

What is data?

Why should we look at data?

Oliver W. Layton

CS251: Data Analysis and Visualization

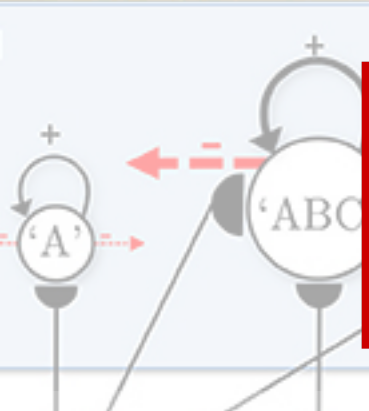
Lecture 1

Spring 2019

My academic journey



Skidmore '09



BOSTON
UNIVERSITY

The Boston University logo, which includes a red square with a white border containing the text "BOSTON UNIVERSITY" in white, serif capital letters. To the left of the square is a circular emblem featuring a stylized 'A' and 'B' with a red arrow pointing from 'A' to 'B'.

Ph.D. Cognitive and
Neural Systems '13



Rensselaer

Postdoc
2013-2018

Colby

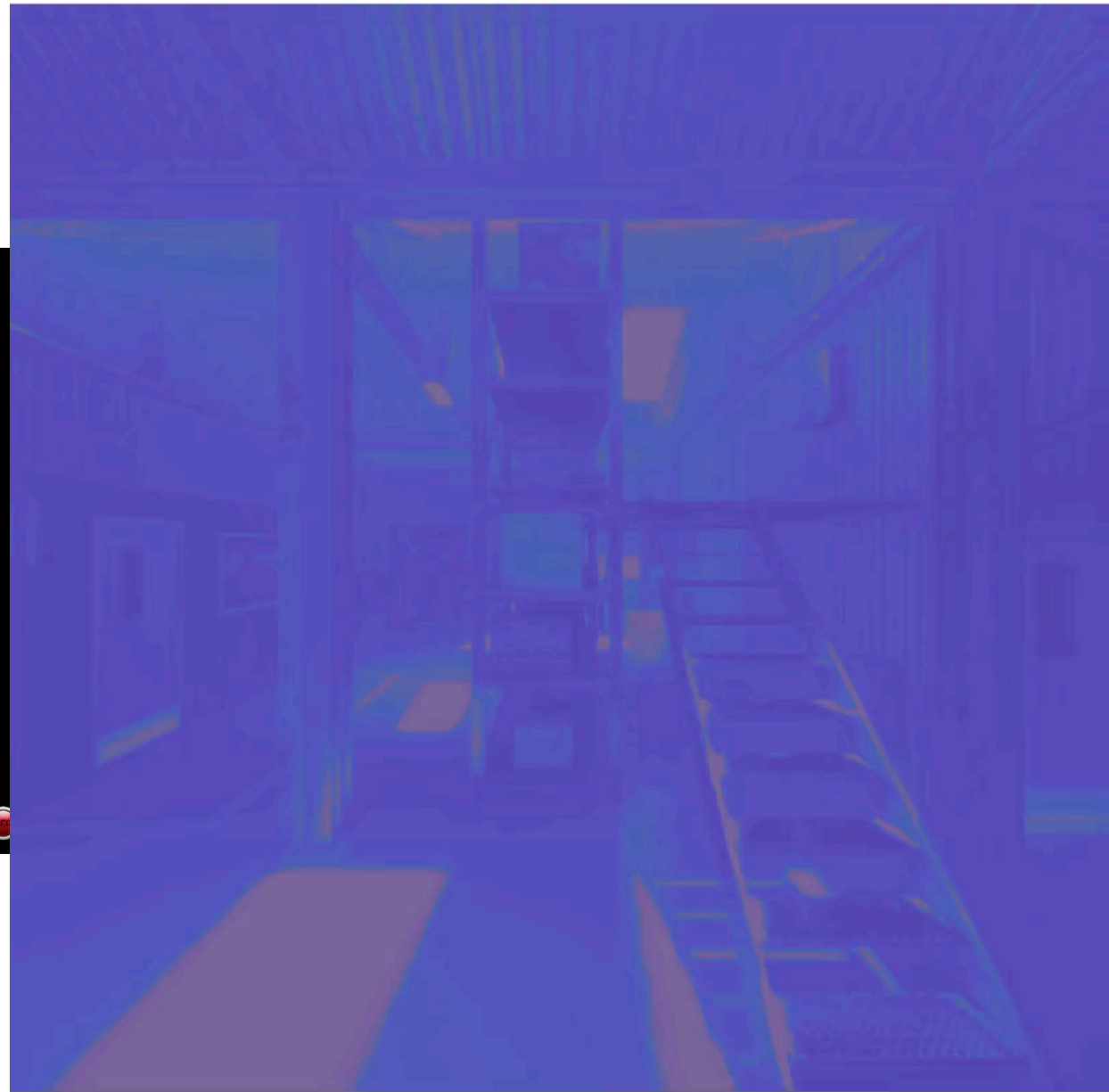
Assistant Professor
Computer Science
2018

Meet Otto Whateley Layton

Born September 2, 2018



My research: Brain-inspired algorithms for autonomous navigation



Overview of the course

**Week 1: Introduction to data
and data visualization**

**Weeks 2-4: Designing a GUI for
interactive data visualization**

**Weeks 5-9: Finding relationships in
high dimensional data**

Weeks 10-14: Machine Learning

Syllabus

- Let's take a look at the course syllabus.

What is data?

- A collection of measurements, often obtained through scientific work, and stored as bits on a computer.
- Real datasets are messy!
- The word is plural.
- Pieces of *information* — implies that symbols are involved that encode meaning. Not “just numbers” in a vacuum. Intended to be interpreted.

Data requires meta-data

Here are some data...

