Comparing two numeric variables: Correlation

Research Methods for Human Inquiry

Andrew Perfors

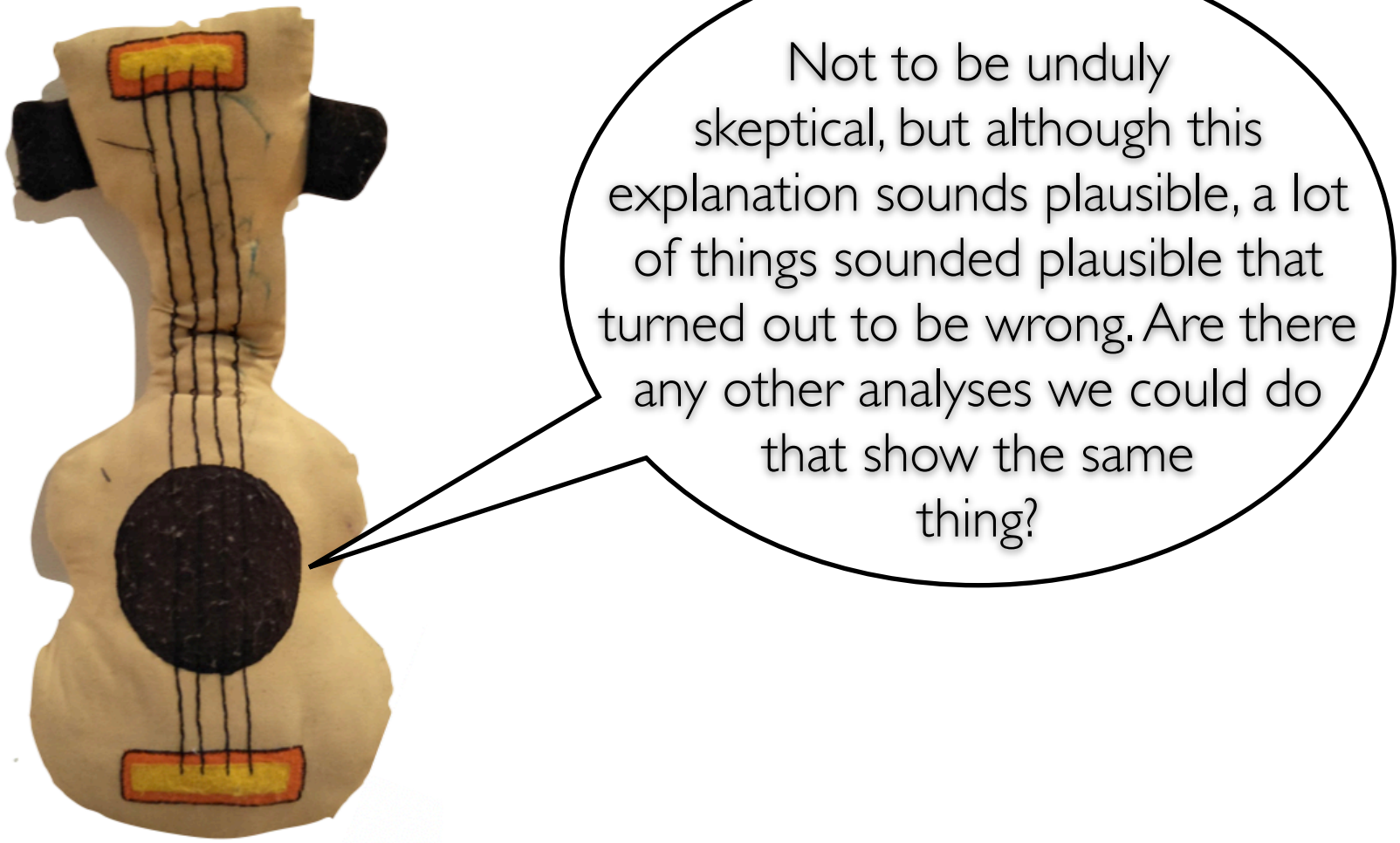
Let’s go to our story…

Everyone from Otherland and Bunnyland has started talking to each other and working together

Let's go to our story…

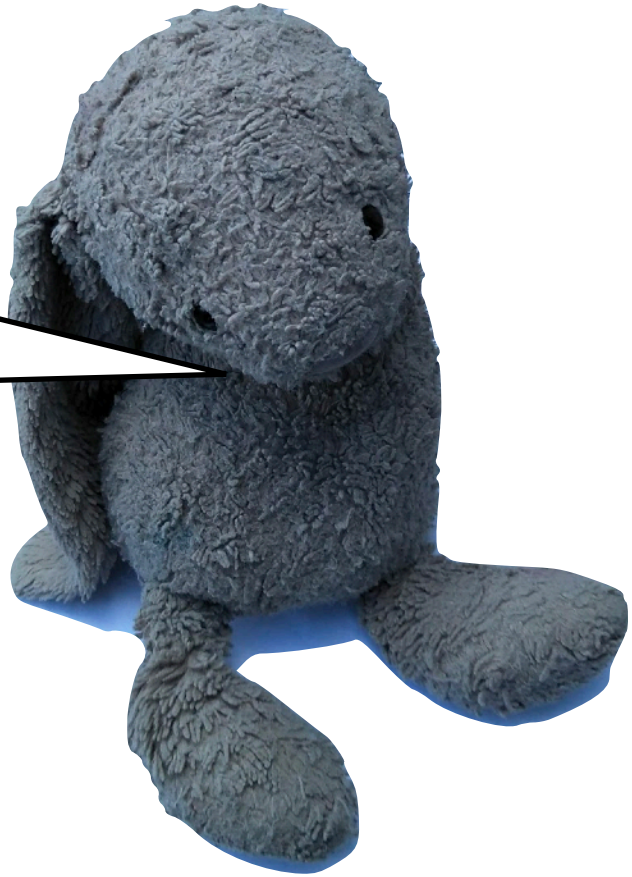
… and figured out that part of the problem is that everyone is using land badly - possibly because there's an imbalance of animals in each place

Let's go to our story…

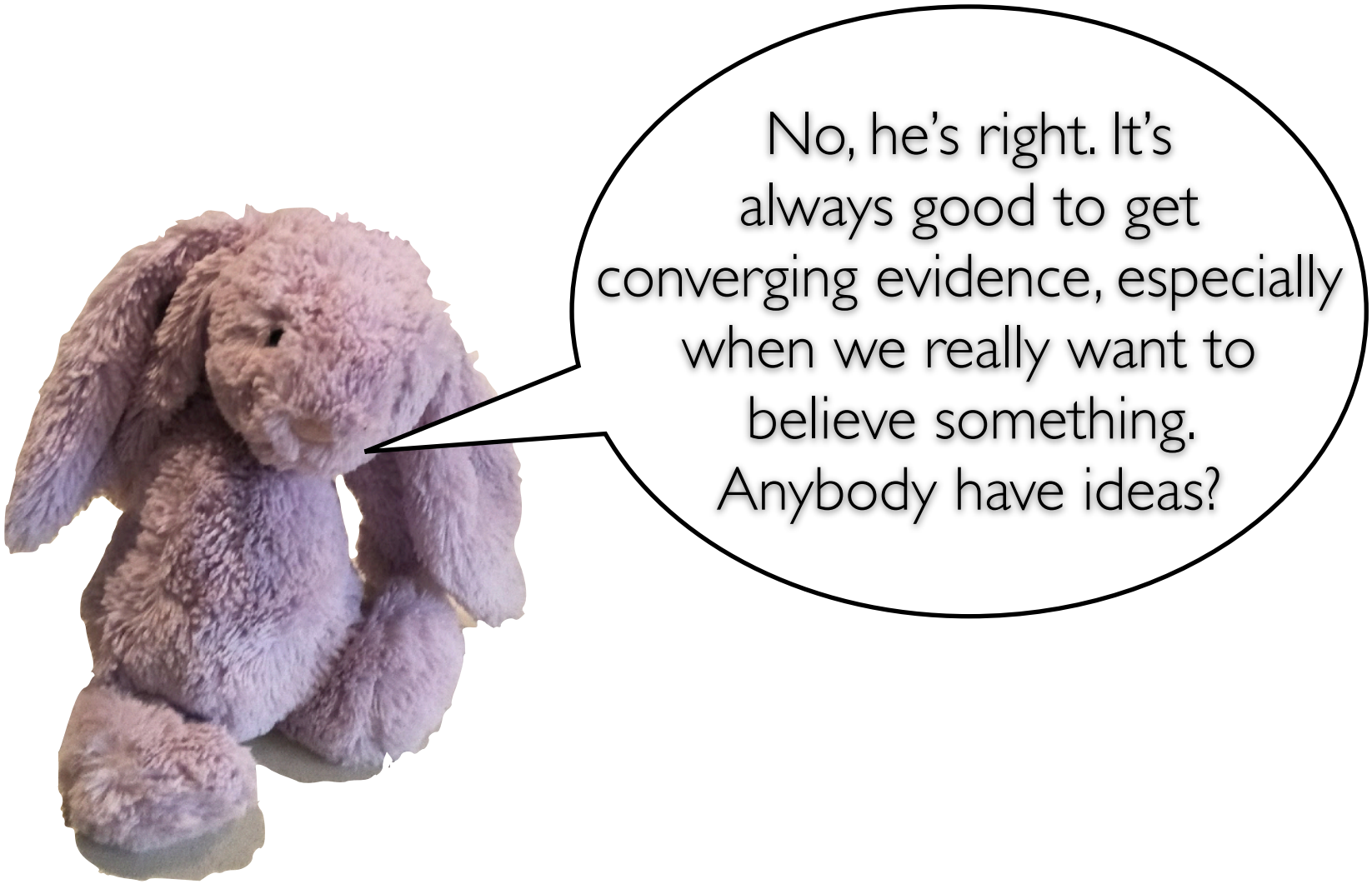


Let's go to our story…

That's kind of mean to say! We worked hard on that, especially Flopsy and Shadow.



