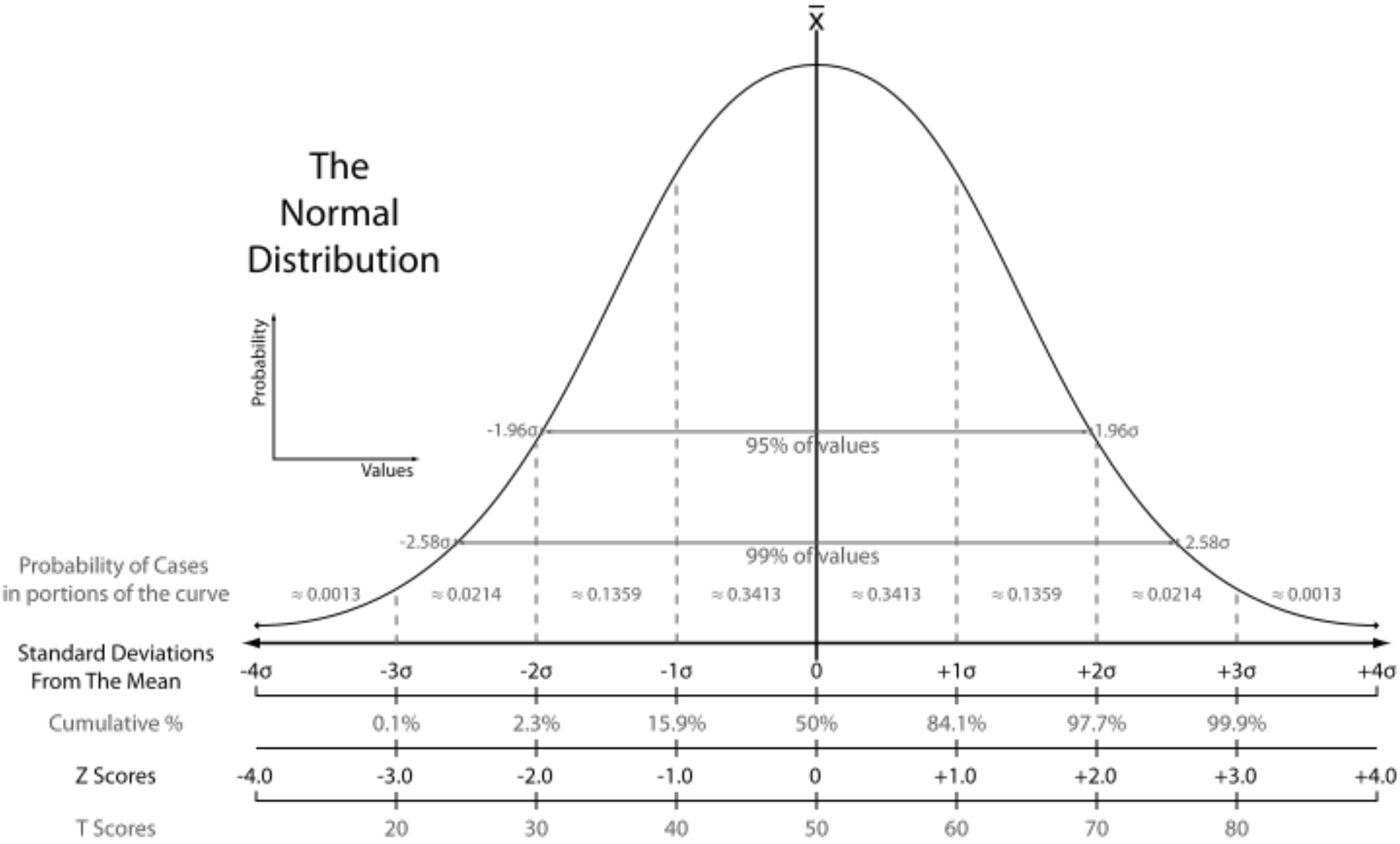
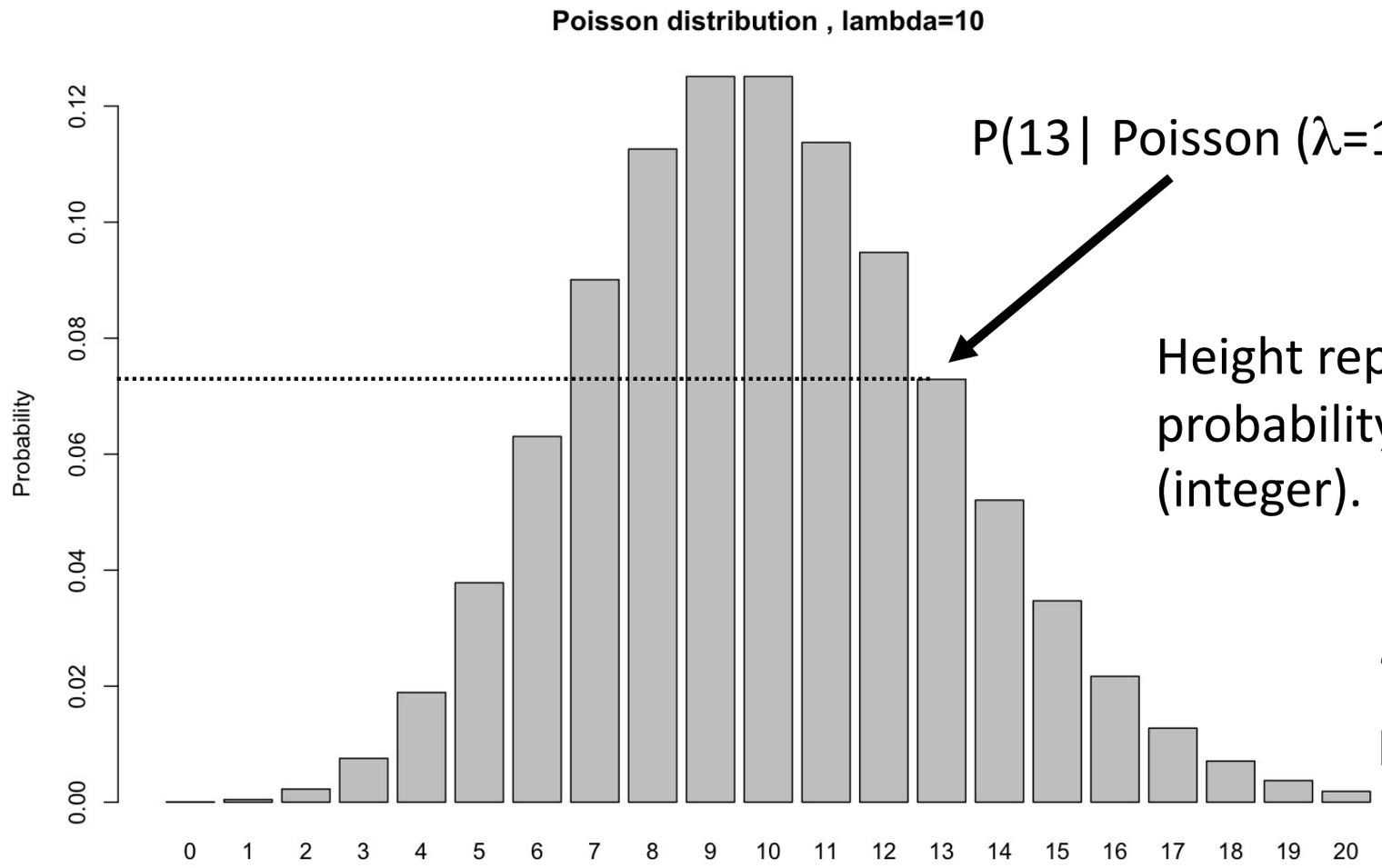
How to Statistics

Amanda Charbonneau

The Normal Distribution



Probability Mass function (For discrete distributions, like read counts)



P(integer) ≥ 0

P(non-integers) = 0.

