all values will be positive, and then we can average and take the square root of the result

$$\sigma_x = \sqrt{\frac{1}{N-1}(d_i)^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Mean and standard deviation

- It might be tempting to average the **deviation** to determine how much measurements deviate from the mean on average
  - **What is the average deviation in our dataset?**
- If we square the **deviation**, all values will be positive, and then we can average and take the square root of the result

$$\sigma_x = \sqrt{\frac{1}{N-1}(d_i)^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- This is the **sample standard deviation**, σ_x

Mean and standard deviation

Measured	Mean	Deviation	Deviation squared
33	34	-1	1
36	34	2	4
32	34	-2	4
35	34	1	1
34	34	0	0

- If we plug in our values we see

$$\sigma_x = \sqrt{\frac{1}{N-1}(d_i)^2}$$

$$= \sqrt{\frac{1}{4}(1+4+4+1+0)}$$

$$= 1.6$$

What does the standard deviation tell us?

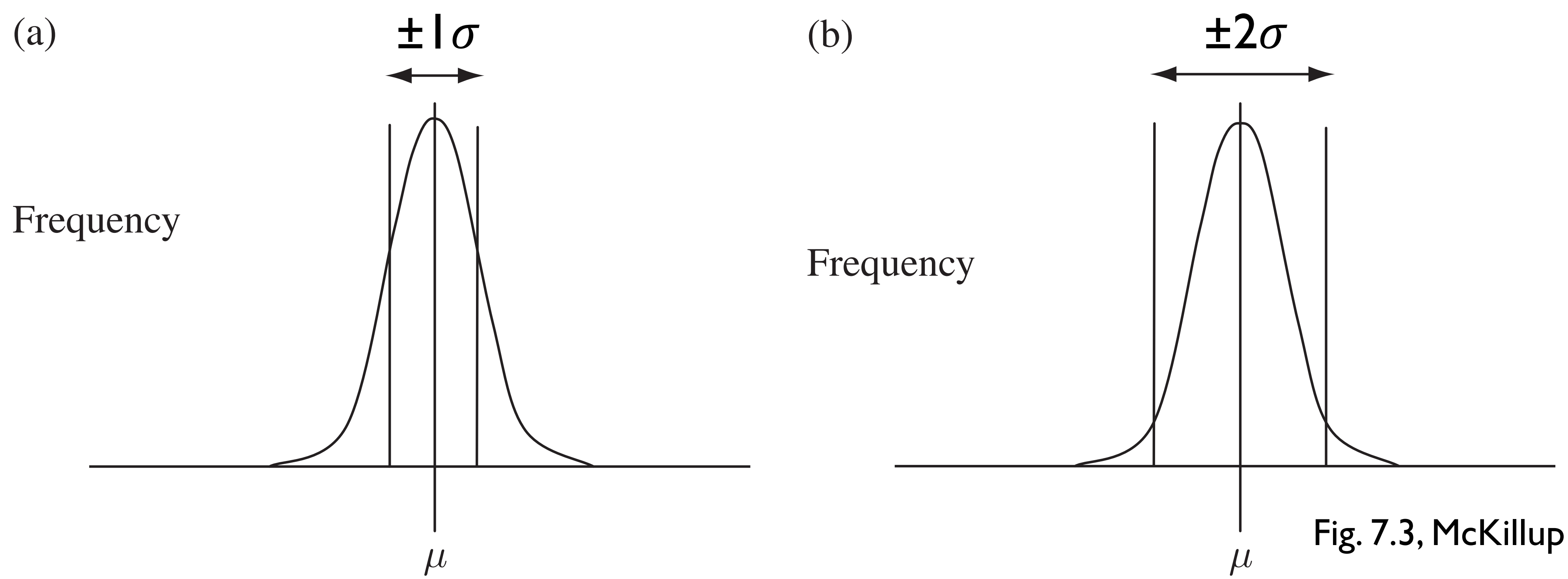
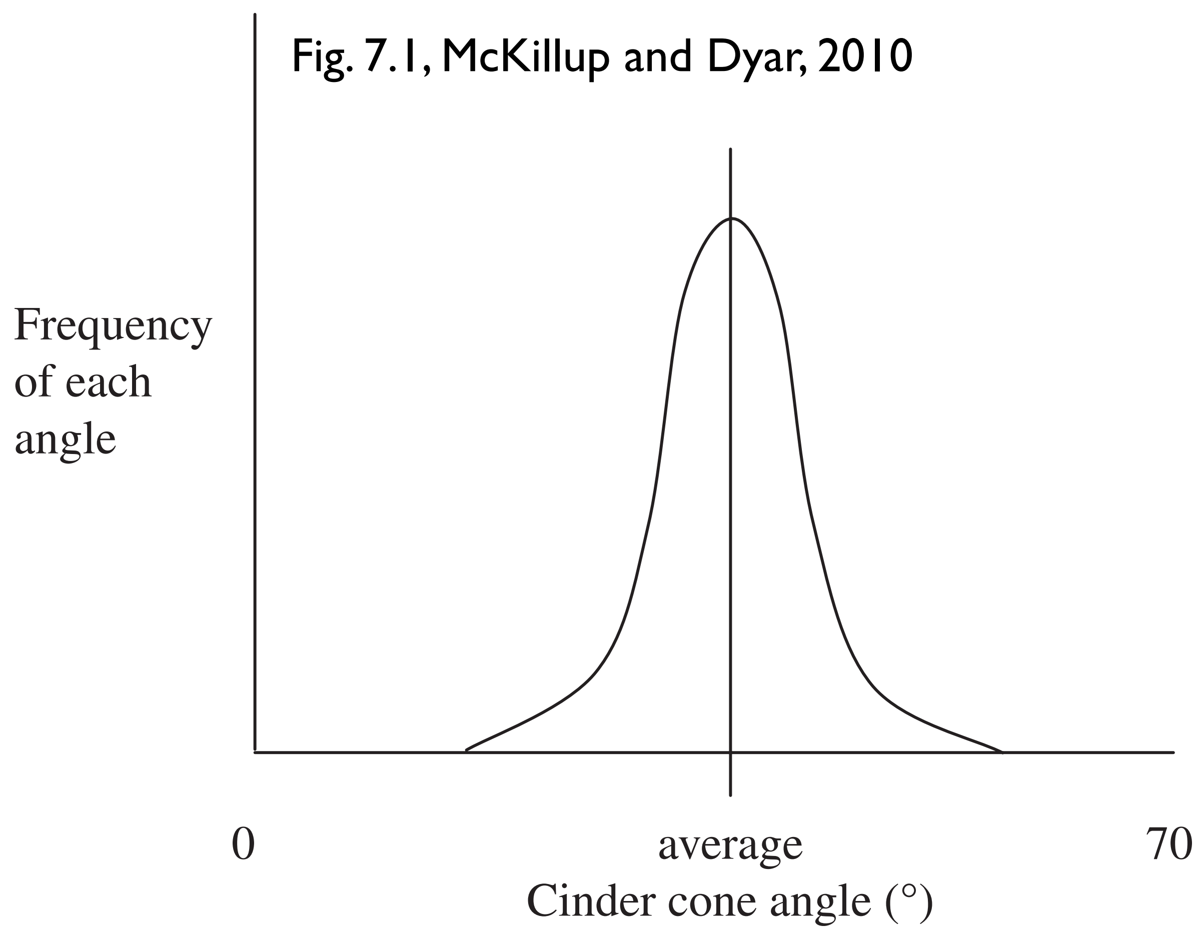