Let's go to our story…



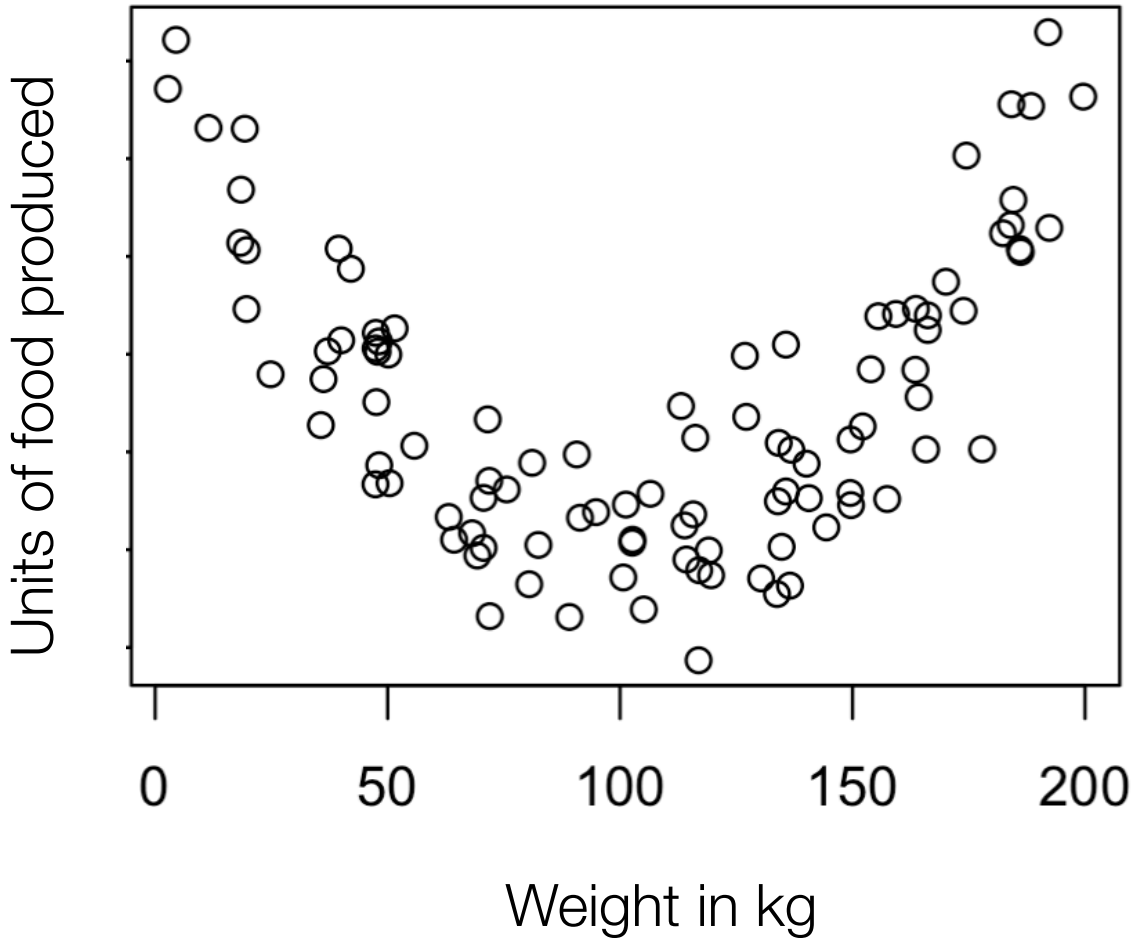
Let's go to our story…

Hmm…
if different species have different skills regarding farming, maybe species size is related to the amount of food they can harvest?

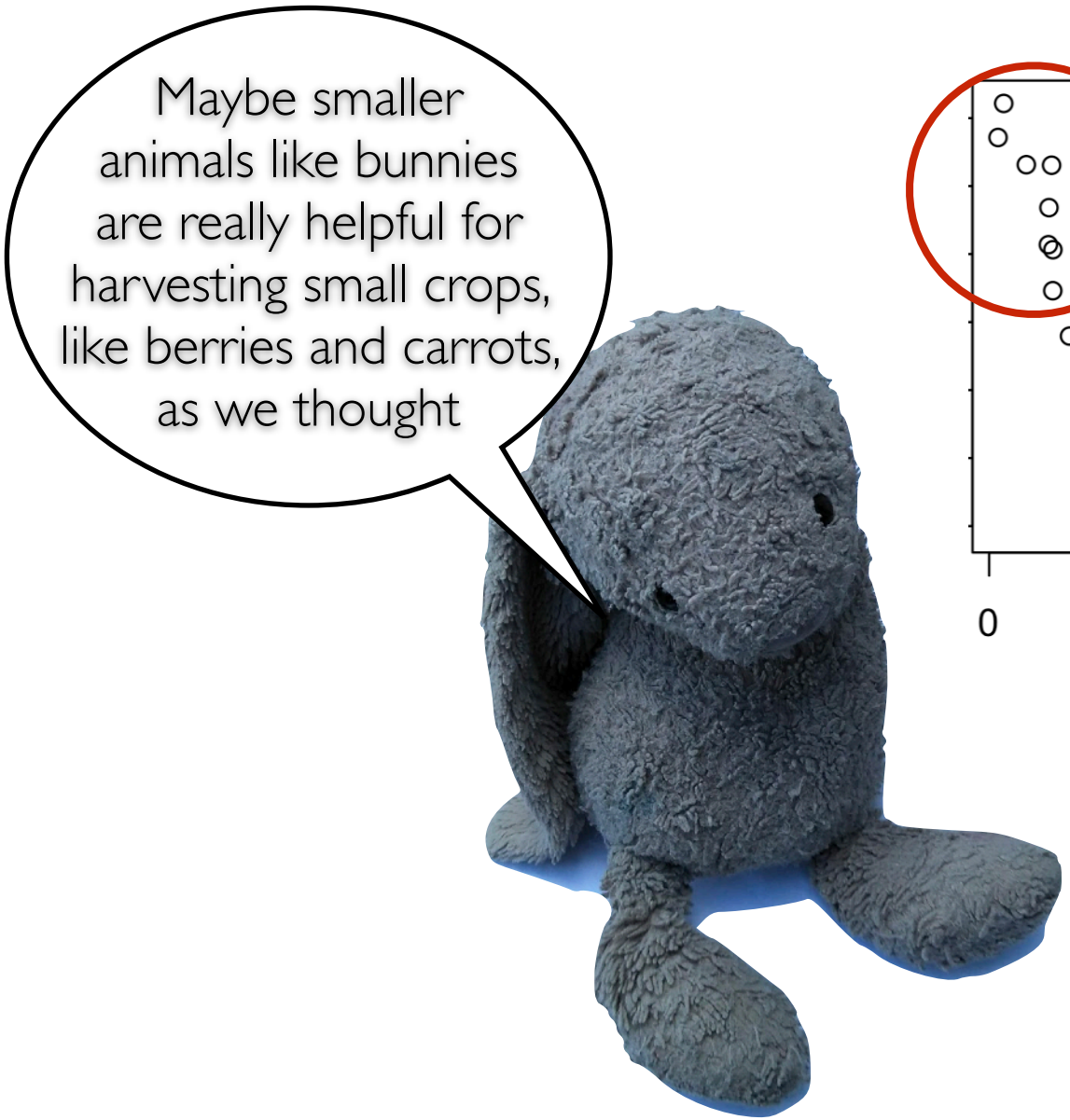
Let's go to our story…

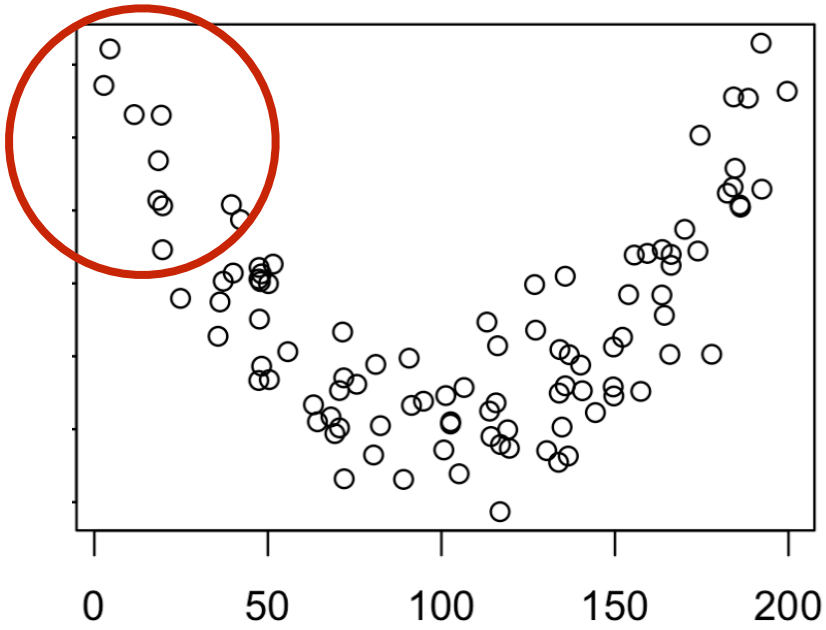
Hmmm….