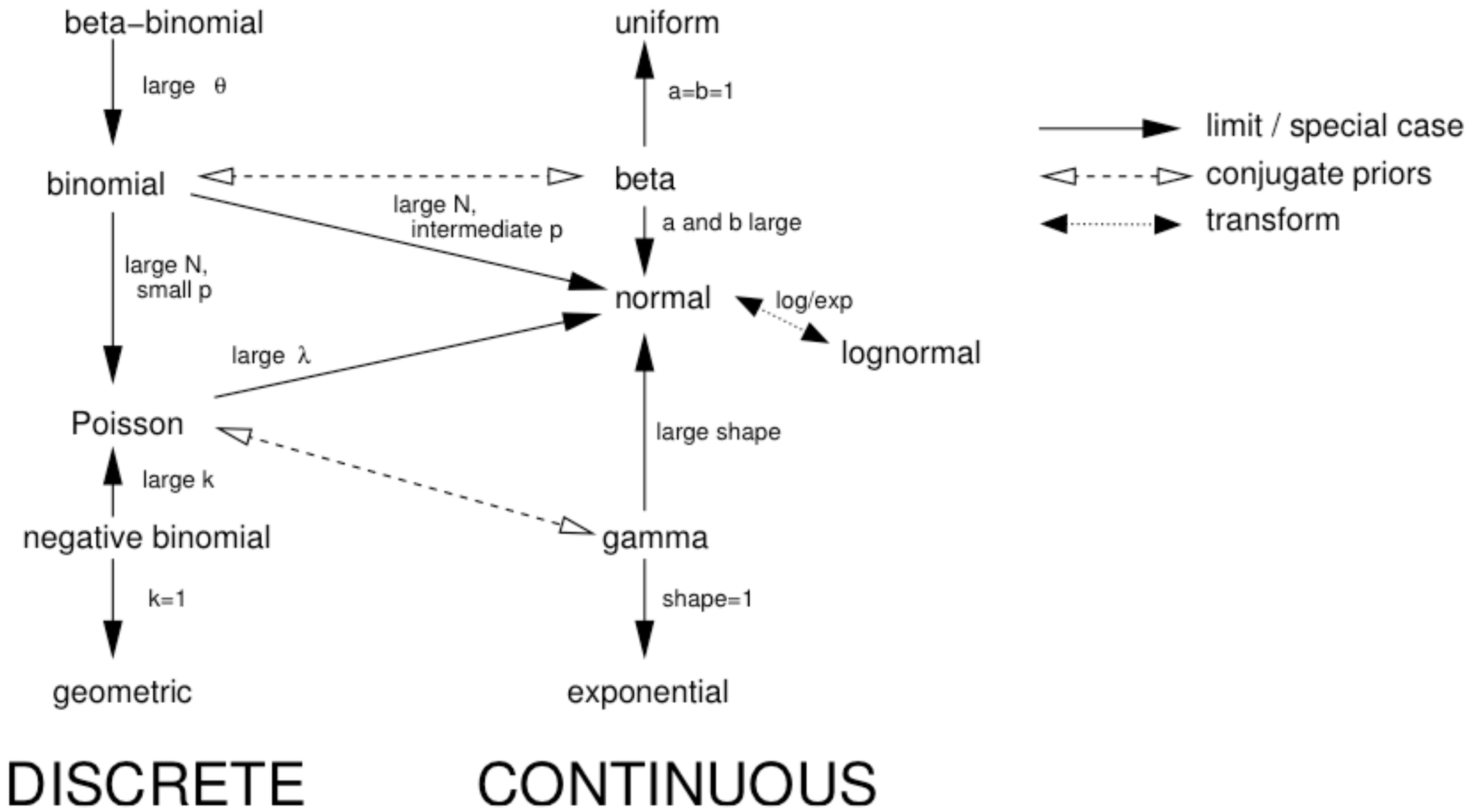


Figure 4.17 Relationships among probability distributions.

Bolker 2007 CH4 page 137

Negative binomial

$$\text{Negative Binomial Distribution} = \frac{\Gamma(k+x)}{\Gamma(k)x!}\left(\frac{k}{k+\mu}\right)^{k}\left(\frac{\mu}{k+\mu}\right)^{x}$$

Expected number of counts =μ
Over-dispersion parameter = k

For our purposes all we care about is that

$$\text{var}(x) = \mu + k\mu^2$$

“To consult the statistician after an experiment is finished is often merely to ask him(her) to conduct a post mortem examination. He(she) can perhaps say what the experiment died of.”

–Ronald Fisher

I have an idea…

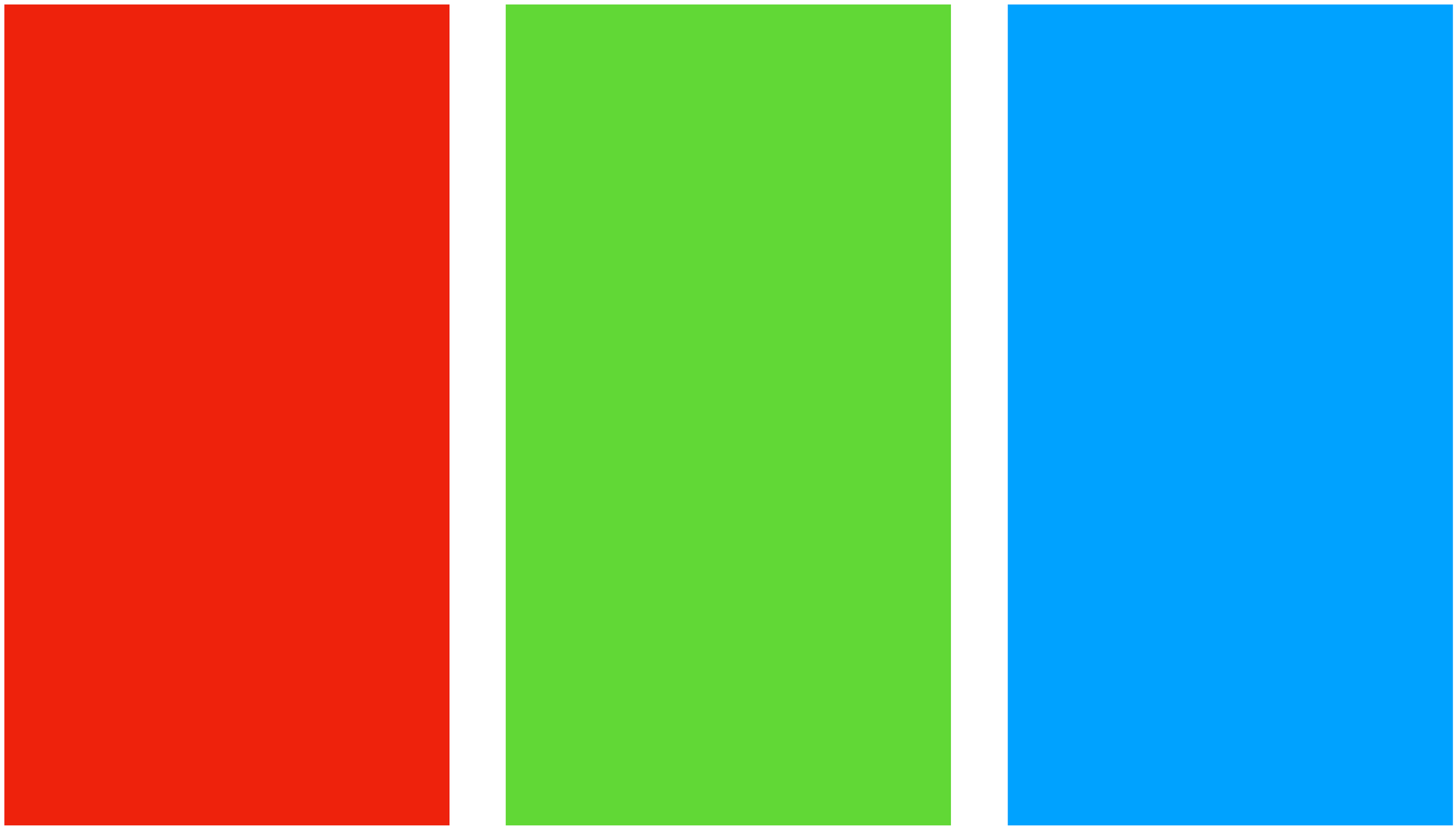
- 144 individuals
- 48 of each treatment
- Treatment lasts 1 week
- We have 3 incubators/ greenhouses/tanks/cages which each hold 48 individuals