4.6	3.1	1.5	0.2	0
5	3.6	1.4	0.2	0
5.4	3.9	1.7	0.4	0
4.6	3.4	1.4	0.3	0
5	3.4	1.5	0.2	0
4.4	2.9	1.4	0.2	0
4.9	3.1	1.5	0.1	0
5.4	3.7	1.5	0.2	0
4.8	3.4	1.6	0.2	0
4.8	3	1.4	0.1	0

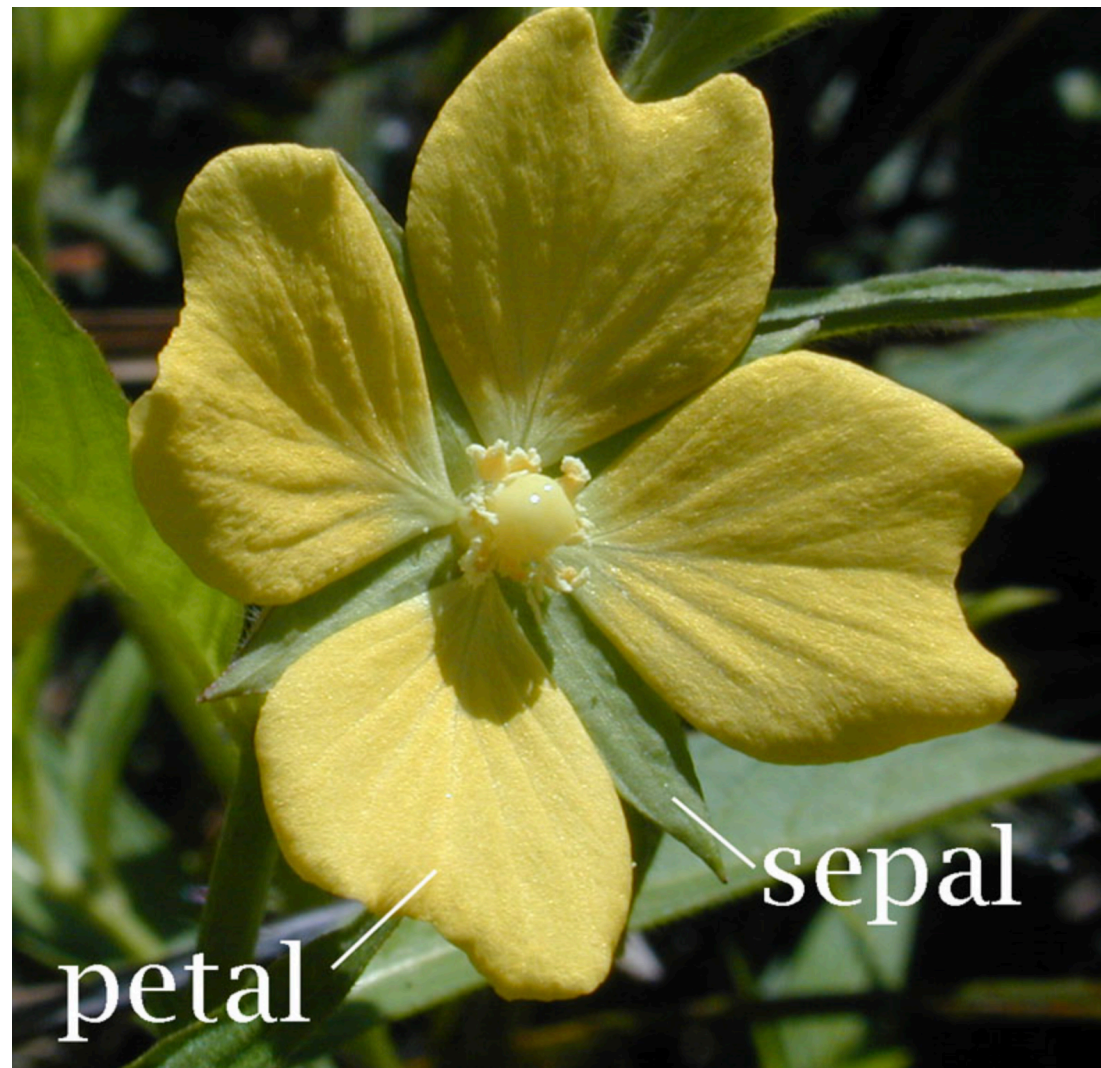
Data requires meta-data

- **Meta-data:** Data about data. Tells us what the numbers, words, and other types of data mean.
- Interpreting data is impossible without meta-data!

Sepallength	Sepalwidth	Petallength	Petalwidth	Class	
4.6	3.1	1.5	0.2	0	
5	3.6	1.4	0.2	0	
5.4	3.9	1.7	0.4	0	
4.6	3.4	1.4	0.3	0	
5	3.4	1.5	0.2	0	
4.4	2.9	1.4	0.2	0	
4.9	3.1	1.5	0.1	0	
5.4	3.7	1.5	0.2	0	
4.8	3.4	1.6	0.2	0	
4.8	3	1.4	0.1	0	

Iris flower data set

- Introduced by famous statistician Ronald Fisher (1936).
- 50 measurements about 3 types of flower petals and sepal.



The anatomy of data 1/2

“N”: Number of measurements

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
6	4.6	3.4	1.4	0.3	setosa
7	5.0	3.4	1.5	0.2	setosa
8	4.4	2.9	1.4	0.2	setosa
9	4.9	3.1	1.5	0.1	setosa
...					

Variable / feature

A label that organizes a set of numbers that have a single meaning. Part of an observation.

Multivariate data?

Missing data

2		3.2	1.3	0.2	setosa
3	4.6		1.5	0.2	setosa

Data point / feature vector

Collection of values that make up a complete observation/measurement. One “point” per labeled thing we could plot on X, Y, Z, etc. axes.

Dimension

Number of variables/features in the data. How many dimensions do these data have here?

Min/max/range of a feature

What are the range of petal_width??