What does *this* mean?



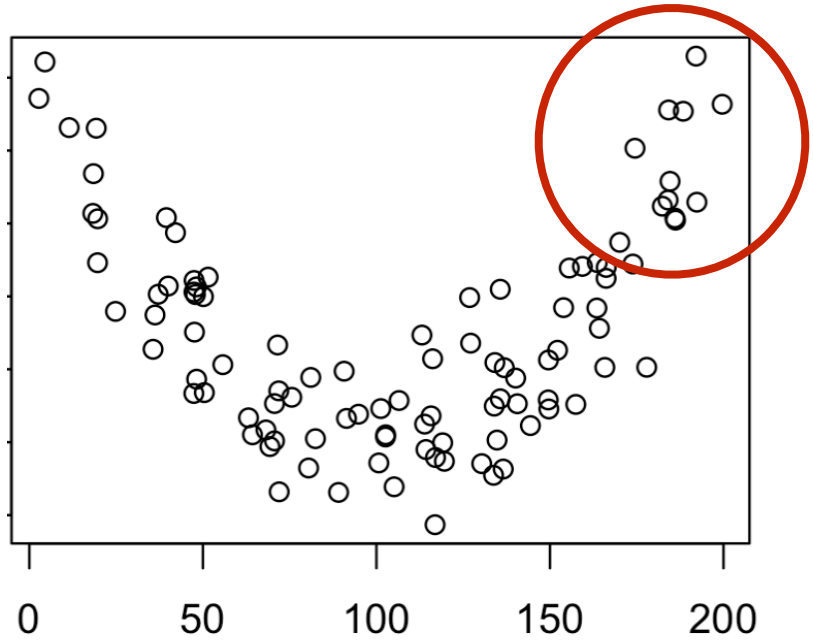
Hmmm….


Maybe smaller animals like bunnies are really helpful for harvesting small crops, like berries and carrots, as we thought


0
50
100
150
200

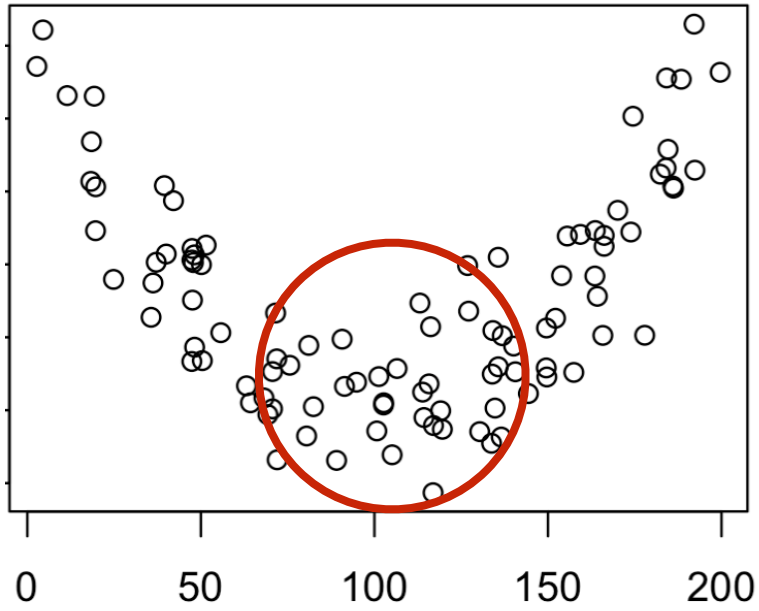
Hmmm….


And larger animals like bears are good for clearing land and pulling plows and ranching cows, like we thought!

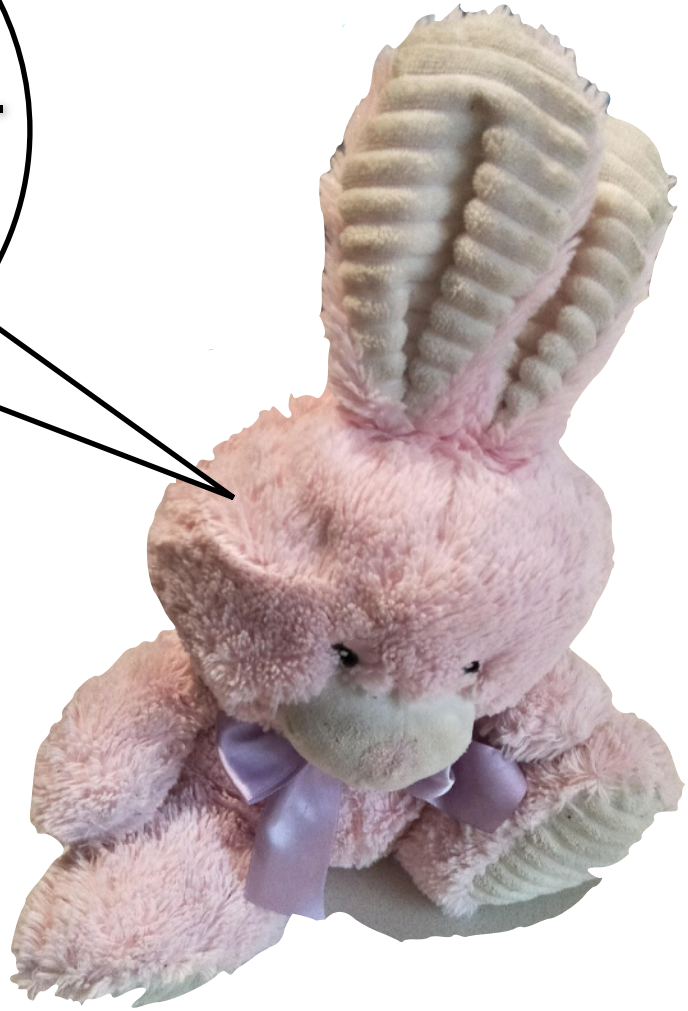

0
50
100
150
200

Hmmm….

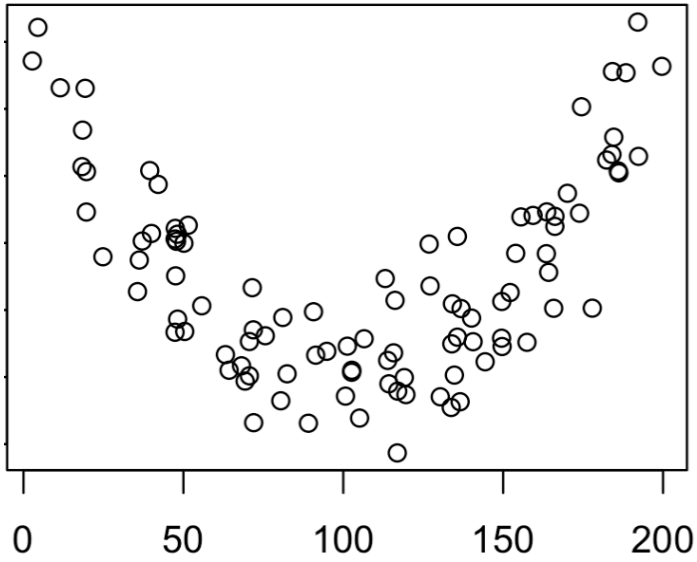
But what about mid-sized animals, like dogs or birds? Are we useless? I'm really sad now.



No way! This only captures one aspect of being useful - food production. We also need leaders, managers, scientists, artists, so many things!



In hindsight, this is misleading in another way too. We think that *diversity* of animals is what we've been lacking. We think we need people of all kinds in order to grow enough food. This doesn't capture that.