Experimental Design

Place 1

Place 2

Place 3



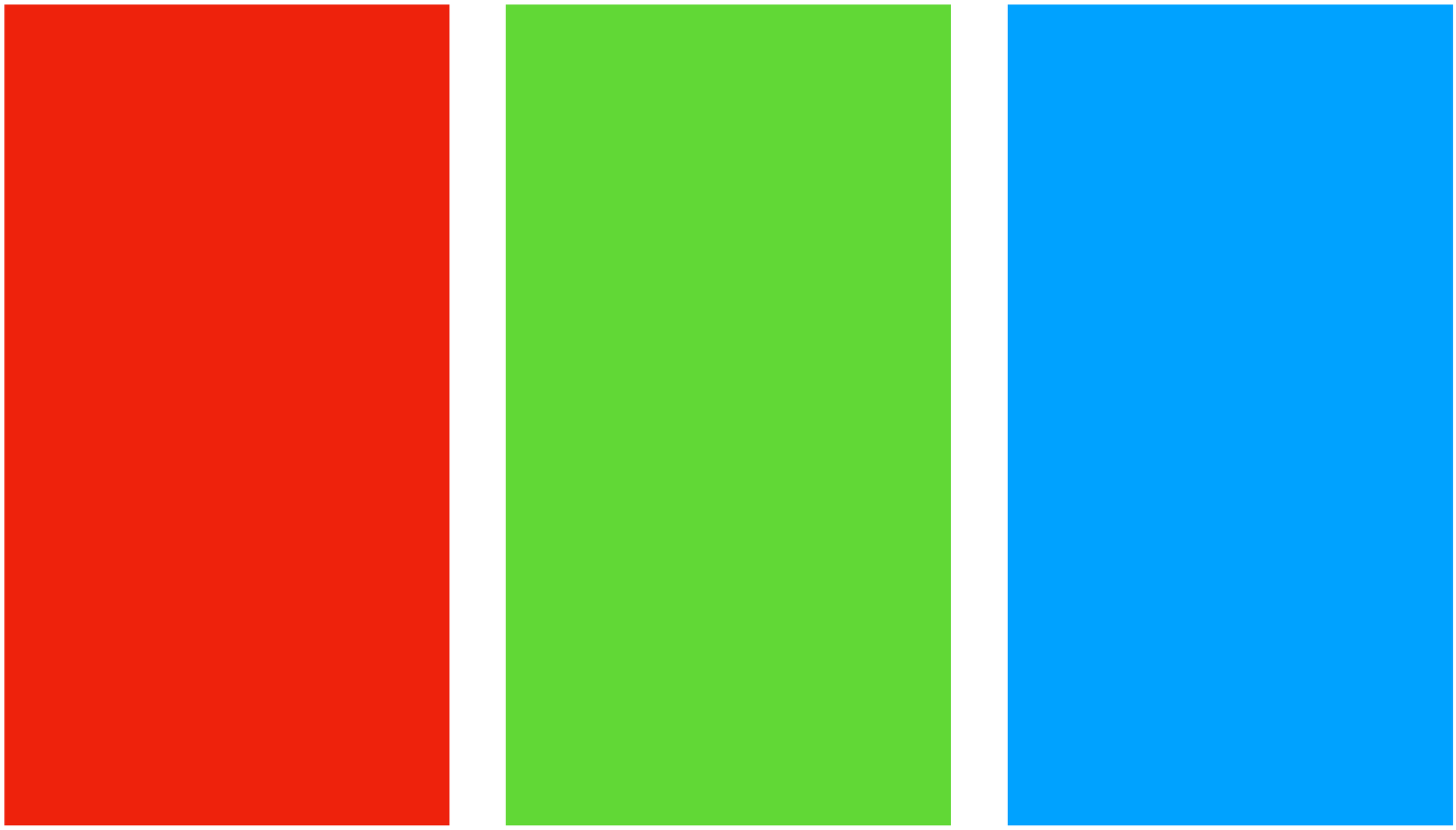
This is the plan. What do you think?