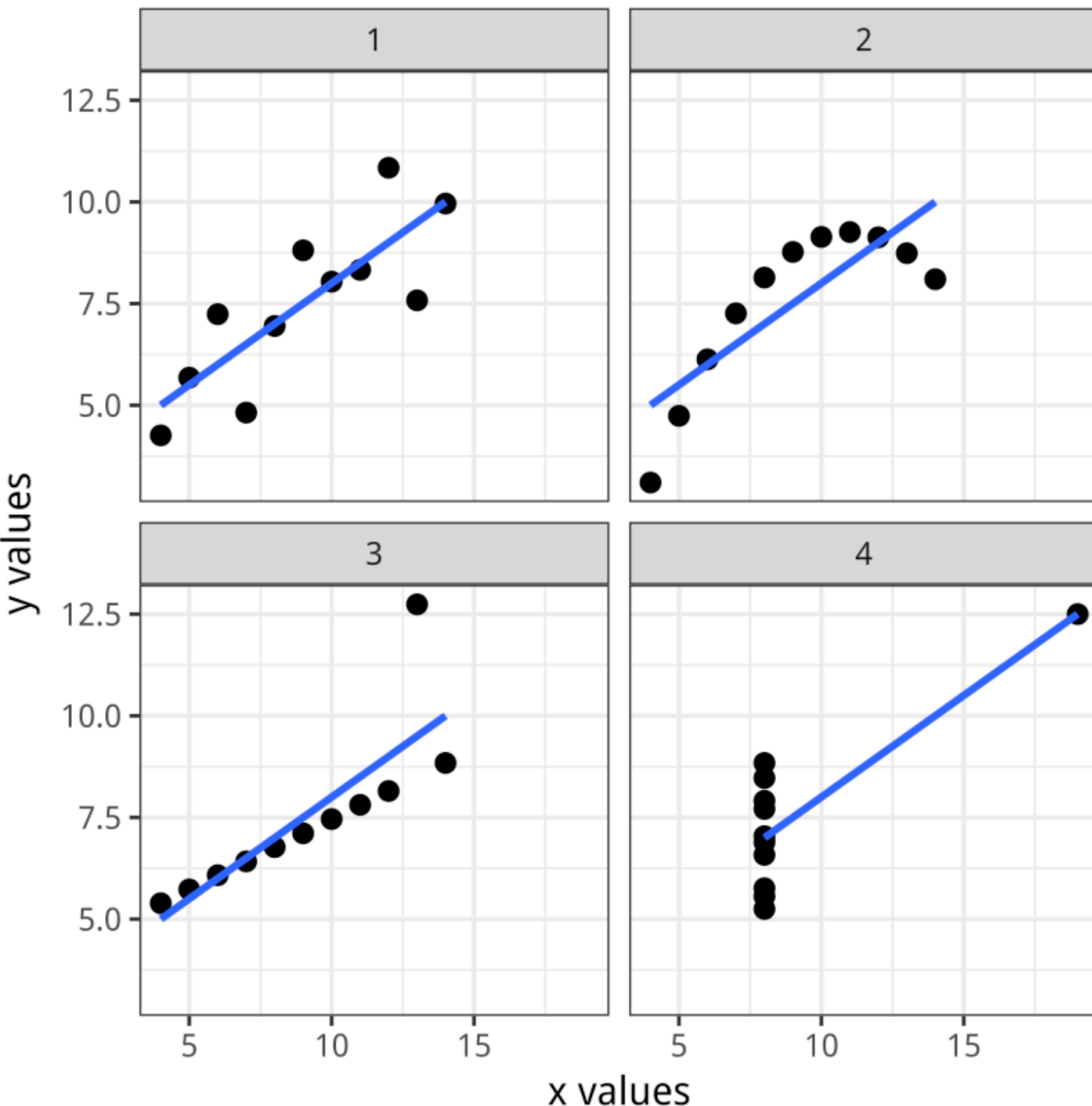
Range: max - min

If we can quickly compute statistics (e.g. range, mean, variance, correlation) to summarize thousands or more data points...

Why look at data?

Anscombe's quartet

Each of these four datasets have the same mean, correlation, and many other statistics!



X mean: 9

X variance: 11

Y mean: 7.5

Y variance: 4.125

Correlation: 0.816

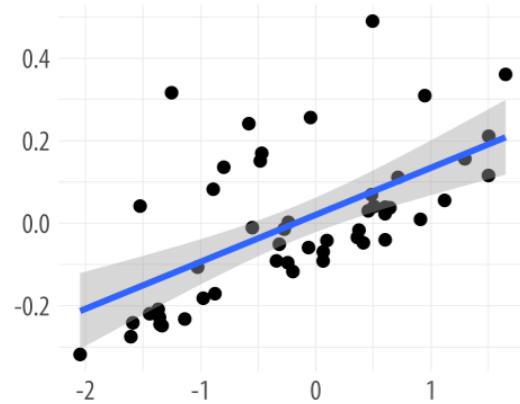
Scatterplot: Assign one feature (variable) to x-axis, assign another to y-axis.

Plot each point that has values for both features.

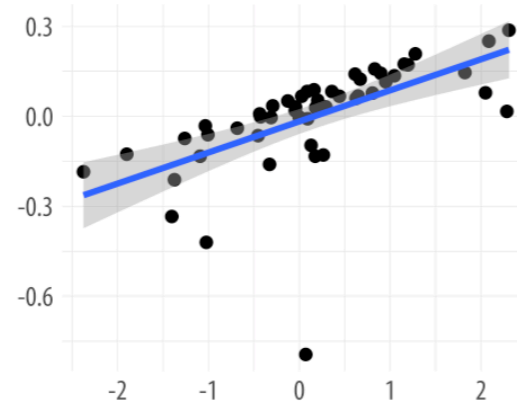
Summary statistics are not enough!

Correlation coefficient = +0.6 in all cases!!!!