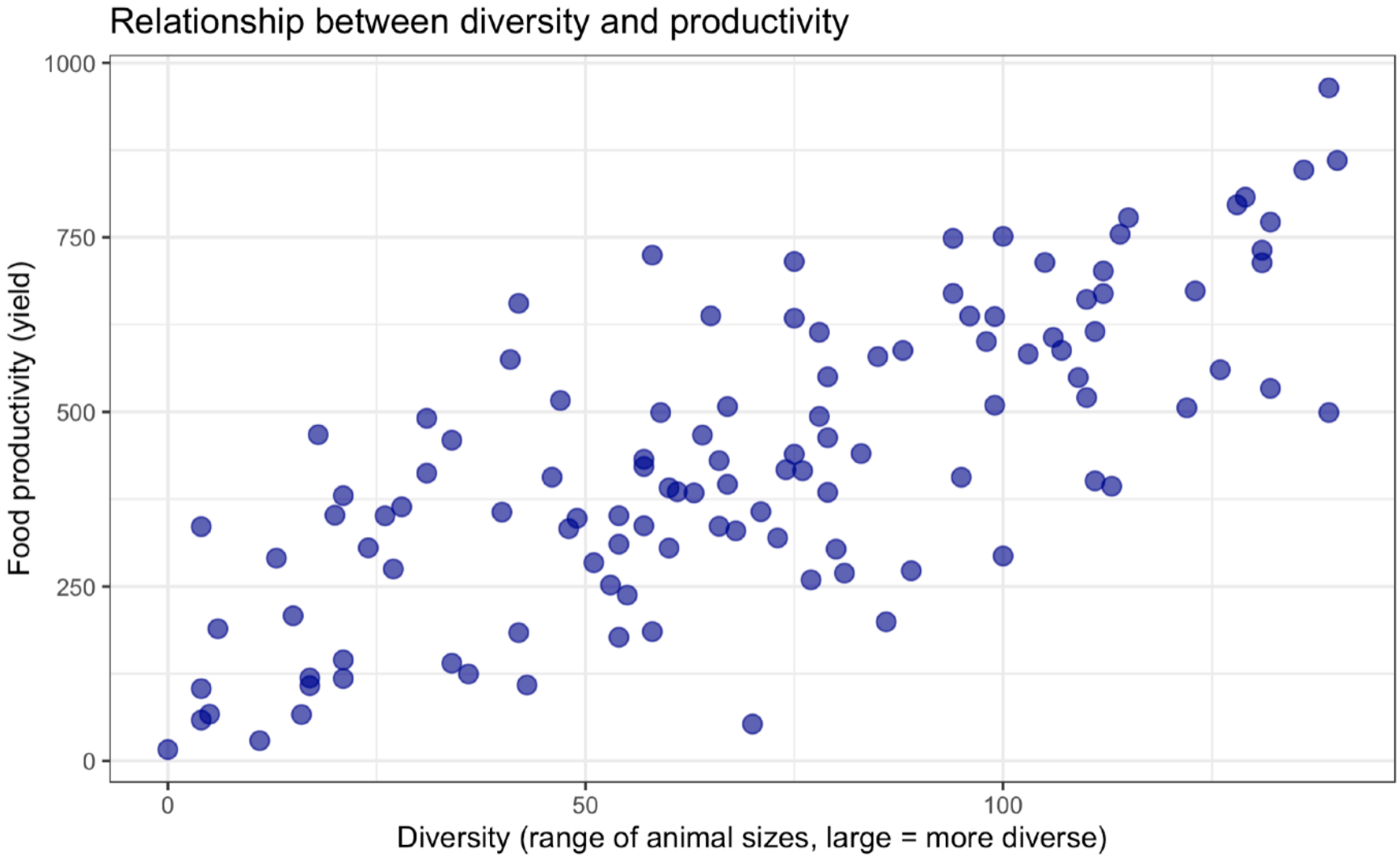
Let's test that directly. We don't have a lot of diversity, but we *do* have some. If that's really our problem, it predicts that the areas of land with the most diverse species farming it should be most productive.

The data

- One tibble, `d`, with four variables
  - `land`... code uniquely identifying each plot of land
  - `area`... size of each plot of land
  - `diversity`... measure of diversity of species farming each land; it's the maximum sized person minus the minimum sized person (so larger is more diverse)
  - `yield`... units of each food this plot of land yielded

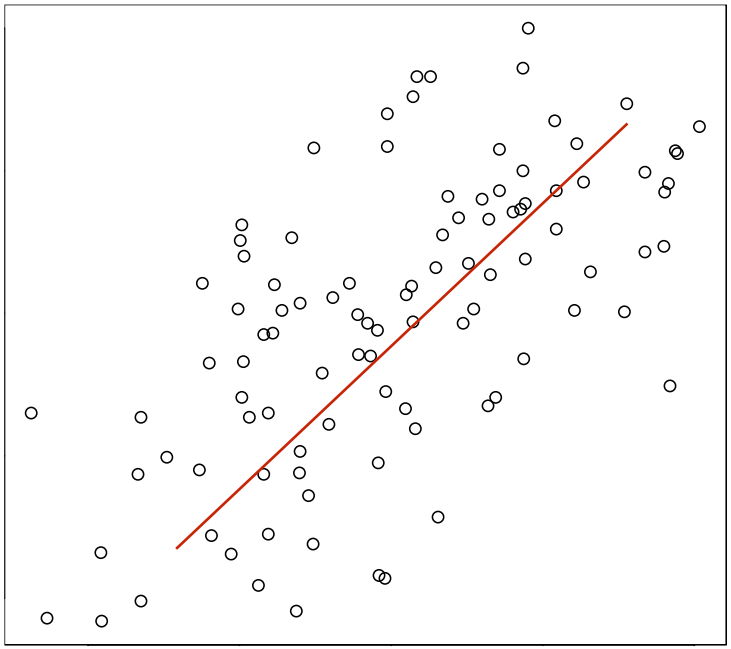
```
> head(d)
      land    area diversity yield
1 iUxbR6F2 578.175       131 713.6
2 1iDku4Bd 491.714       128 796.7
3 4wTCpKQy 285.915        68 329.7
4 cj47Q9GZ 443.127        95 406.4
5 hnnu0h26 189.164        54 177.2
6 zvbOpHaV 368.642        51 284.0
```

The data


Relationship between diversity and productivity
Food productivity (yield)
1000
750
500
250
0
0
50
100
Diversity (range of animal sizes, large = more diverse)

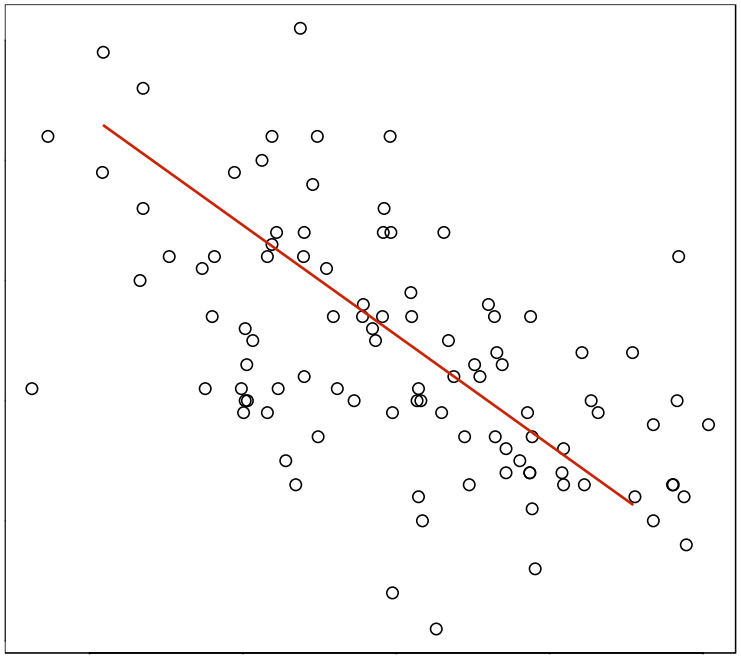
How can we characterise this?

Positive relationship: when one variable goes up, the other one goes up too



positive = going up a hill!

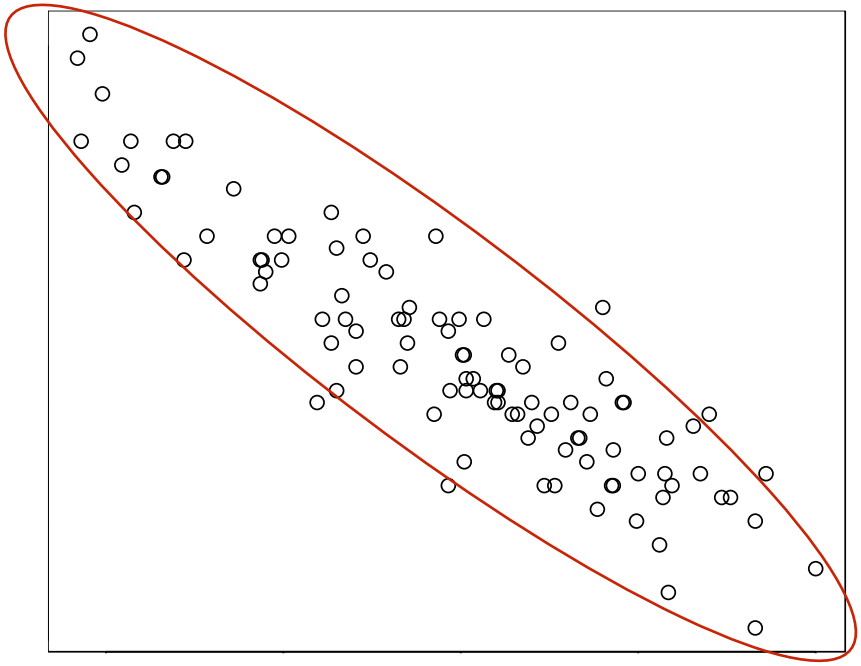
Negative relationship: when one variable goes up, the other one goes down



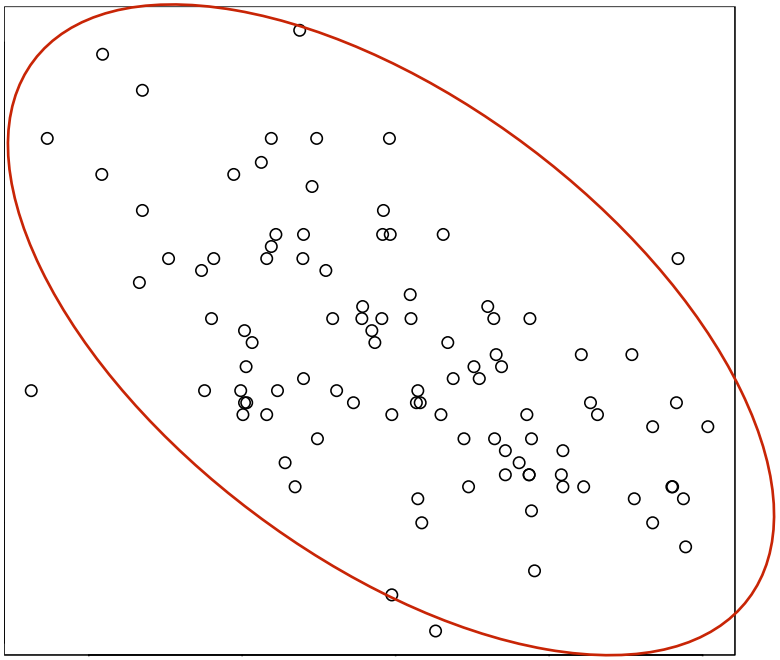
negative = going down a hill!