Fig. 7.3, McKillup and Dyar, 2010

~68% of measurements within ±1σ ~95% of measurements within ±2σ

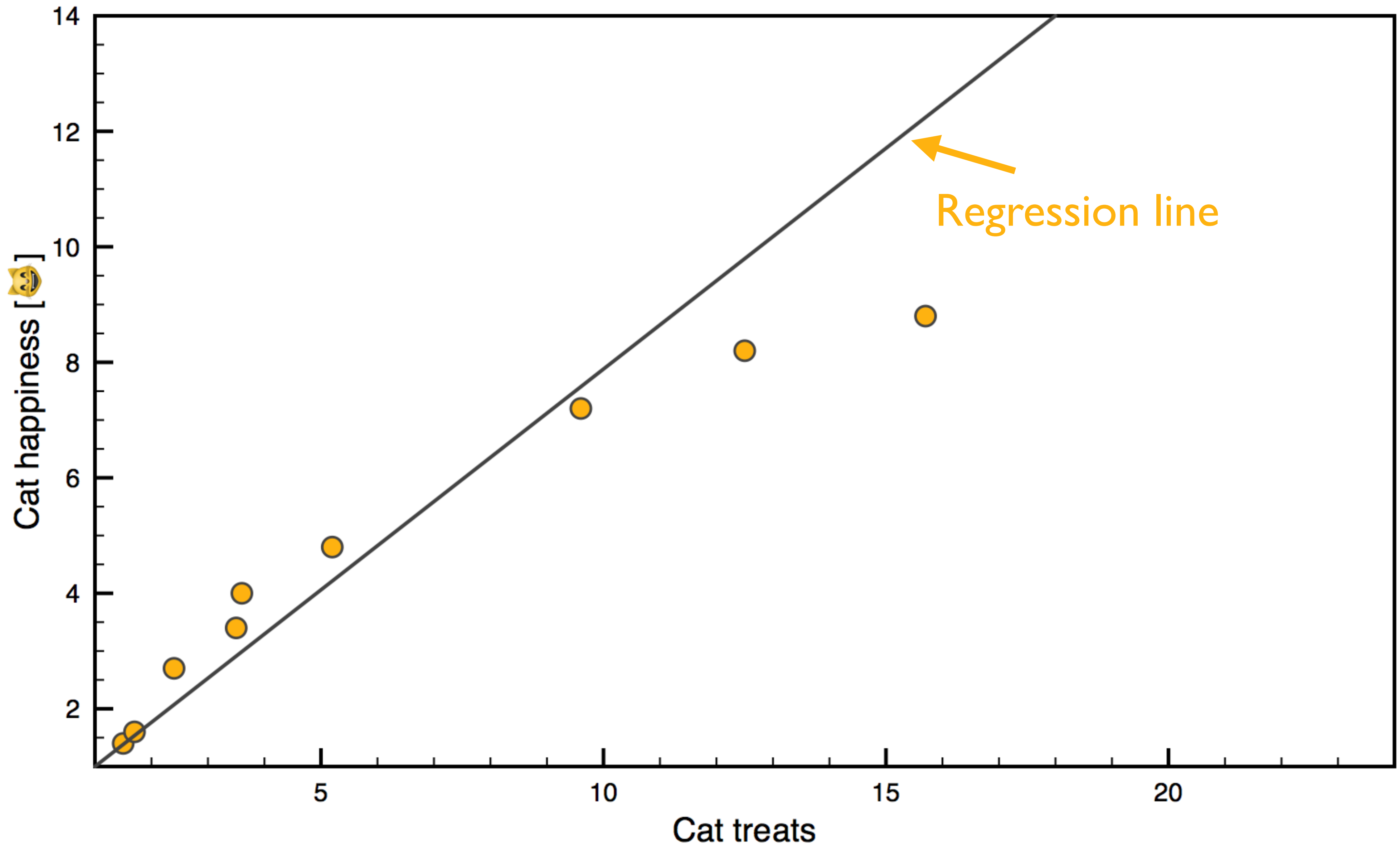
- The 2σ uncertainty is often reported for geological measurements
  - For the measured values, ~95% of the measurements are within ±2σ of the mean μ (μ is the population mean, $\bar{x}$ is the sample mean)