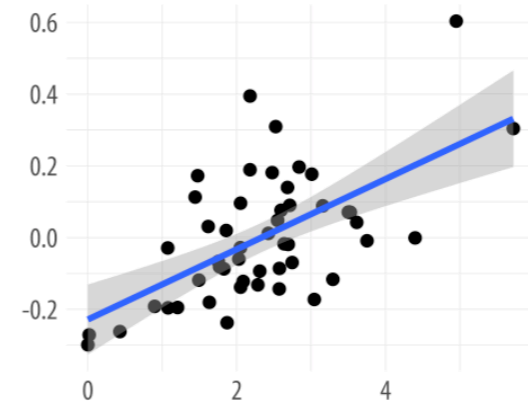
5. Normal x, + skewed residuals



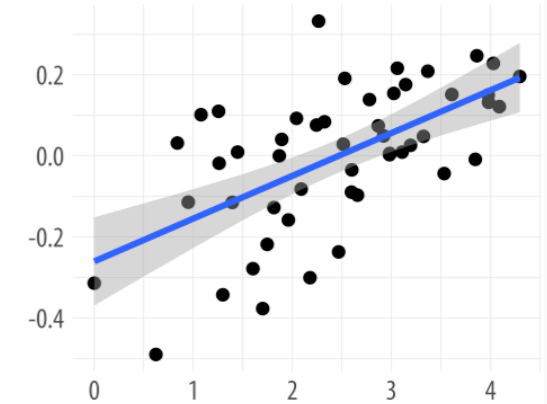
6. Normal x, - skewed residuals



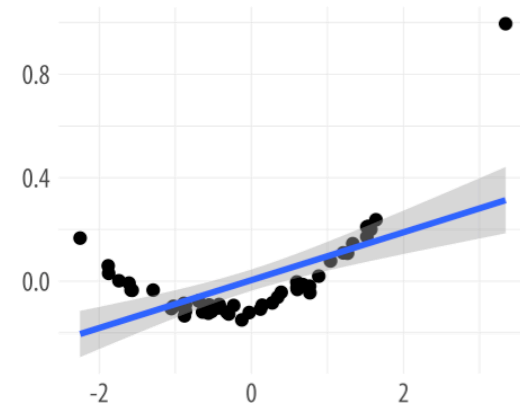
7. Increasing spread



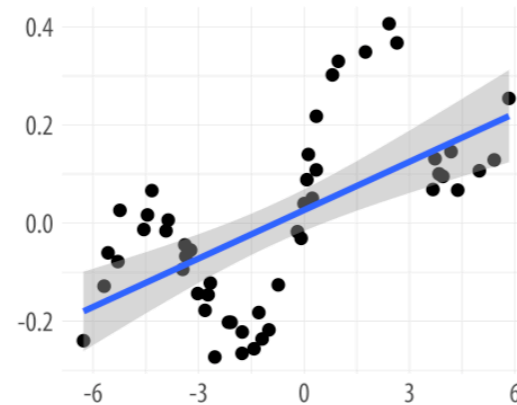
8. Decreasing spread



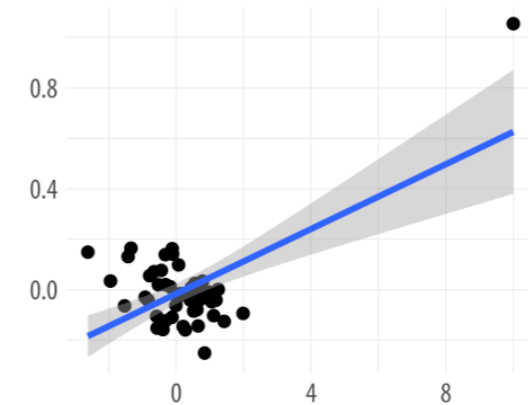
9. Quadratic trend



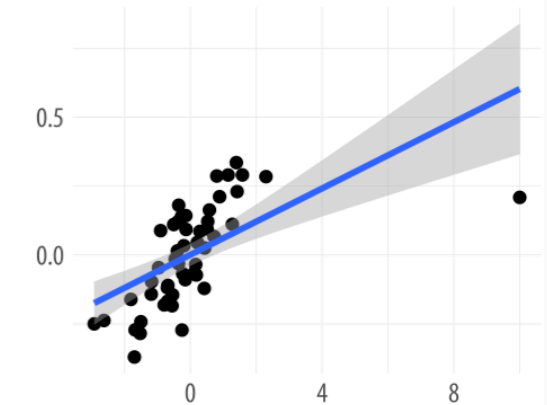
10. Sinusoid relationship



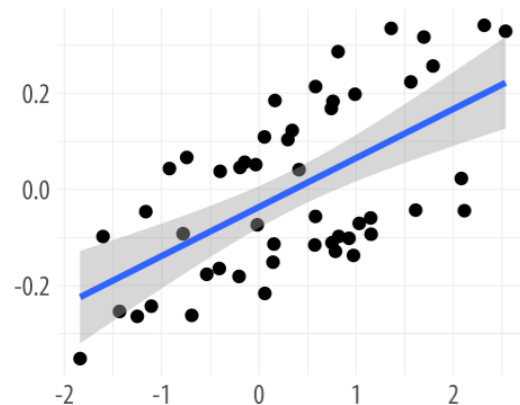
11. A single positive outlier



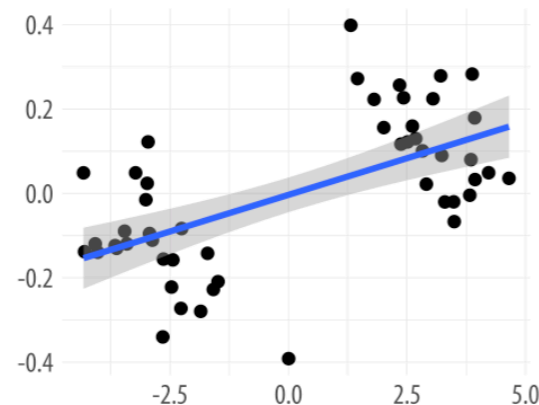