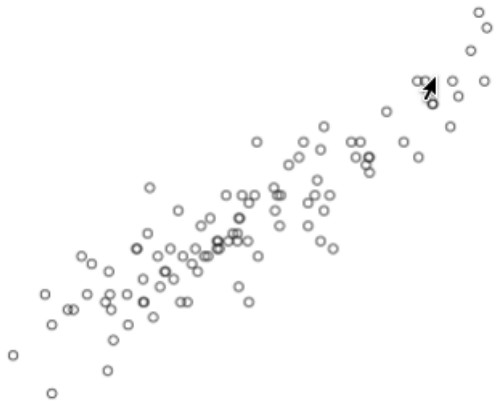
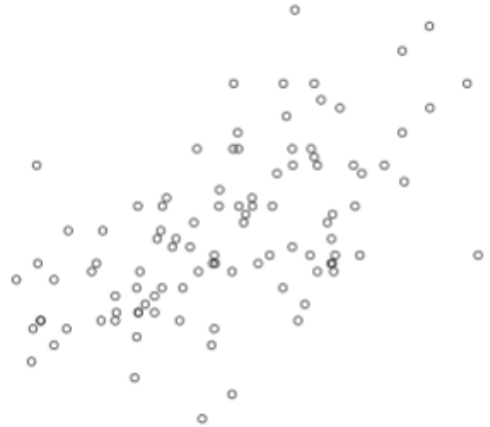
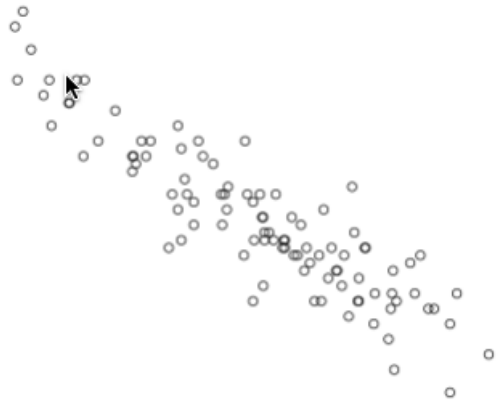
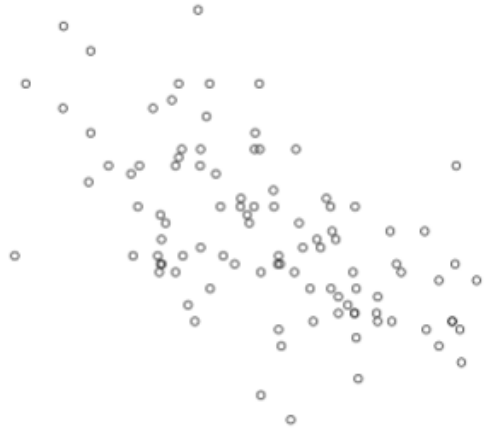
How can we characterise this?

Strong relationship: knowing the value of one variable tells you a lot about the value of the other one

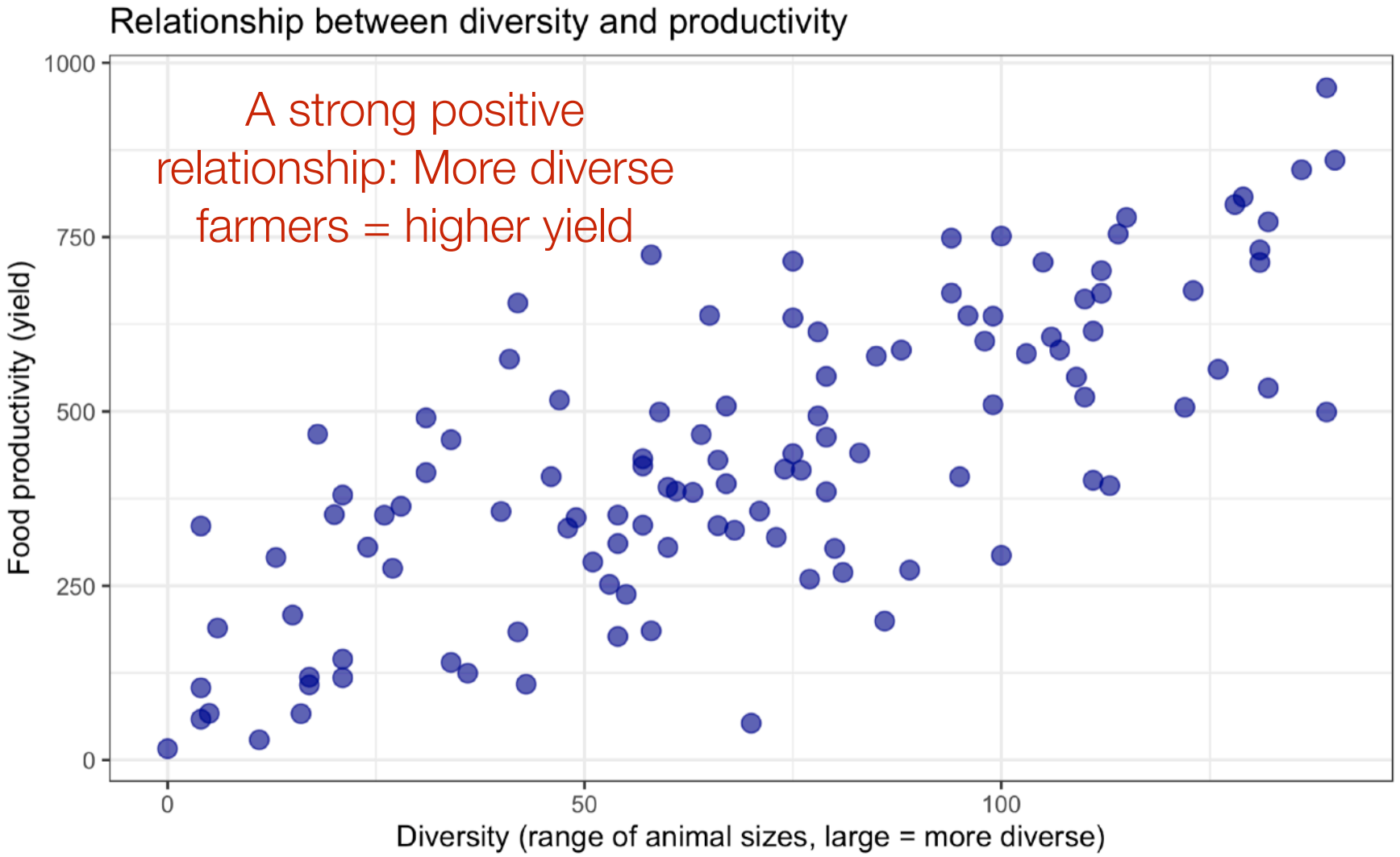


Weak relationship: knowing the value of one variable tells you a little about the value of the other one



	Strong	Weak
Positive		
Negative		

How can we characterise this?



Quantify it with a **Pearson correlation**

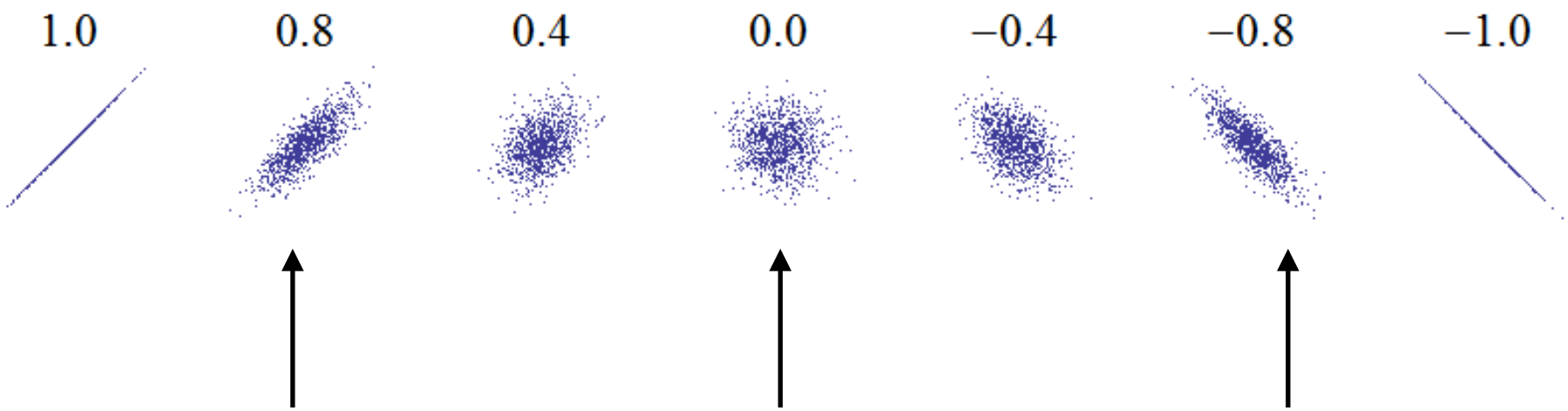
- A Pearson correlation is denoted r
  - Measures the strength and direction of a relationship
  - It assumes the relationship is linear