Normal distribution

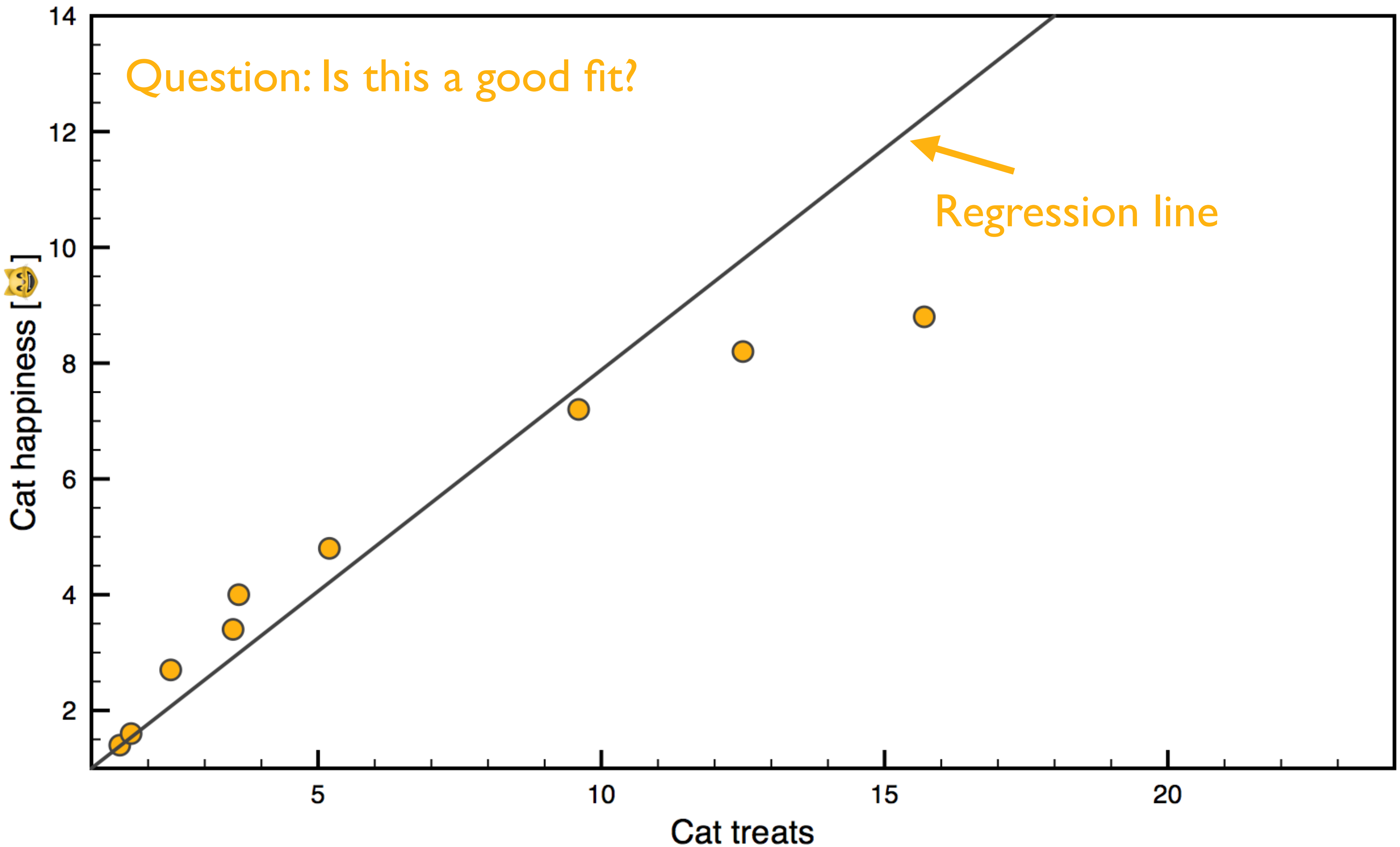


- The **normal distribution** refers to a bell-shaped curve that mathematically describes the frequency that a measured value is observed
    - This is also known as a **Gaussian curve**
- Many geological variables follow a **normal distribution**
- We will look at more on Wednesday