Experimental Design

Place 1

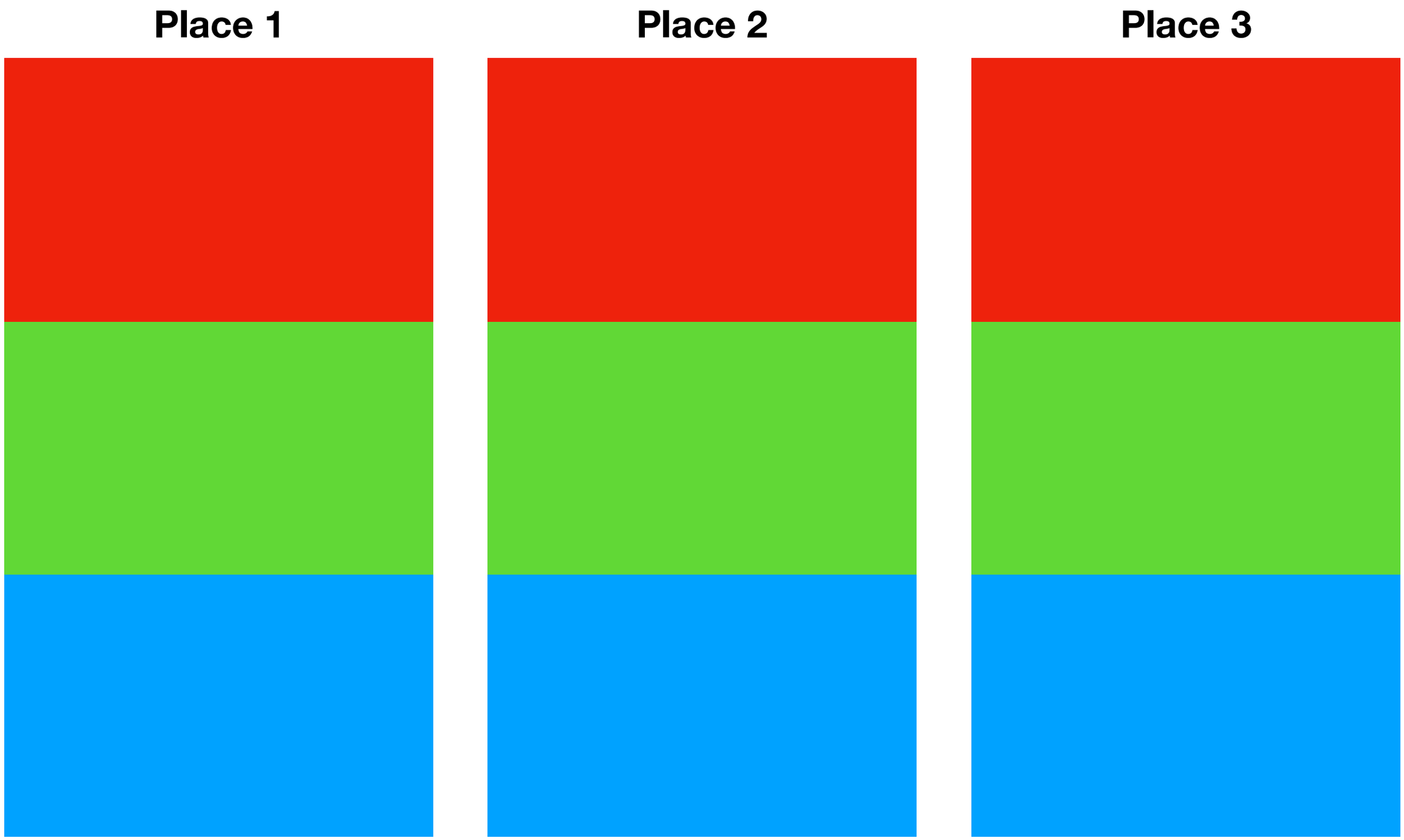
Place 2

Place 3



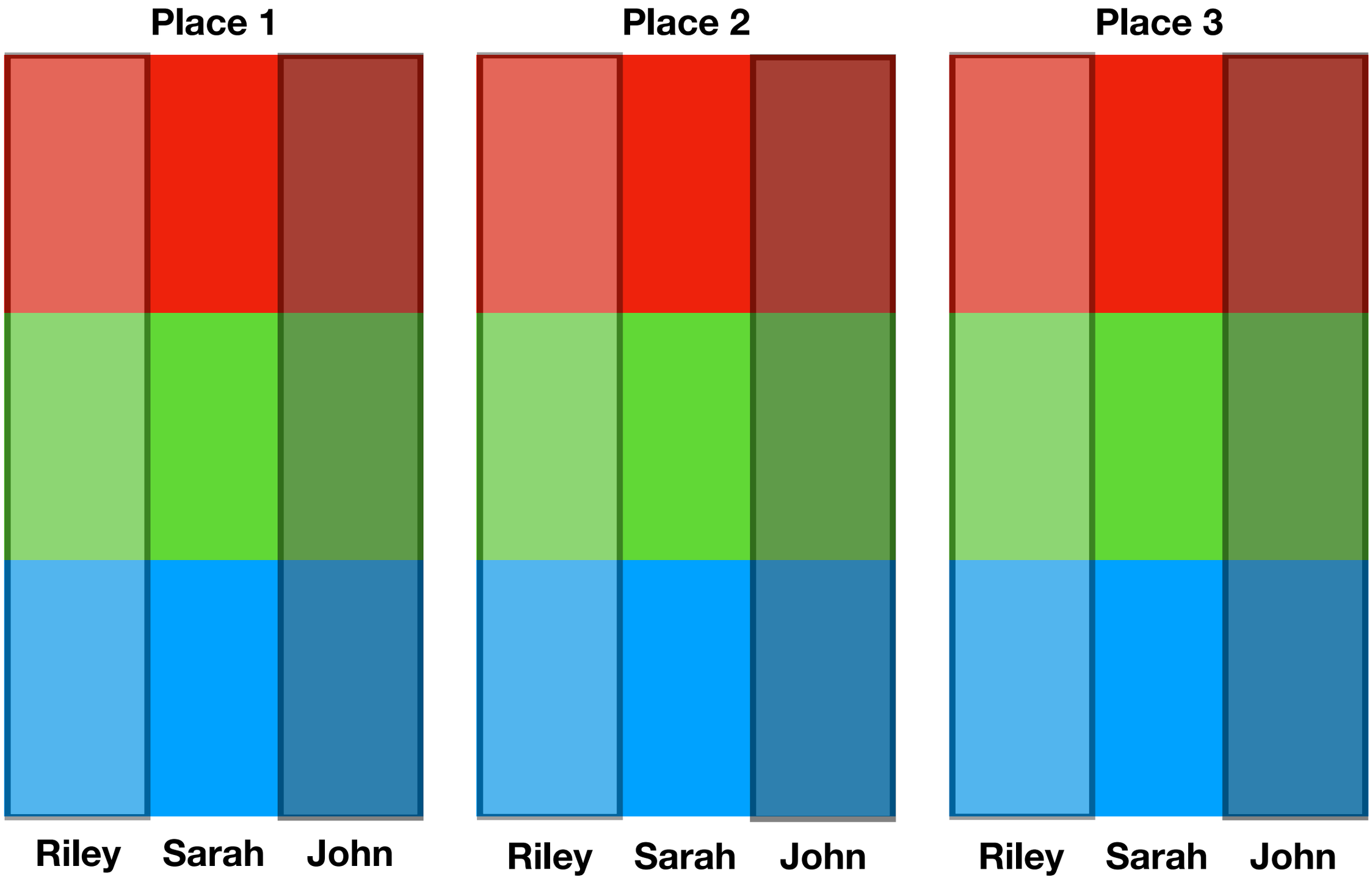
This is the plan. What do you think?

Experimental Design



You have 3 undergrads. How should they split the data collection work?