- The formula for calculating is this…

$$r_{XY} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{(N-1)s_X s_Y}$$

  - I won't explain it here because it's very related to what I'll explain regarding linear regression in the next videos

A visual explanation



Strong positive relationship

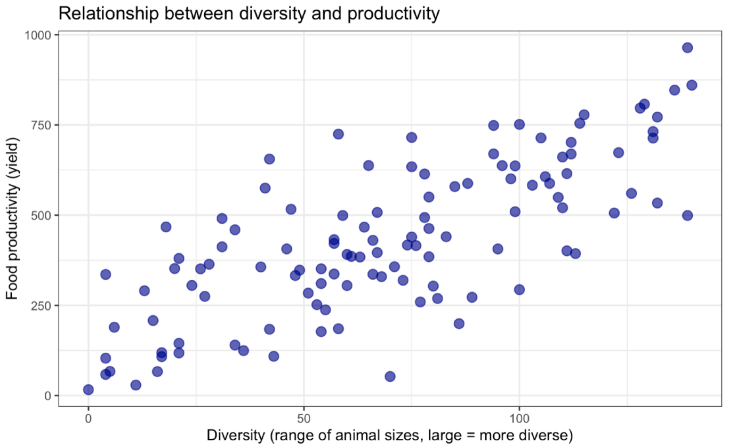
No relationship

Strong negative relationship

Doing it in R

```
> cor.test(d$diversity,d$yield)

    Pearson's product-moment correlation

data:  d$diversity and d$yield
t = 11.979, df = 113, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6544954 0.8189172
sample estimates:
      cor
0.7479659
```

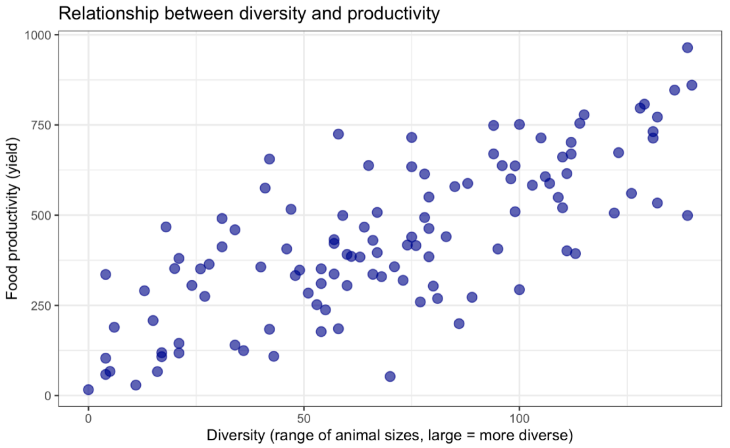
This is telling you what it's calculating and what data it's using



Doing it in R

```
> cor.test(d$range,d$yield)

    Pearson's product-moment correlation

data:  d$diversity and d$yield
t = 11.979, df = 113, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6544954 0.8189172
sample estimates:
      cor
0.7479659
```

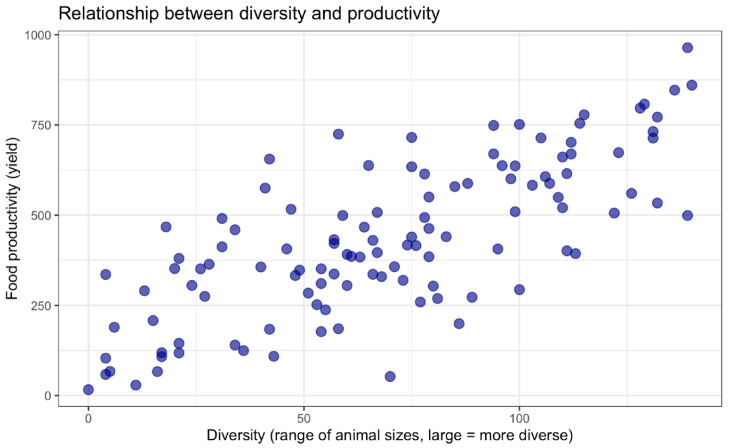
Correlation is significant!



Doing it in R

```
> cor.test(d$range,d$yield)

    Pearson's product-moment correlation

data:  d$diversity and d$yield
t = 11.979, df = 113, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6544954 0.8189172
sample estimates:
      cor
0.7479659
```

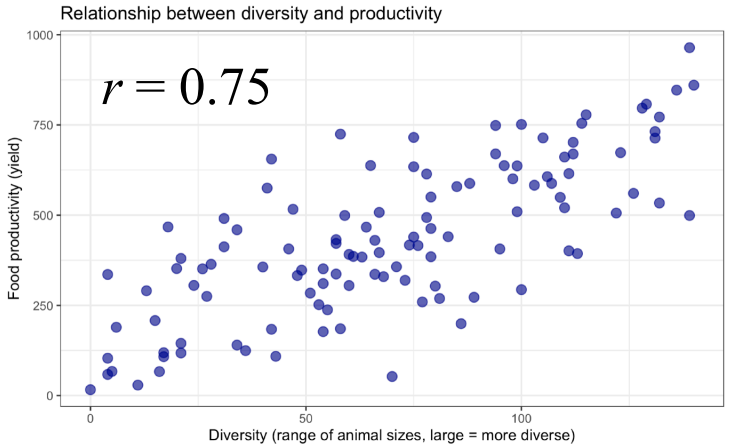
Why is this a t-statistic and this our hypothesis? (I'll answer later, for now let's ignore)



Doing it in R

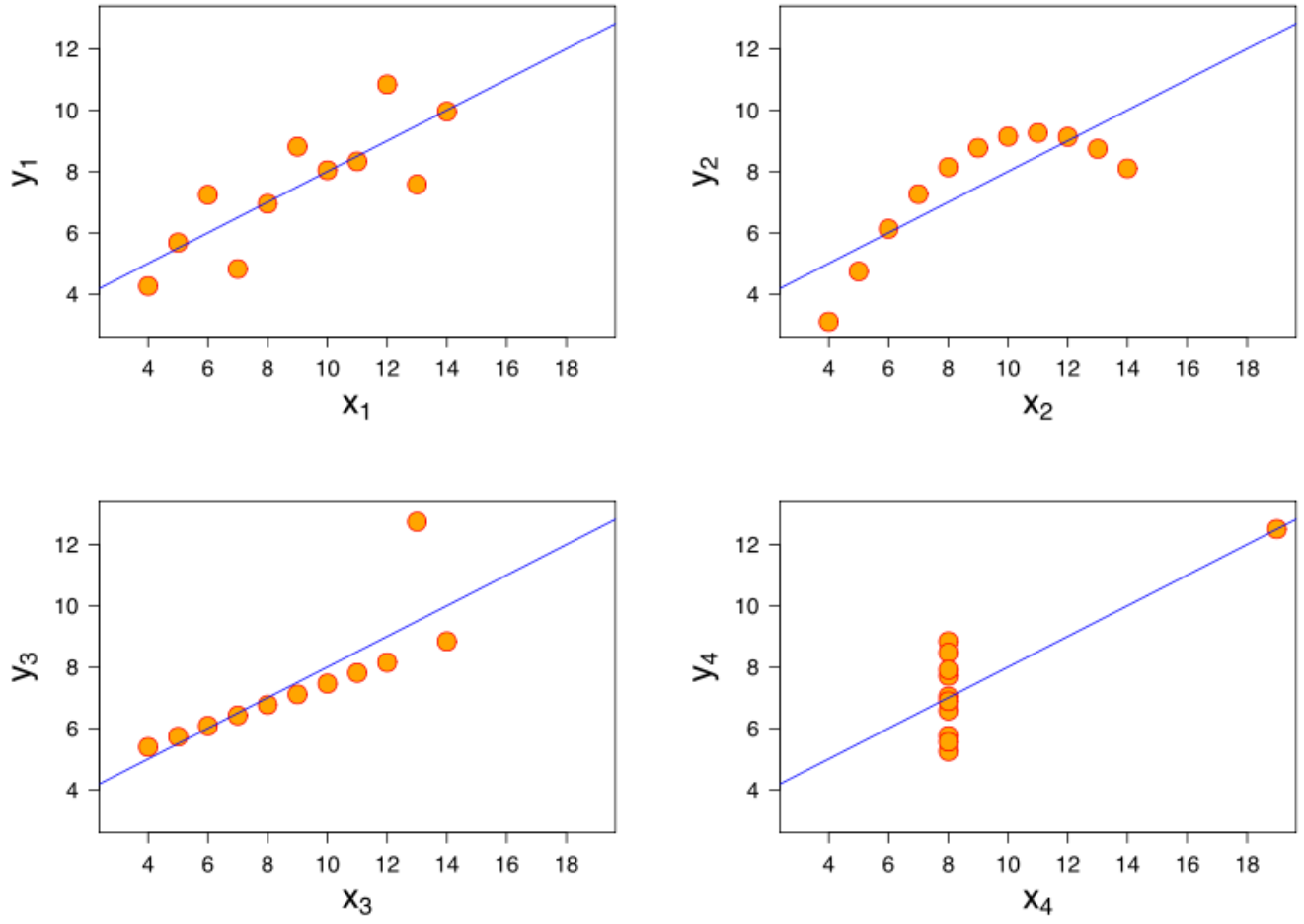
```
> cor.test(d$range,d$yield)

    Pearson's product-moment correlation

data:  d$diversity and d$yield
t = 11.979, df = 113, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6544954 0.8189172
sample estimates:
      cor
0.7479659
```