12. A single negative outlier



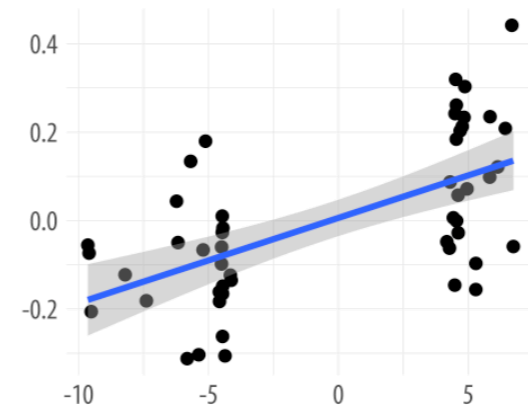
13. Bimodal residuals



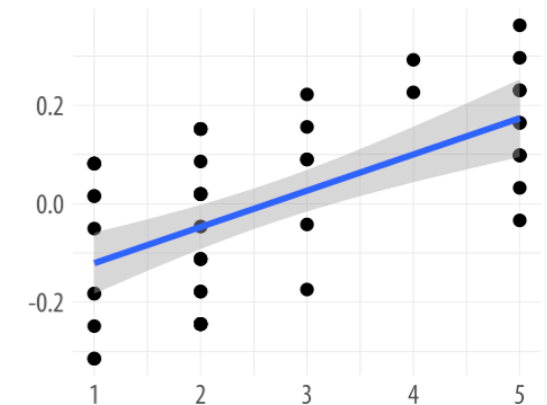
14. Two groups



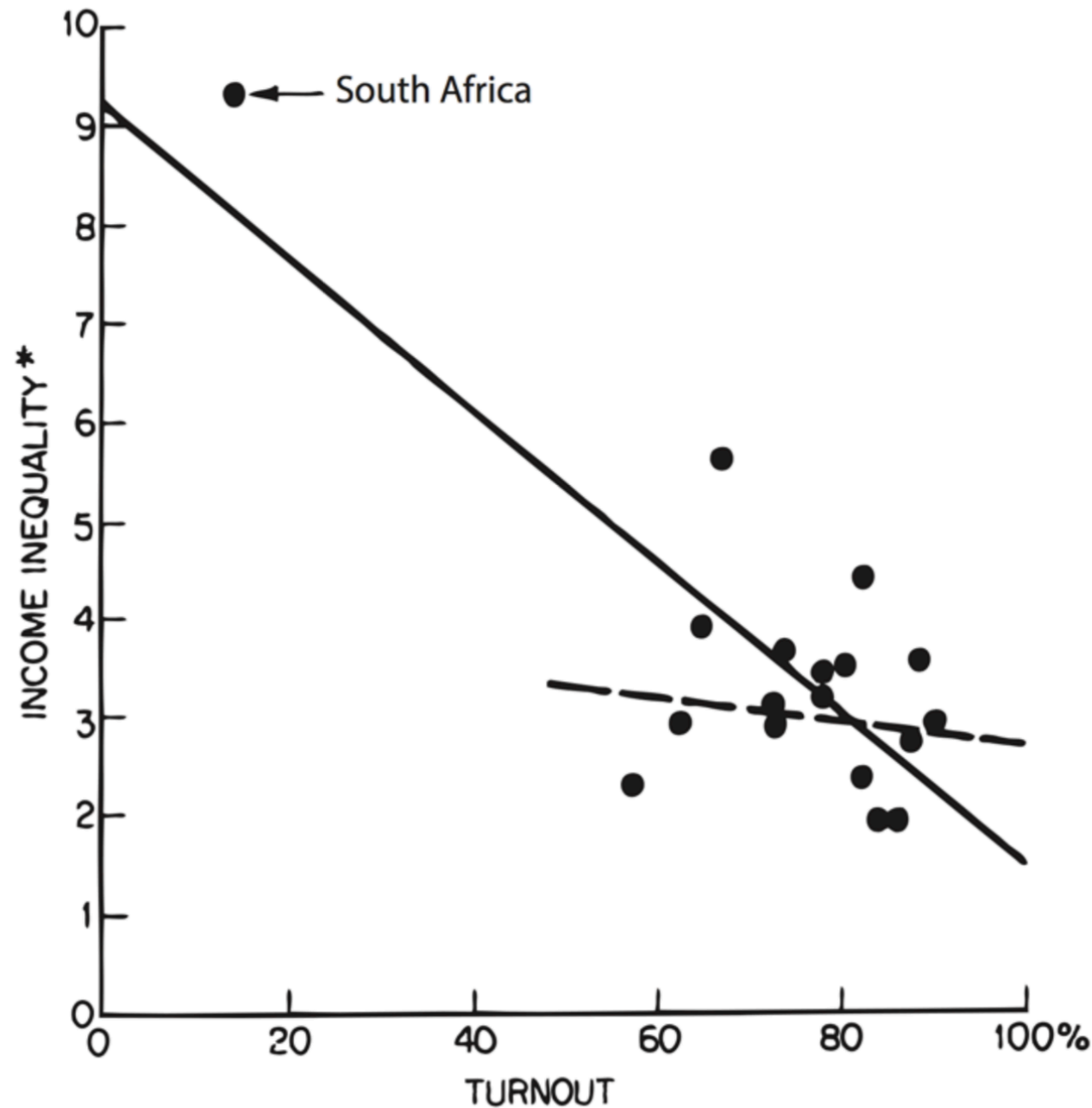
15. Sampling at the extremes



16. Categorical data



Voter turn-out data across 18 countries



Jackman: Numeric analysis shows significant association between voter turn-out and income inequality (solid line).

What's the problem?

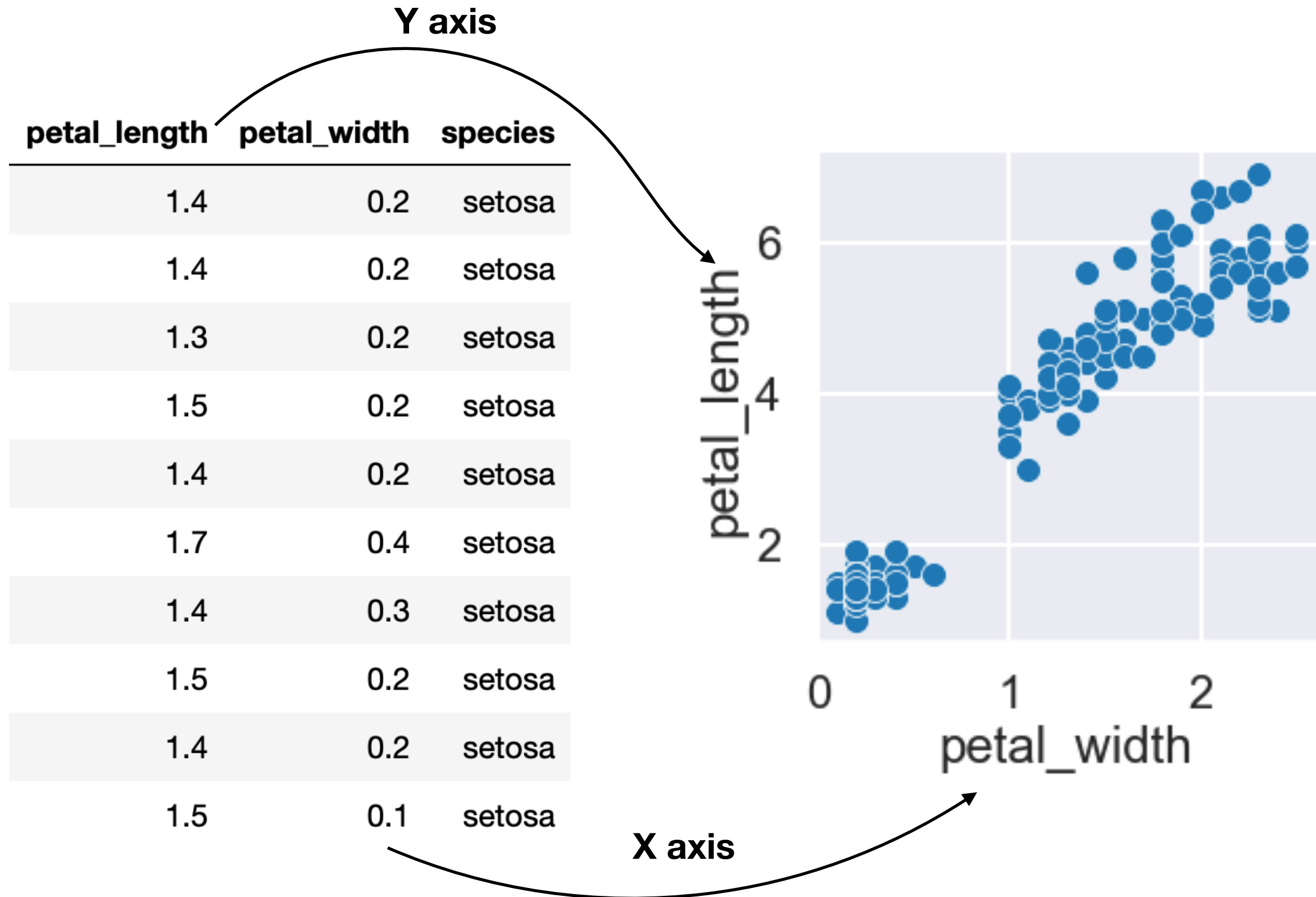
South Africa has high **leverage**.