Experimental Design

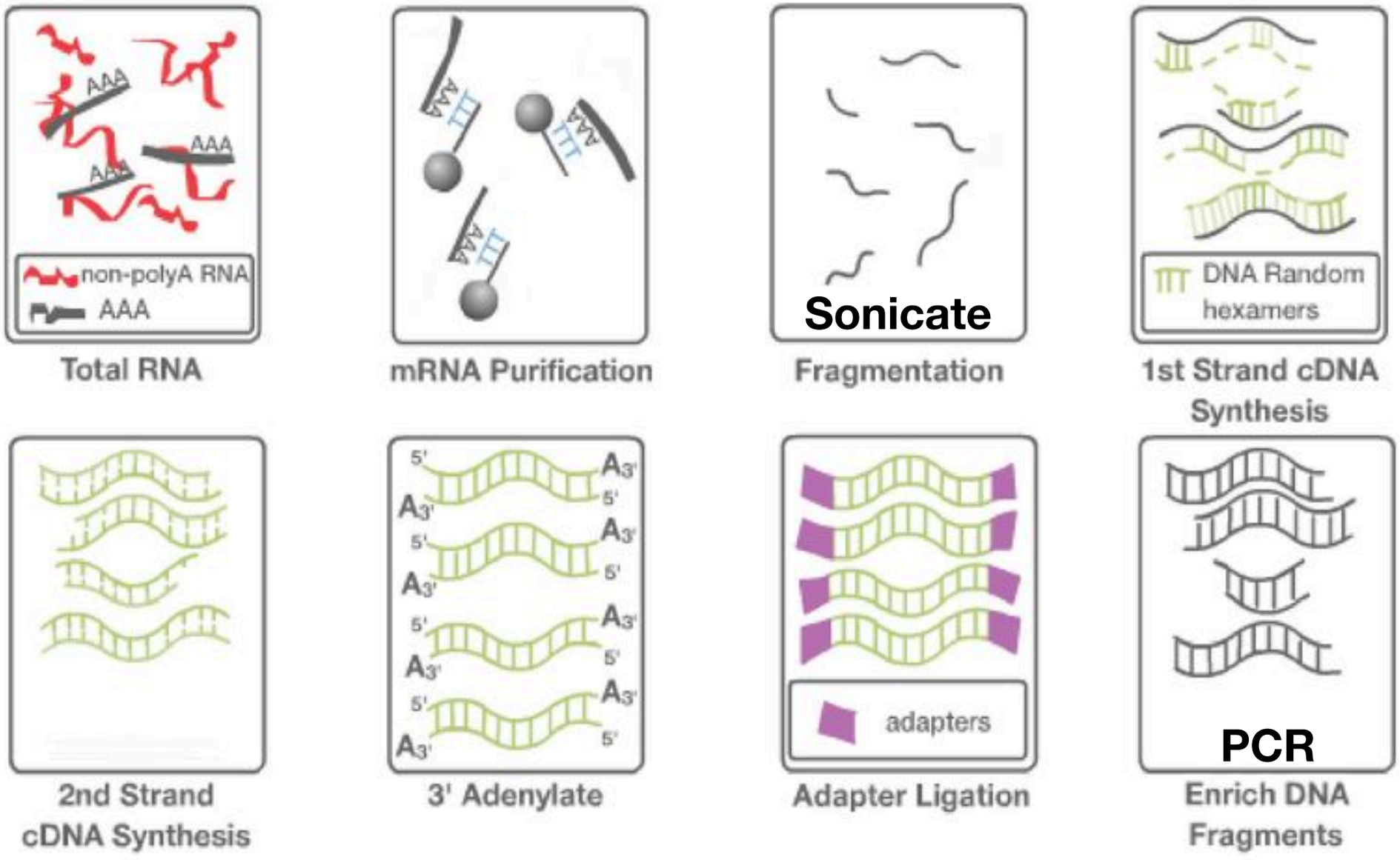


Experimental Design

Sample	Treatment	Place	Student	Measurement
1		1	Riley	92
2		1	Sarah	56
3		1	John	21
4		2	John	77
5		2	Riley	35
6		2	Sarah	26
7		3	Sarah	68
8		3	John	41
9		3	Riley	42

Let’s say you collect RNA for sequencing from all your organisms, what other variables should we record?

1. TruSeq® RNA Sample Prep Kit V2

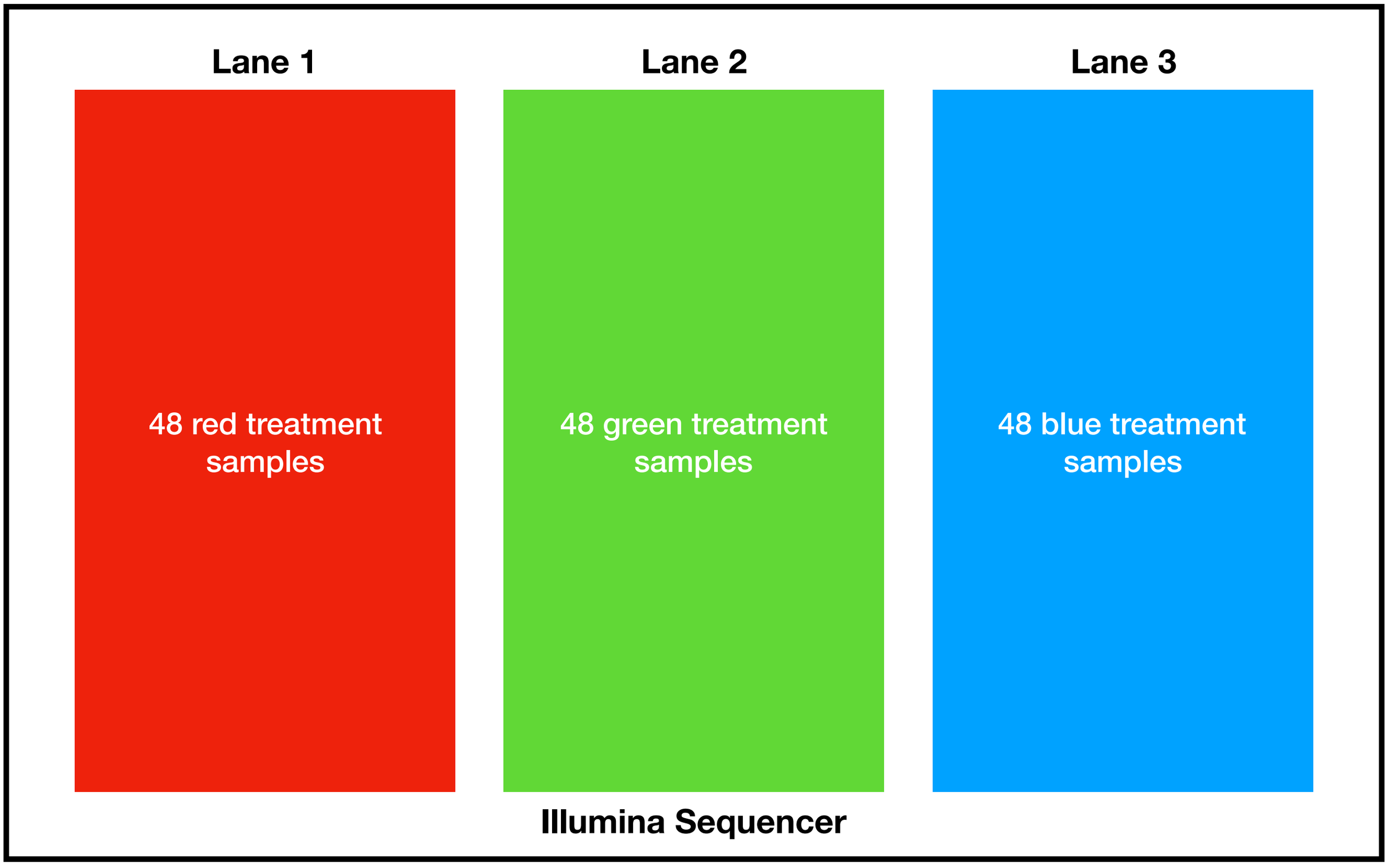


Most people can prep 8 to 16 samples at a time. A robot can do 96.