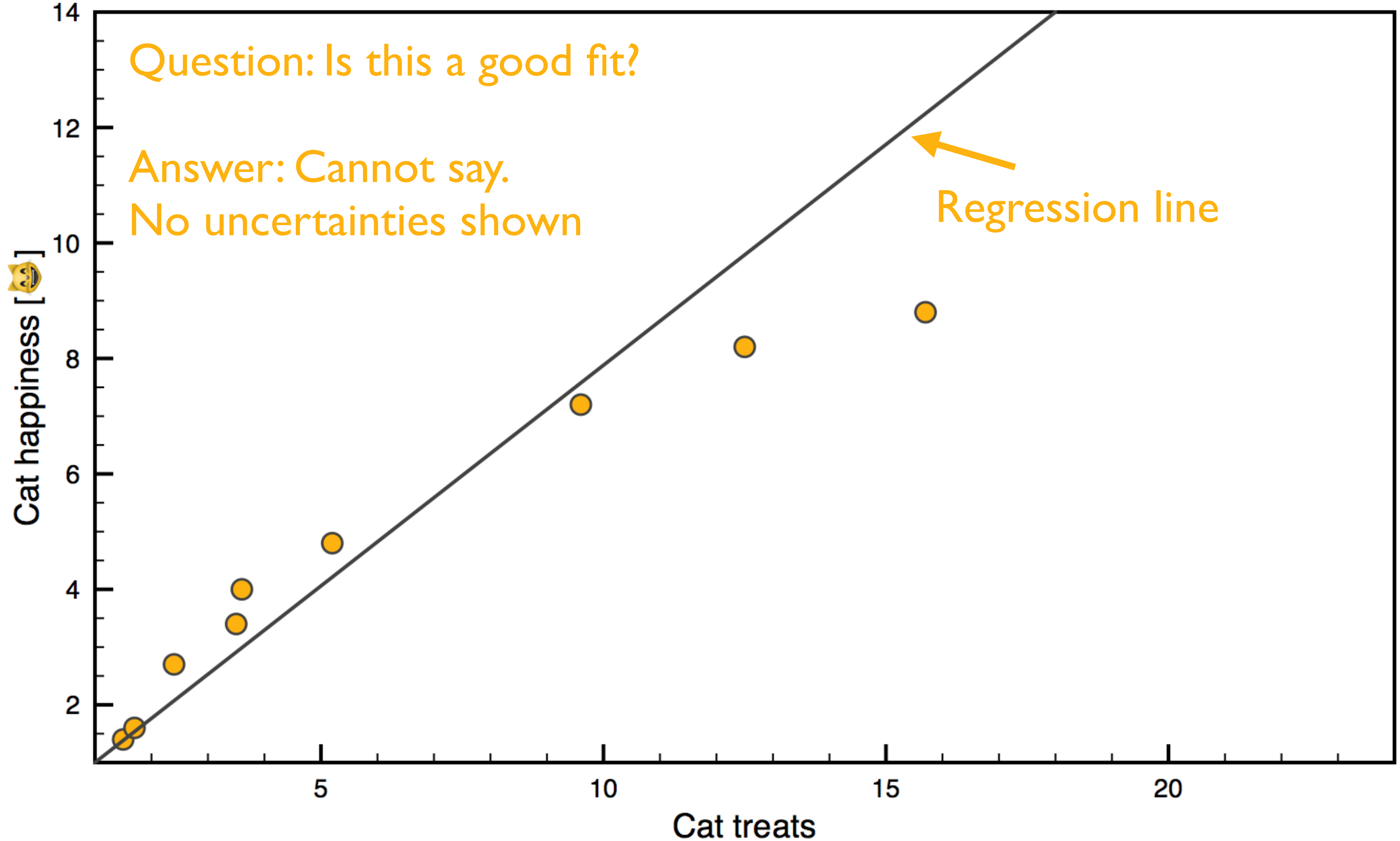
A silly example of why uncertainties matter