Key. — Bivariate slope including South Africa (N = 18)
- - - Bivariate slope excluding South Africa (N = 17)

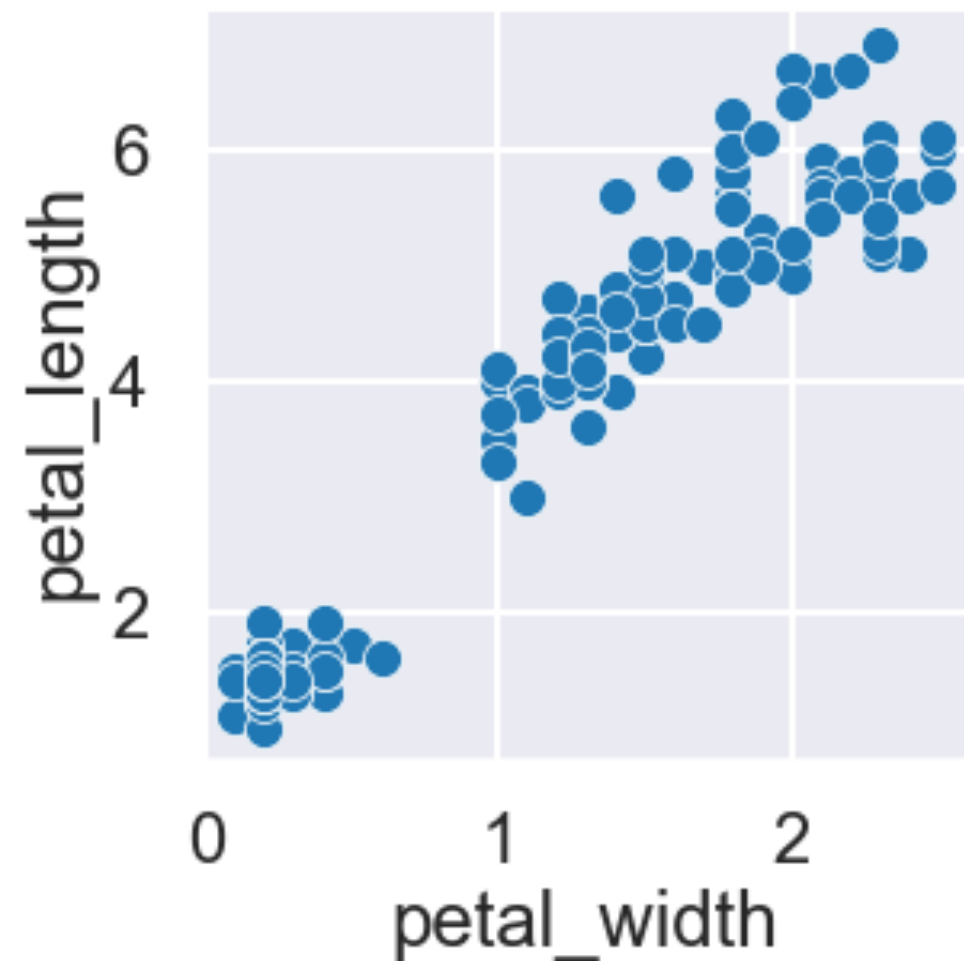
Data visualization

- Vision is the dominant means by which humans take in information about the world ($\geq 30\%$ of neurons are involved in processing visual information!).
- We are exceptional at identifying patterns — computers probably process information much faster, but are comparably utterly horrendous at pattern recognition.
- **Data visualization:** Take advantage of human pattern recognition to suggest meaningful patterns and hypotheses about entities that you are interested in understanding.
- Goal is NOT making pretty pictures, it is about extracting knowledge. Often the best data visualization minimizes “data-to-ink” ratio.