Correlation coefficient and confidence interval around it



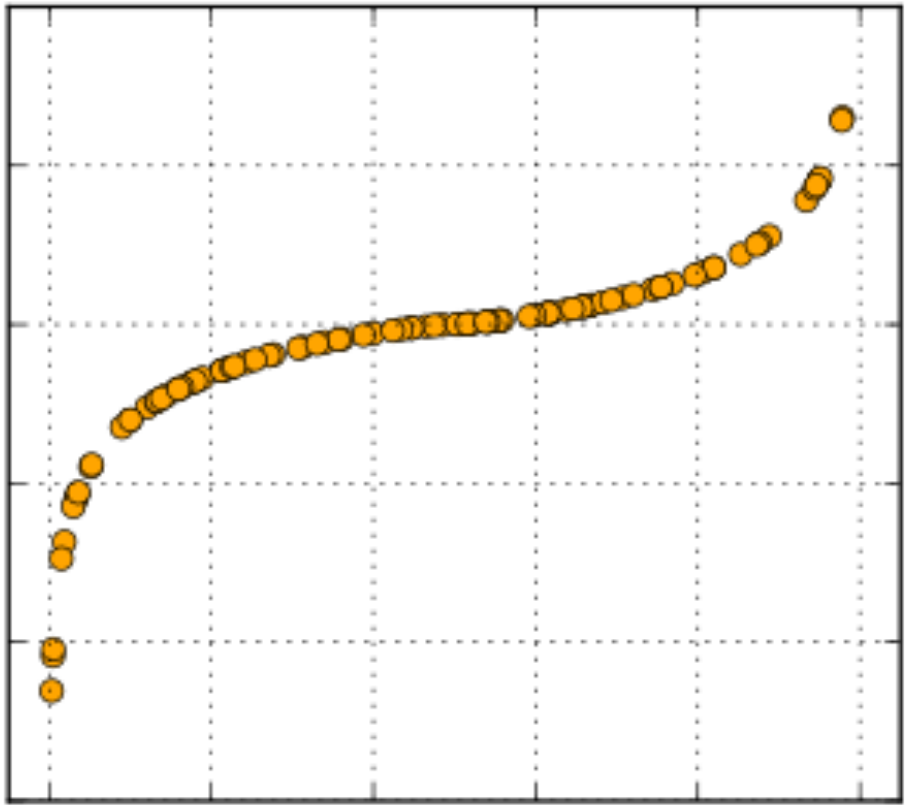
You should always plot your data part 3437847

These all have the same Pearson correlation!



Spearman correlations

This is a perfect **non-linear** relationship

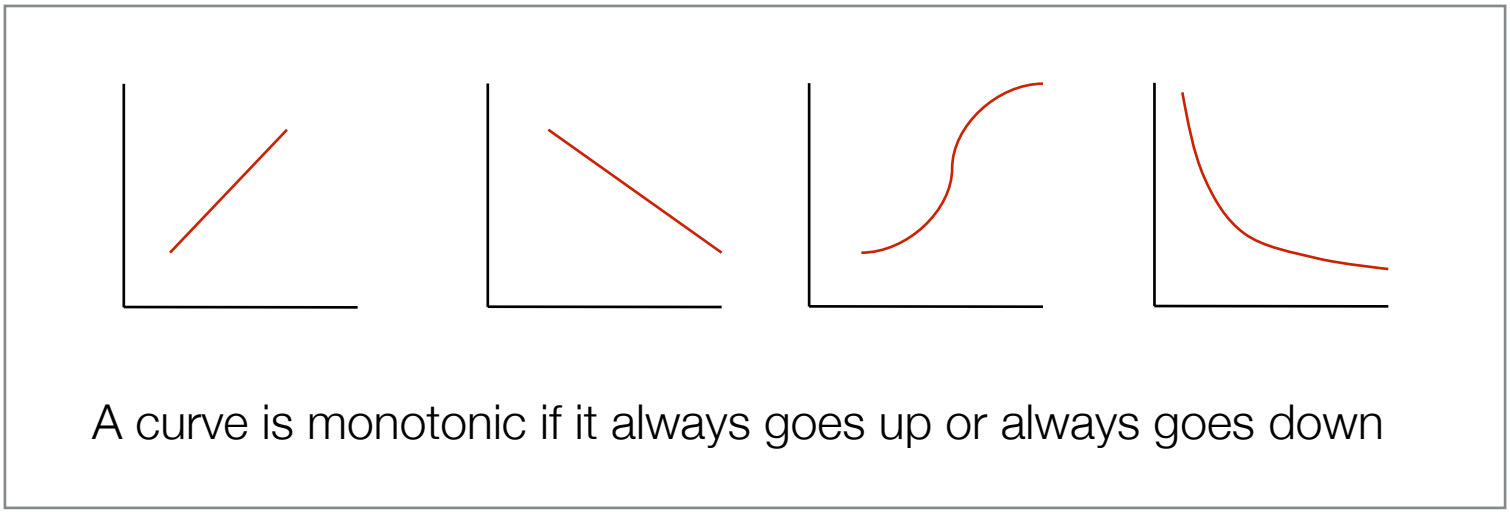


A Pearson's r of .88 misrepresents what's really going on here

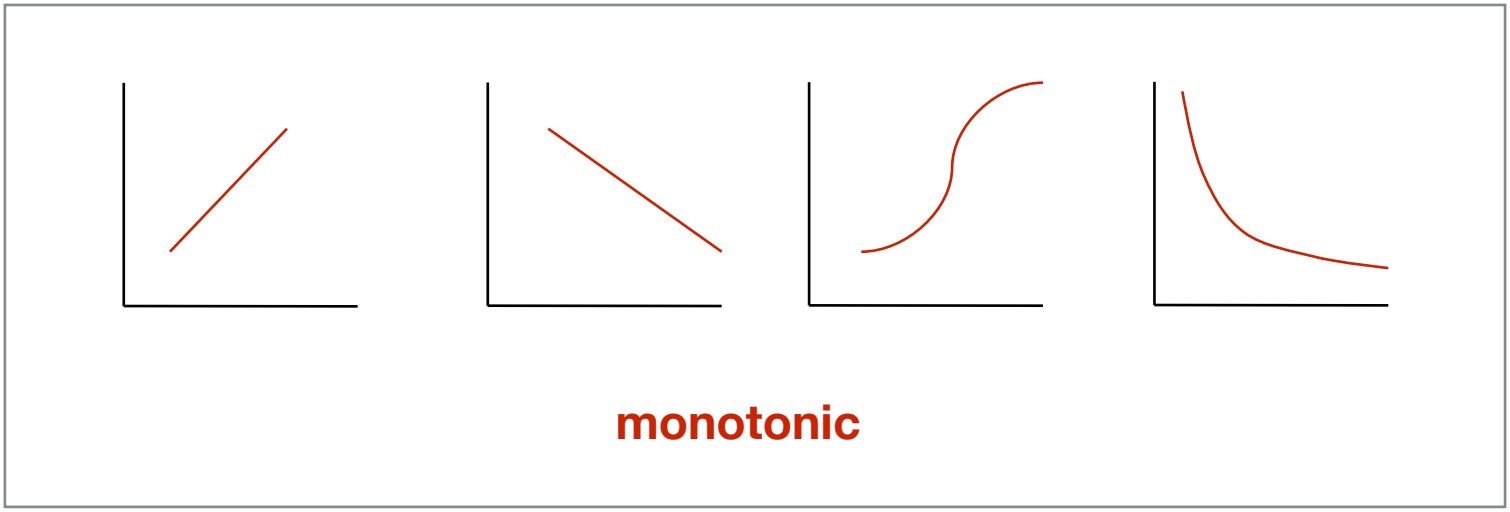
What we want is a statistic that says that this is a relationship of strength 1… it just happens to be a non-linear relationship