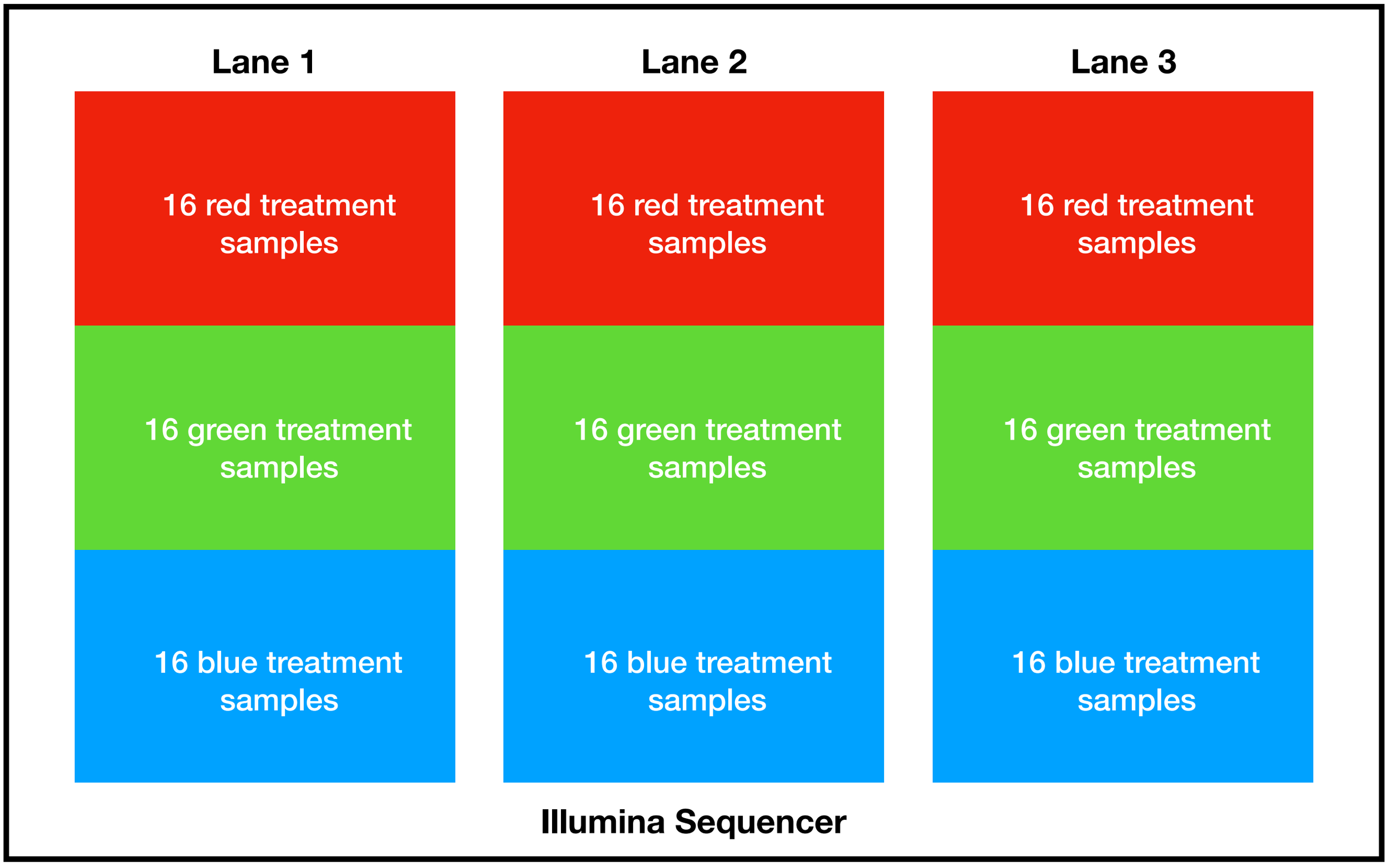
If this is going to be an RNAseq experiment, what variables should we try to account for?

- Time collected RNA
- Extraction method/batch
- Time in storage
- PCR differences
- When/where samples collected
- Kit number

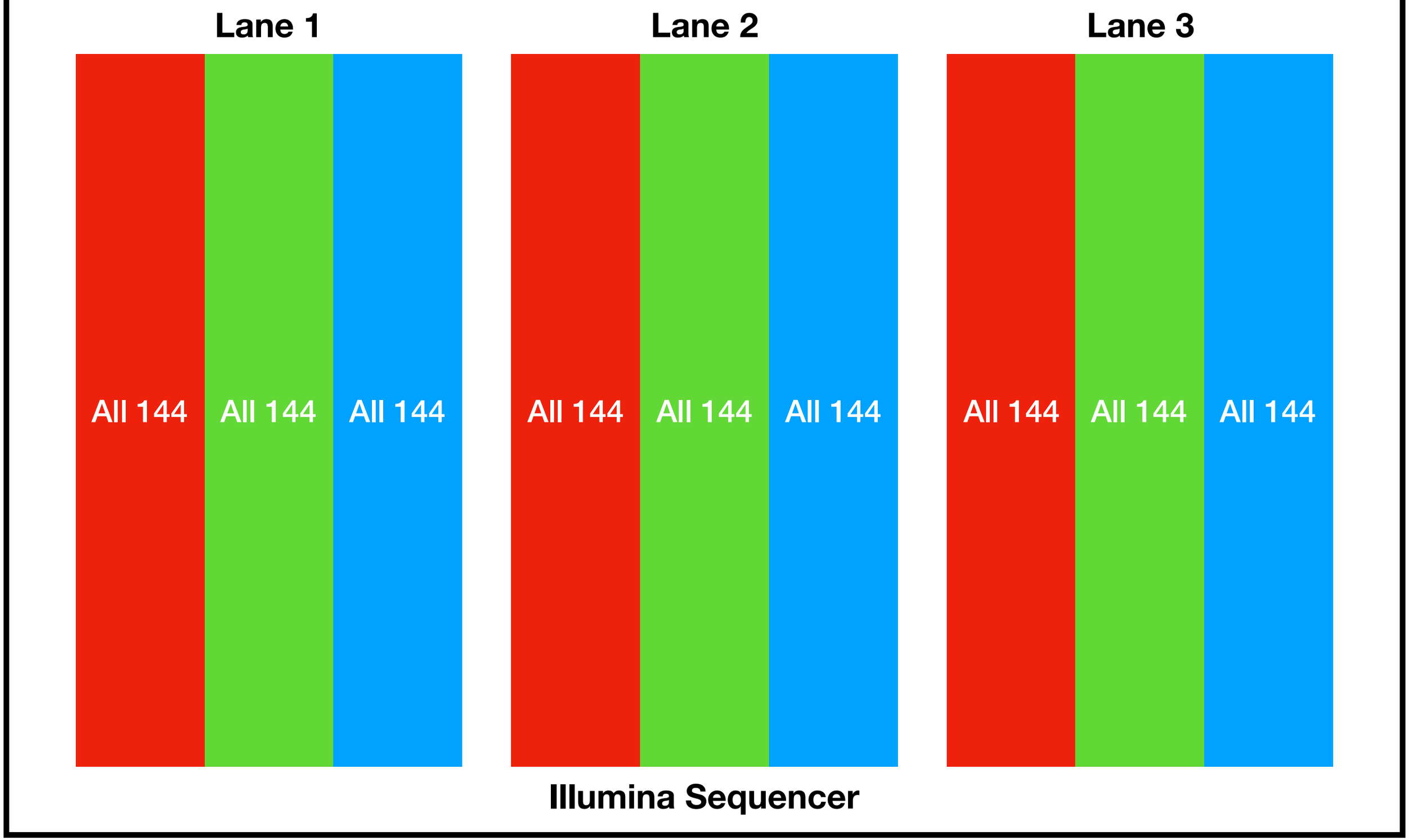
Is this how you load your sequencer?