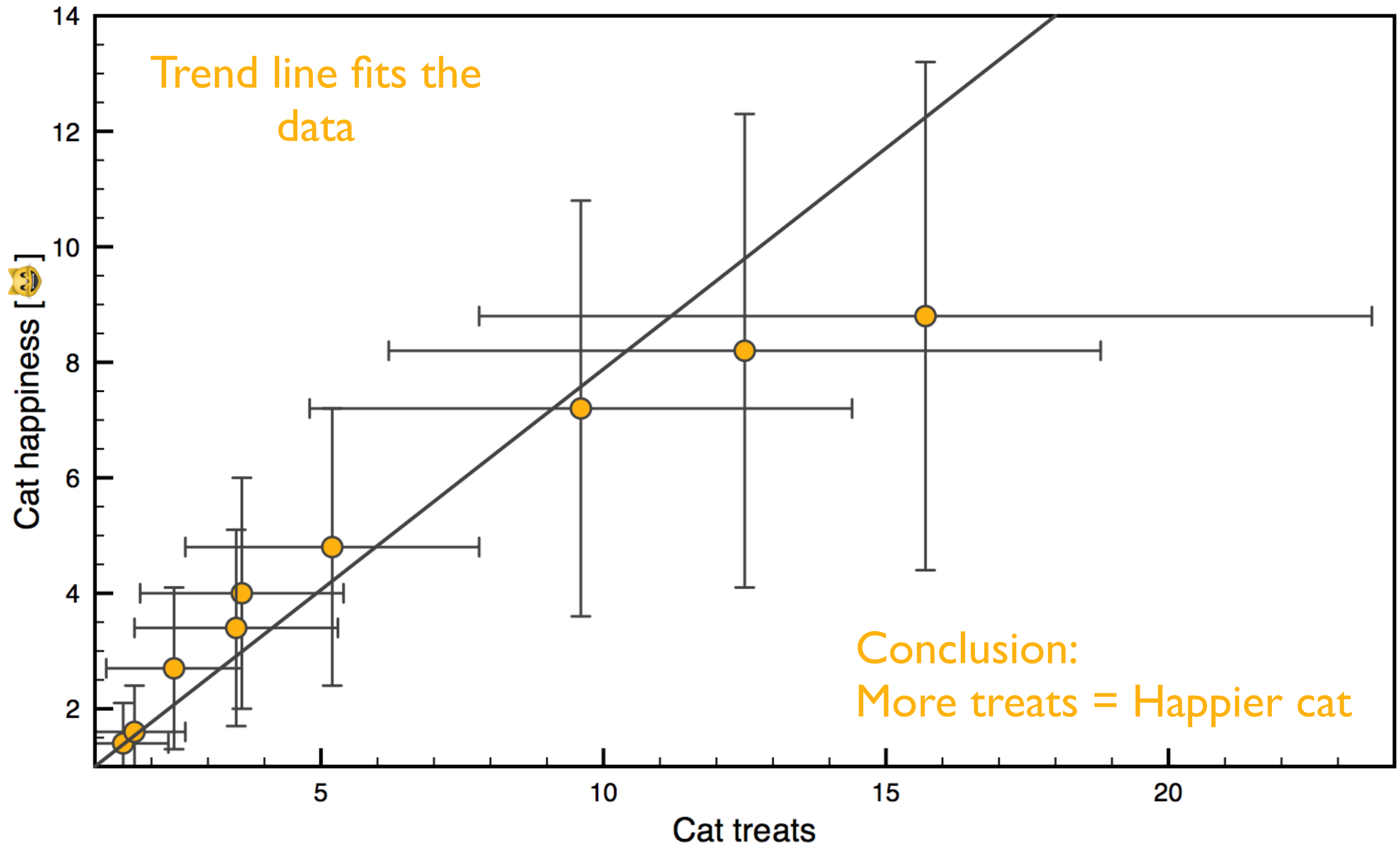
A silly example of why uncertainties matter