Example: Identifying hypotheses about flower petal dimensions

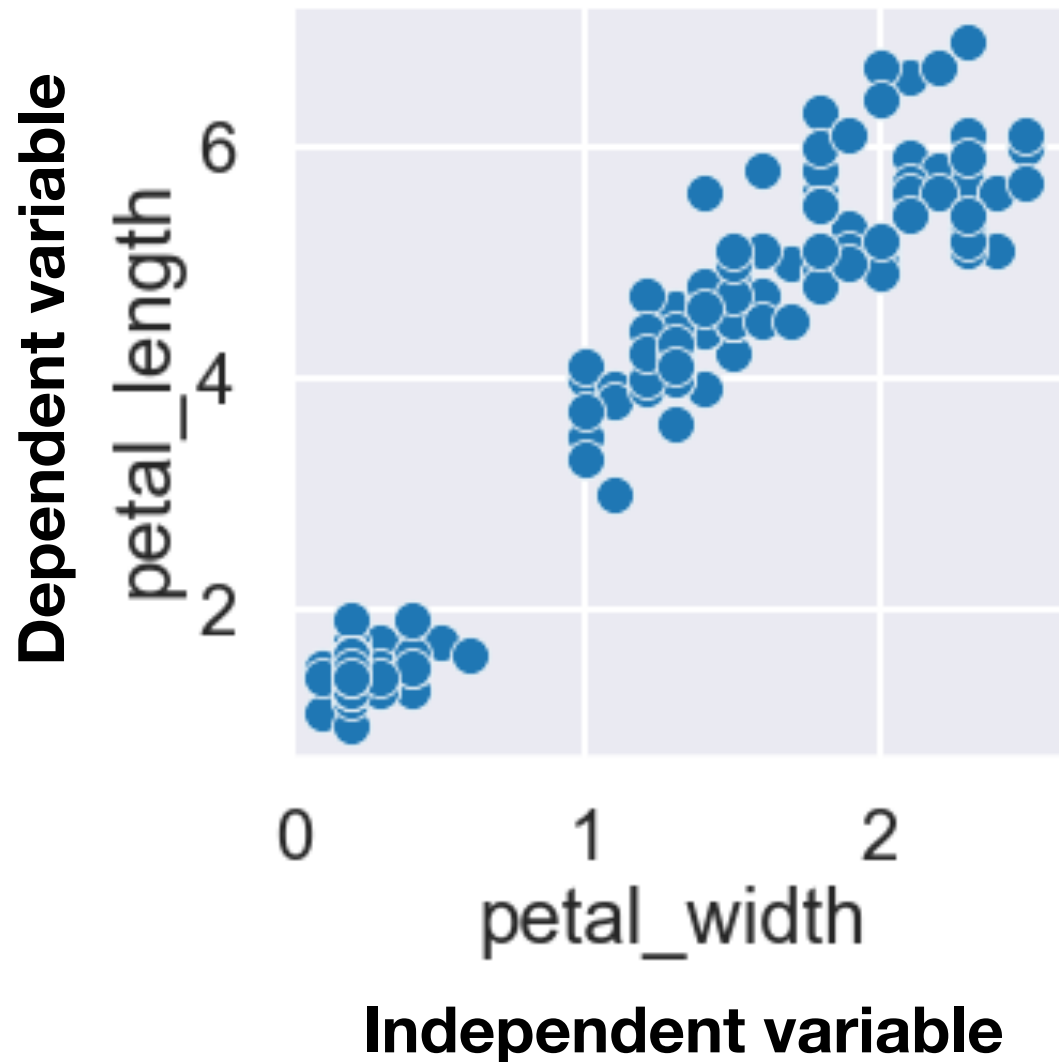


Data analysis

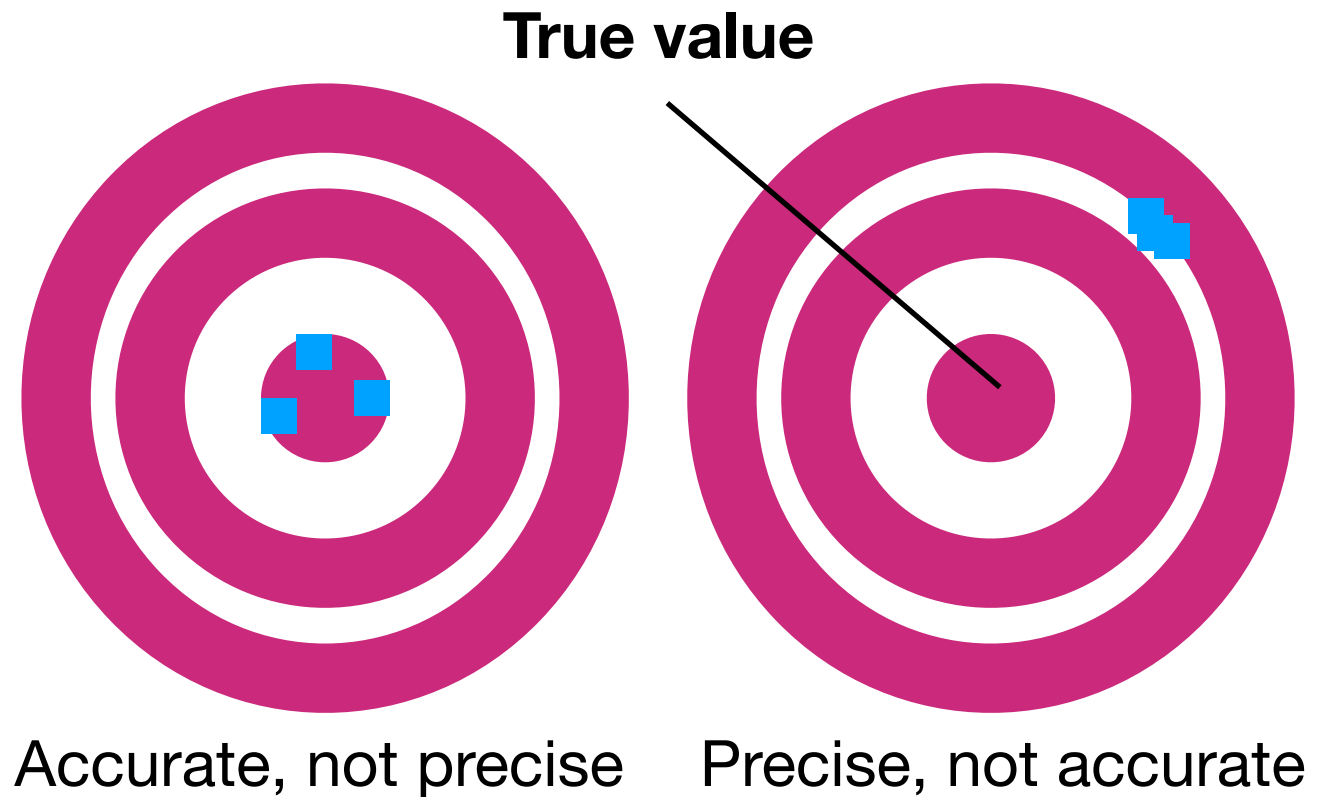


- The process of quantifying the answers to questions we ask about data (hypotheses, suspected relationships, etc.). Often involves statistics.
- Goal: Synthesize knowledge from the data.
- In practice, involves crunching numbers, which computers can help out with!

The anatomy of data 2/2



Which picture depicts an **accurate** measurement?
A **precise** measurement?



$$Y = mx + b$$

Independent var?

Dependent var?

Data are not “truth”: How we measure and store data contributes error in precision.