Is this how you load your sequencer?



Experimental Design



If this is going to be an RNAseq experiment, what variables should we try to account for?

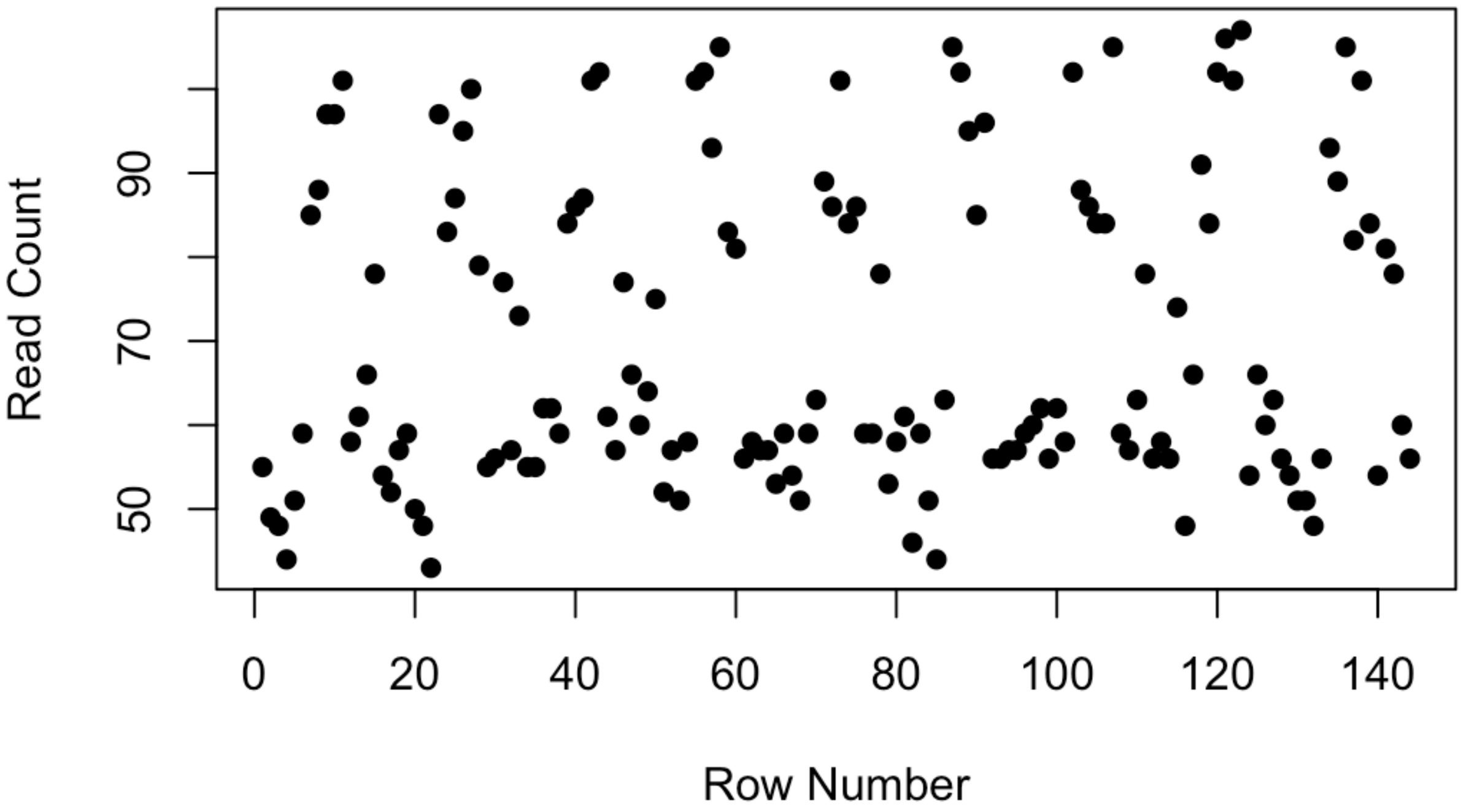
- Time collected RNA
- Extraction method/batch
- Time in storage
- PCR differences
- When/where samples collected
- Kit number
- Sequencing date
- Sequencer lane
- Index/custom barcodes
- Pooling
- Sequencing center
- Sequencer