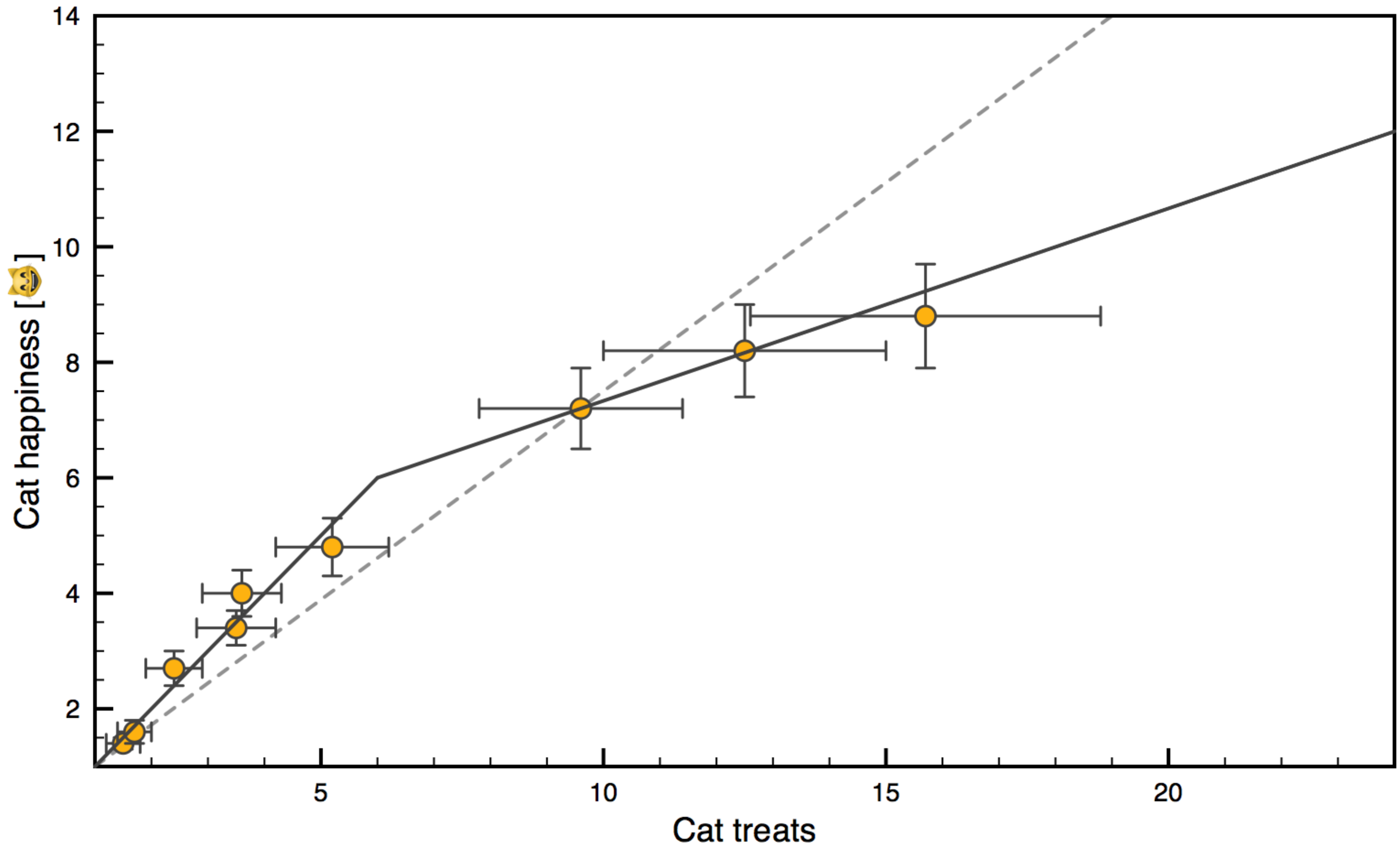
A silly example of why uncertainties matter