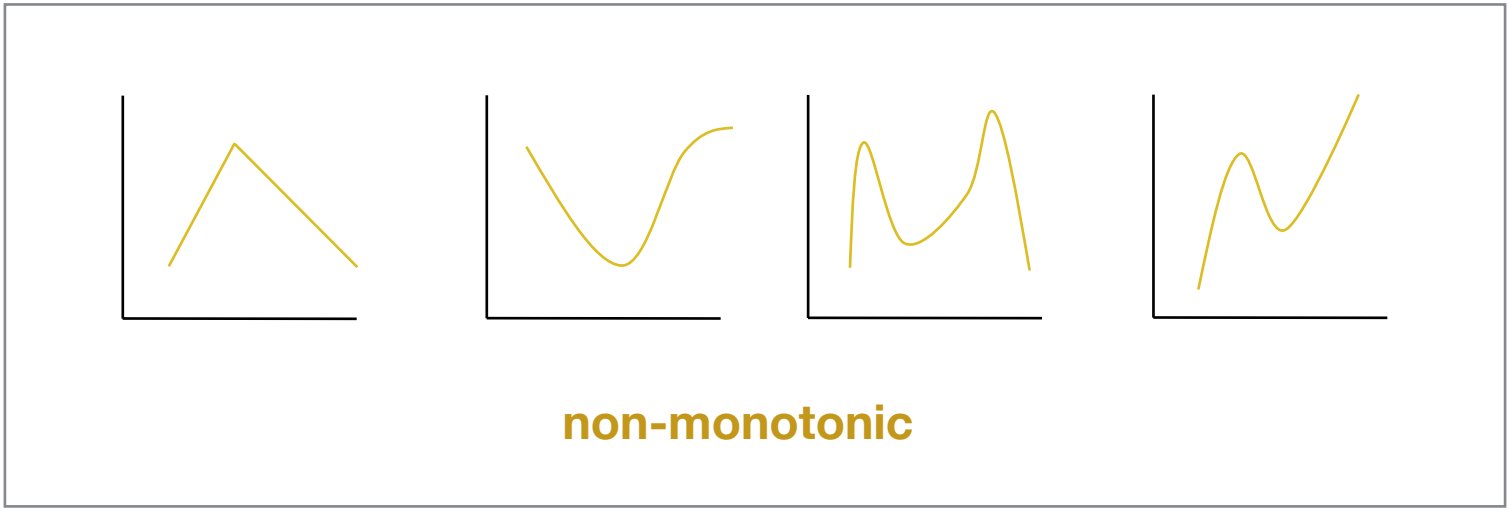
Spearman's rho (written: ρ)

- Spearman's correlation is similar to Pearson's
  - It's a number from -1 (perfect negative) to 1 (perfect positive)
  - However, it doesn't assume the relationship is linear.
  - All it assumes is monotonicity:



A curve is monotonic if it always goes up or always goes down

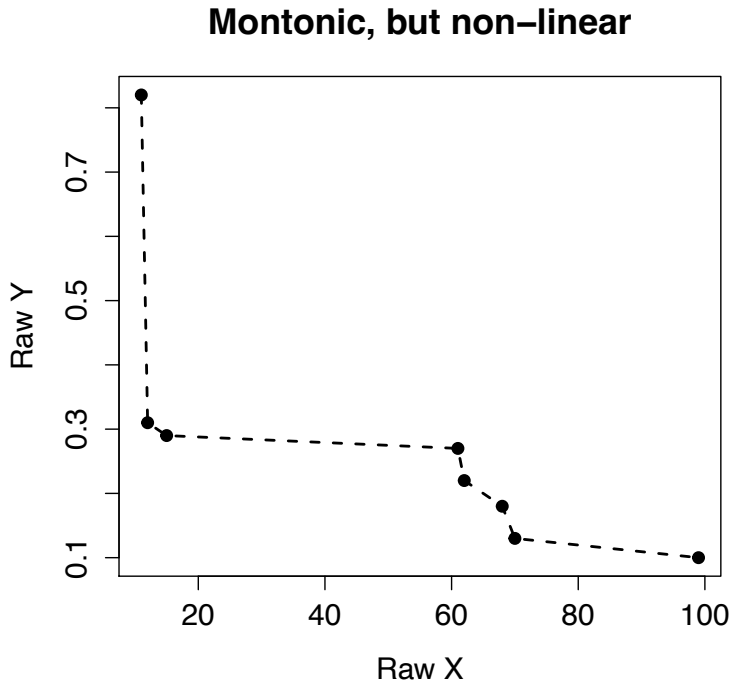

monotonic


non-monotonic

How it works

- Convert all the raw data to ranks
- Then apply Pearson correlations to the ranks

raw X	raw Y
11	0.82
12	0.31
15	0.29
61	0.27
62	0.22
68	0.18
70	0.13
99	0.10

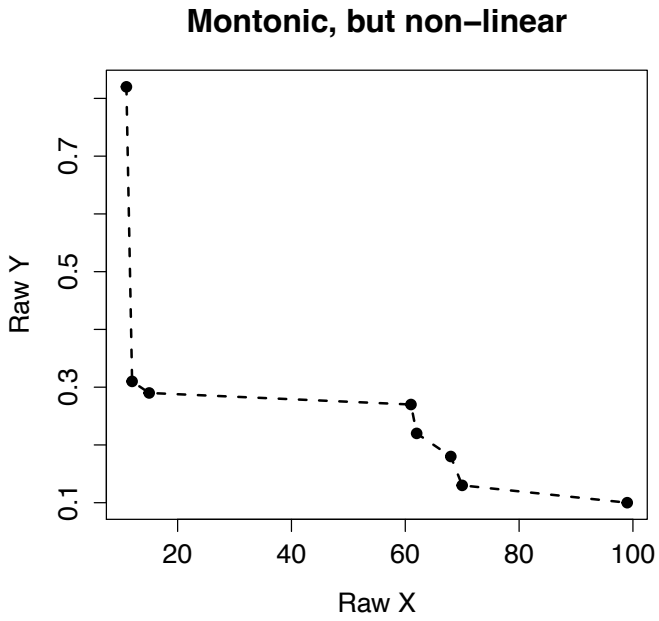


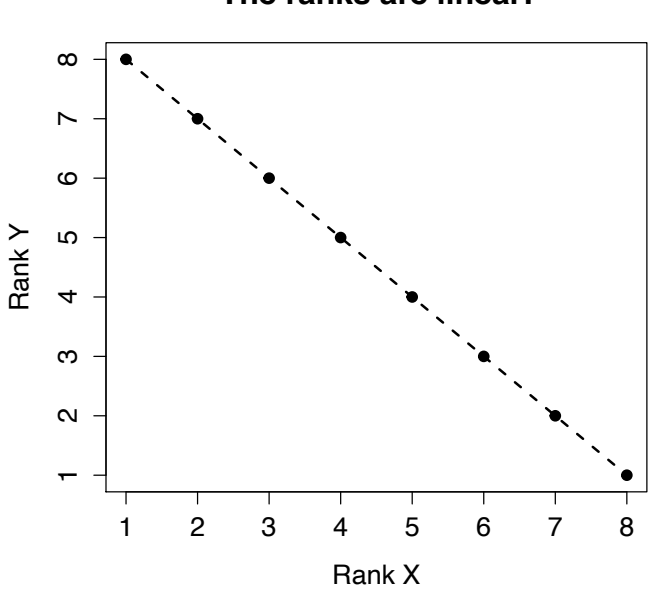
How it works

- Convert all the raw data to ranks
- Then apply Pearson correlations to the ranks

raw X	raw Y	rank X	rank Y
11	0.82	1	8
12	0.31	2	7
15	0.29	3	6
61	0.27	4	5
62	0.22	5	4
68	0.18	6	3
70	0.13	7	2
99	0.10	8	1

Rank all the values on for both variables





Pearson:

If we correlate **the raw data**, we get a value of -0.71

raw X	raw Y	rank X	rank Y
11	0.82	1	8
12	0.31	2	7
15	0.29	3	6
61	0.27	4	5
62	0.22	5	4
68	0.18	6	3
70	0.13	7	2
99	0.10	8	1

Spearman:

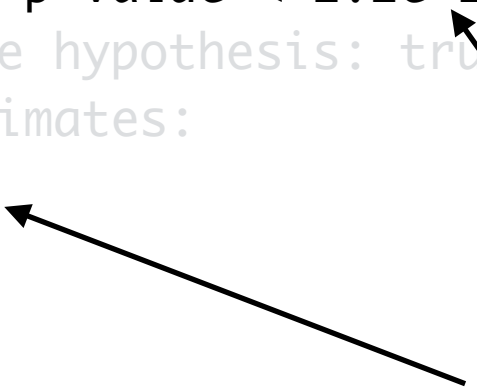
If we correlate **the ranks**, we get a value of -1

Spearman's correlation in R

```
> cor.test(d$diversity, d$yield, method="spearman")

   Spearman's rank correlation rho

data:  d$diversity and d$yield
S = 69664, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7251483
```

Spearman's correlation in R

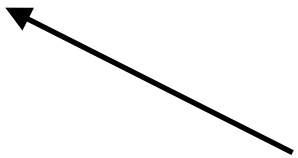
```
> cor.test(d$diversity, d$yield, method="spearman")

    Spearman's rank correlation rho

data:  d$diversity and d$yield
S = 69664, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7251483
```



As before these are the important bits

Spearman's correlation in R

```
> cor.test(d$diversity,d$yield,method="spearman")

    Spearman's rank correlation rho

data:  d$diversity and d$yield
S = 69664, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7251483

Warning message:
In cor.test.default(d$diversity,d$yield,method="spearman") :
  Cannot compute exact p-value with ties
```



This warning occurs because it had some values that were exactly the same, so it got tied ranks. Not an issue unless you have lots of this — good to check the dataset to see, but mostly don't worry hugely at this point

Exercises are in w9day1exercises.Rmd