Example: $\pi \rightarrow 12$ sig. digits
 $\rightarrow 3$ sig. digits $\rightarrow 12$ sig digital

3.14000264991 12 sig figures
 3 digits of precision

Types of data

	Example Values	Example Application	Other names
Quantitative Continuous	5×10^{-3}	Temperature	
Quantitative Discrete	1,2,3,4	Product units sold	
Qualitative Unordered	Sudan, SUV, motorcycle	Type of car	Variable: Factor Specific values: Levels
Qualitative Ordered	Excellent, good, poor	Survey rankings	Ordered factors
String/text	Green eggs and ham	DNA sequences	
Date/time	Jan 1, 1970	Transaction times	

Id,Longitude,Latitude,Continent,Basin_id,Cell_locat,evapproxy,premin,premax,salmin,salmax,minairtemp,maxairtemp
string,string,string,string,enum,enum,numeric,numeric,numeric,numeric,numeric,numeric,numeric,numeric,numeric,n
36313,132.5,-10.5,Australia,771,-1,11939.9,2.4,350,3423,3448,287,301.5,299953,302817,0.09,6.39,1872.04,0.2,0.51
36323,142.5,-10.5,Australia,0,-1,10951.2,8,350,3384,3487,290,301.2,298990,302280,1.6,0,396.521,0.49,0.8,0,2,4.5
36671,130.5,-11.5,Australia,771,1,11933.5,3,350,3472,3483,289.5,301.5,299153,303625,0,0,52.0,0.56,1.25,0,1,8
36672,131.5,-11.5,Australia,771,1,11937.2,5,350,3472,3483,289.5,301.5,299153,303625,0,0,52.0,0.56,1.25,0,1,8
36673,132.5,-11.5,Australia,771,1,11585.1,3,350,3418,3459,288,301,299936,302890,0,5,754.002,0.66,1.53,300,1,6
36674,133.5,-11.5,Australia,771,1,11640.5,2.25,375,3392,3472,287,302,299746,302767,0,9,911.529,0.56,0.75,0,1,6
36675,134.5,-11.5,Australia,0,-1,11933.5,3,350,3472,3483,289.5,301.5,299153,303625,0,0,52.0,0.56,1.25,0,1,8
36676,135.5,-11.5,Australia,0,-1,12830.5,3.71,435.71,3354,3483,288.9,302.4,299530,302733,0,0,733.104,0.61,0.82,
36677,136.5,-11.5,Australia,771,1,12696.5,3,450,3349,3480,289,302,299402,302775,0,9,731.905,0.64,0.97,0,1,4.5
36682,141.5,-11.5,Australia,0,-1,12730.1,6.33,350,3369,3467,289.9,301.4,299473,302847,2,0,2851.74,0.32,0.68,0,2
36683,142.5,-11.5,Australia,783,1,11259.8,8,350,3378,3482,290,301,299099,302422,1,14,661.522,0.41,1.11,0,2,4.5
36684,143.5,-11.5,Australia,0,-1,11633.6,8,350,3400,3505,289.9,301.4,298653,301970,2,0,2646.32,0.31,0.5,0,2,4.5
37031,130.5,-12.5,Australia,0,-1,13357.9,2.77,411.54,3430,3485,285.7,302,300153,303427,0.06,0,677.418,0.85,1.54
37032,131.5,-12.5,Australia,771,1,13281.3,3,350,3432,3483,285,302,300103,303326,0,10,1853.16,1.48,2.5,300,1,8
37033,132.5,-12.5,Australia,771,2,12742.1,0,350,3415,3471,285,302,299951,303016,1,12,2345.57,-9999,2.43,300,1,6
37035,134.5,-12.5,Australia,771,1,12336.9,3,450,3366,3480,288,303,299518,302771,0,13,1358.36,1.27,1.95,300,1,6
37036,135.5,-12.5,Australia,771,1,12436.4,3,450,3341,3478,290,302,299227,302826,0,6,800.45,1.04,1.68,150,1,6
37037,136.5,-12.5,Australia,771,1,13096.3,4.43,421.43,3317,3473,288,303,298886,302907,0,8,28.967,0.5,0.91,0,2,4
37042,141.5,-12.5,Australia,0,-1,13226.9,5.22,350,3346,3449,289.5,301.8,299713,303057,2.42,0,2734.71,0.63,0.86,
37043,142.5,-12.5,Australia,783,1,12100.6,8,350,3367,3475,290,302,299193,302556,4,13,1241.48,-9999,1.91,600,2,4
37044,143.5,-12.5,Australia,0,-1,12594,7,330,3386,3500,289.5,301.8,298600,301867,2.35,0,3427.91,0.3,0.47,0,2,4.
37386,125.5,-13.5,Australia,772,-1,11875.9,0,335.71,3465,3484,284.1,302.8,300083,302460,0,2.9,2496.7,0.29,0.71,
37387,126.5,-13.5,Australia,0,-1,12040.1,0,350,3473,3491,284.3,303.3,300147,302677,0,0,1625.46,0.42,1.25,0,1,10
37388,127.5,-13.5,Australia,772,-1,11986.5,0,350,3479,3498,284,303.3,300147,302860,0,2.6,4411.86,0.27,0.85,138,
37390,129.5,-13.5,Australia,0,-1,13803.7,1.33,438.89,3449,3506,282.6,303.1,300003,303327,0,0,1996.64,0.28,1.17,
37391,130.5,-13.5,Australia,771,1,12587.3,3,450,3444,3497,283,303,300051,303361,0,9,2988.15,0.69,2.21,0,1,10
37396,135.5,-13.5,Australia,771,1,13342.3,8,350,3308,3471,289,303,298652,302973,0,8,2015,-9999,-9999,150,2,4.5
37397,136.5,-13.5,Australia,0,-1,13467.4,5.22,361.11,3292,3466,288.4,302.9,298593,302970,0,0,492.279,0.51,0.66,
37402,141.5,-13.5,Australia,0,-1,14095,3,350,3340,3453,289,302,299713,303057,2,11,2948.66,0.93,1.3,0,2,4.5
37404,143.5,-13.5,Australia,783,1,13287.5,5.22,283.33,3368,3490,289,302,298846,302295,2,13,2927.28,0.39,0.59,0,
37746,125.5,-14.5,Australia,772,1,12835.6,0,350,3469,3488,285,302,300109,302571,0,3,1675.19,0.51,0.99,0,1,8
37747,126.5,-14.5,Australia,772,1,11237.0,0,350,3475,3492,284,304,299902,302728,0,6,2992.77,1.2,1.12,300,1,10

Data storage (in this class)

CSV (comma separated file)

HW0

- No quiz this week, but homework assignment (due tomorrow) is find a dataset online *with at least 5 numeric features*.
- Briefly identify (in a text file): meta-data, features, dimensions, N, range, min, max, example missing data (if any), example feature vector, precision of data.
- If you find a sample plot of the dataset: what are the independent and dependent variables?
- **DO NOT UPLOAD THE DATASET TO FILER!**
 - You can include screenshots and/or written summaries of the highlights.