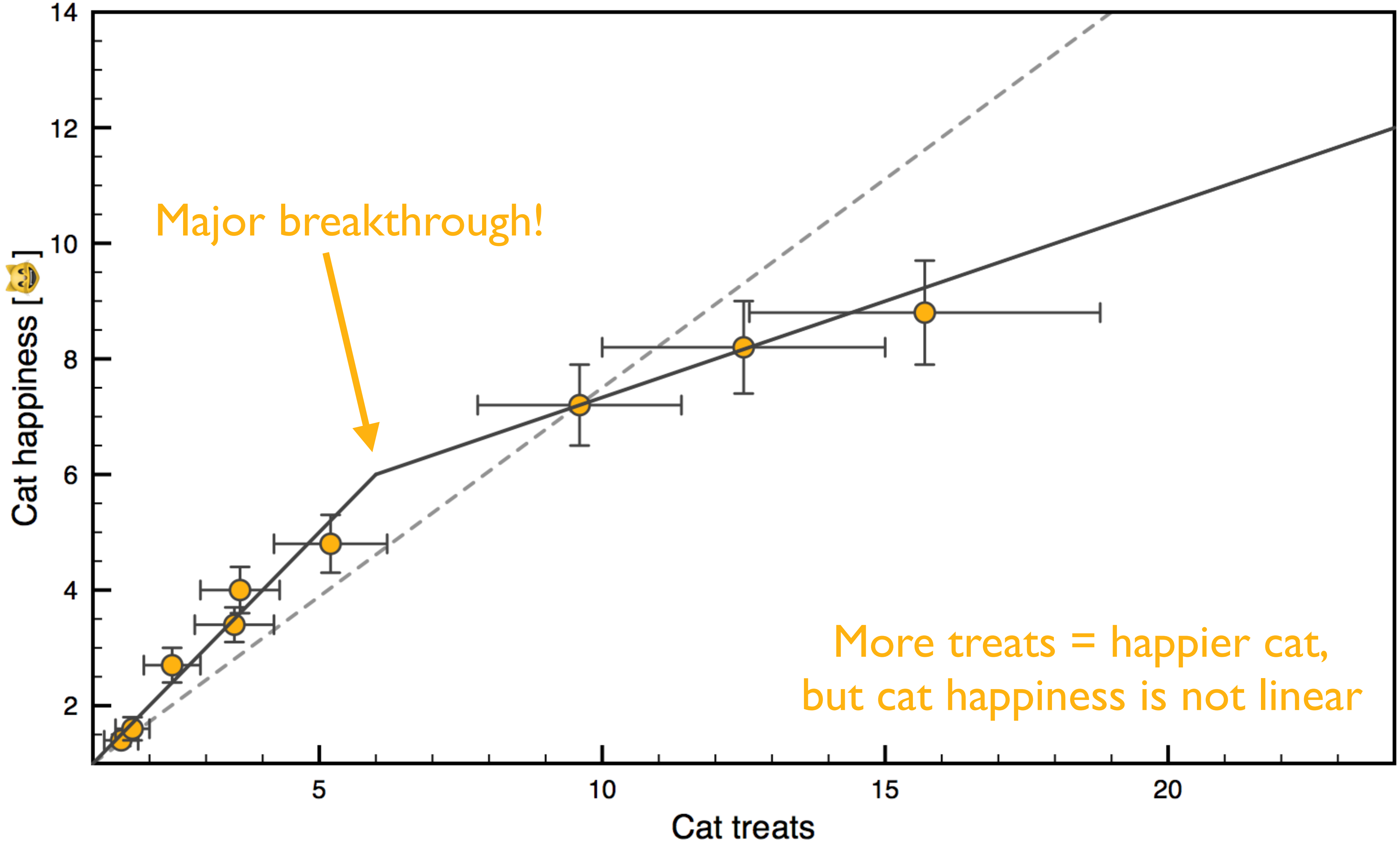
A silly example of why uncertainties matter



A silly example of why uncertainties matter



A silly example of why uncertainties matter



Reporting uncertainties

- The point is uncertainties matter
  - **What our data can say is a function of the uncertainty**
  - A model fit to data is meaningful only if the predictions lie within the uncertainty in the data

- A geological age of 1.88 Ga for an event might be interesting, but an age of 1.88±0.02 Ga is far more powerful

Recap

- What are some of the challenges collecting geological data samples?
- What are some common sources of uncertainty in Earth science data?
- How should you mathematically represent uncertainty in any measurement?

Recap

- What are some of the challenges collecting geological data samples?

- What are some common sources of uncertainty in Earth science data?

- How should you mathematically represent uncertainty in any measurement?

Recap

- What are some of the challenges collecting geological data samples?
- What are some common sources of uncertainty in Earth science data?
- How should you mathematically represent uncertainty in any measurement?

References

McKillup, S., & Dyar, M. D. (2010). *Geostatistics explained: an introductory guide for earth scientists*. Cambridge University Press.

Taylor, J. R. (1997). *An introduction to error analysis: the study of uncertainties in physical measurements*. Sausalito, CA, USA: University science books.