RNAseq is an experiment, not an assay

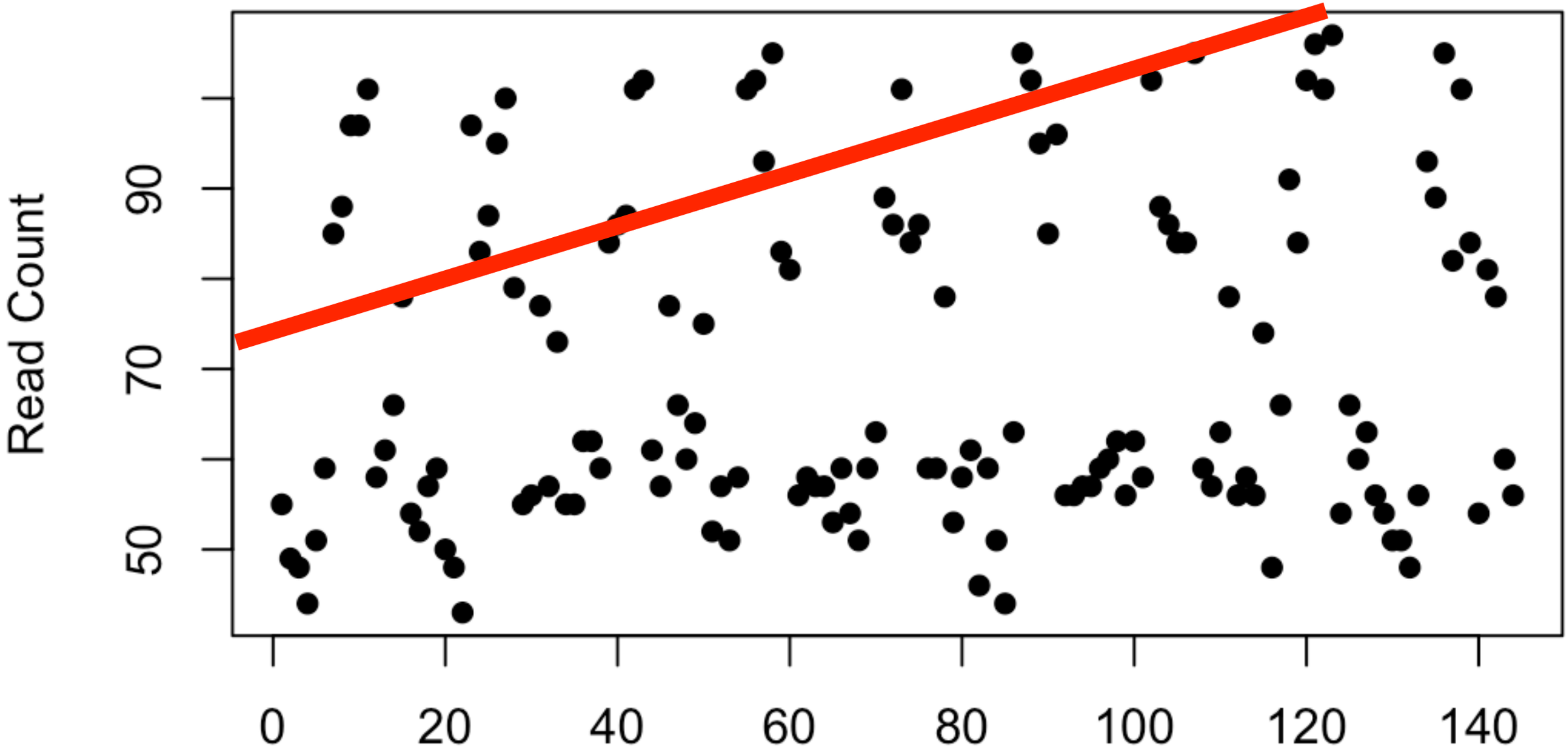
Experiment Matrix

Sample	Treatment	Place	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18

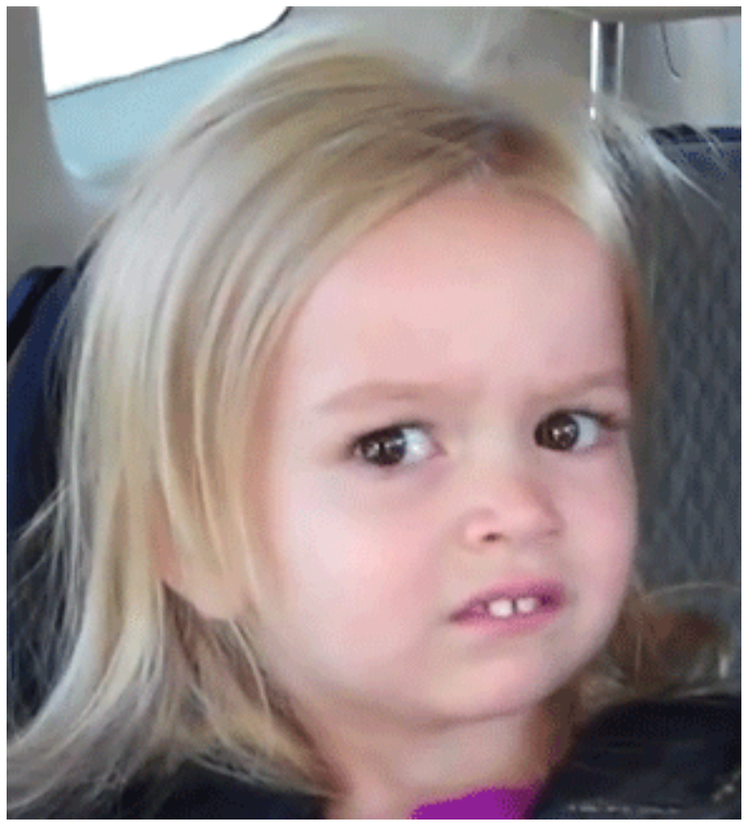
Our experimental data in a scatterplot



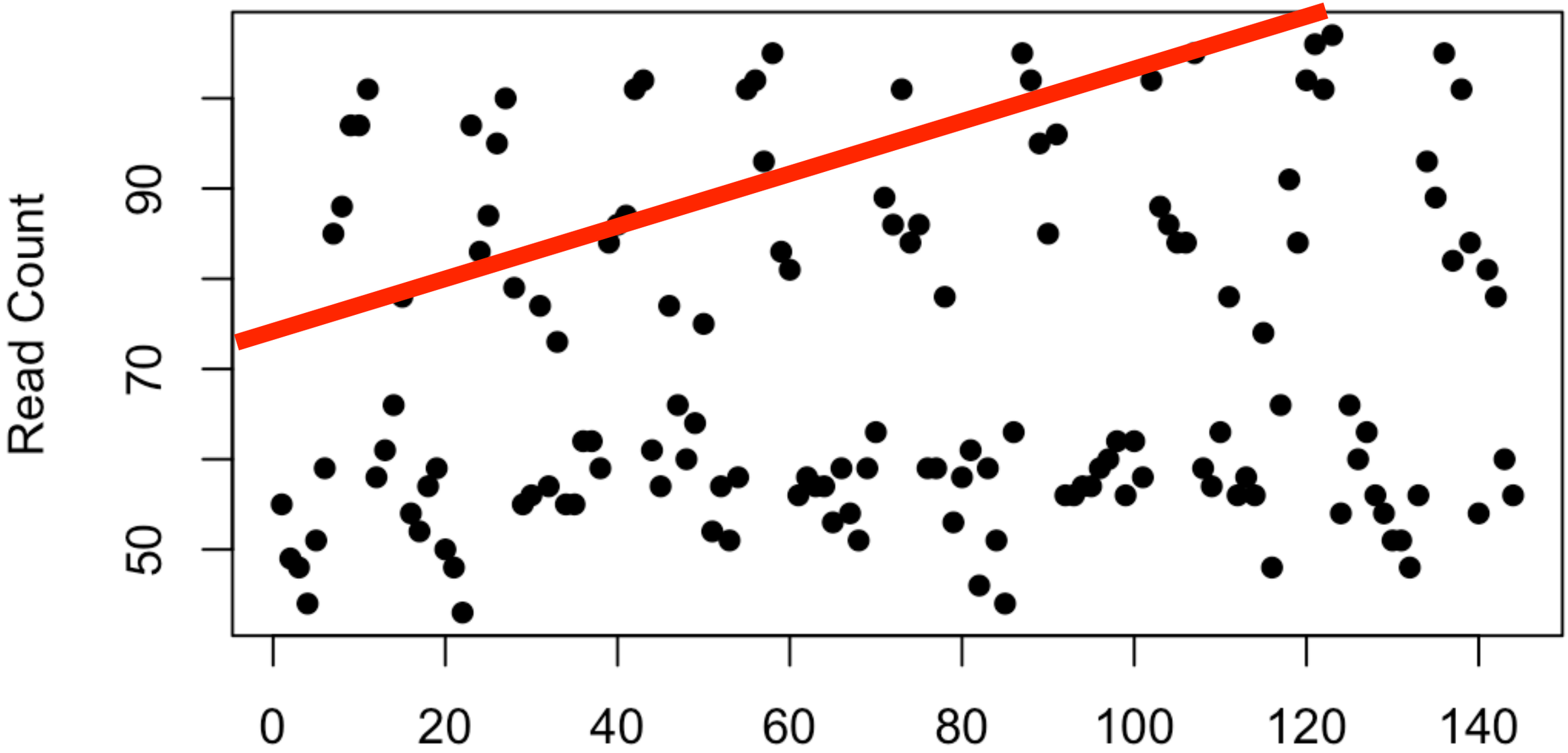
Our experimental data in a scatterplot, but now there's a line on it



How do you *feel* about that line?

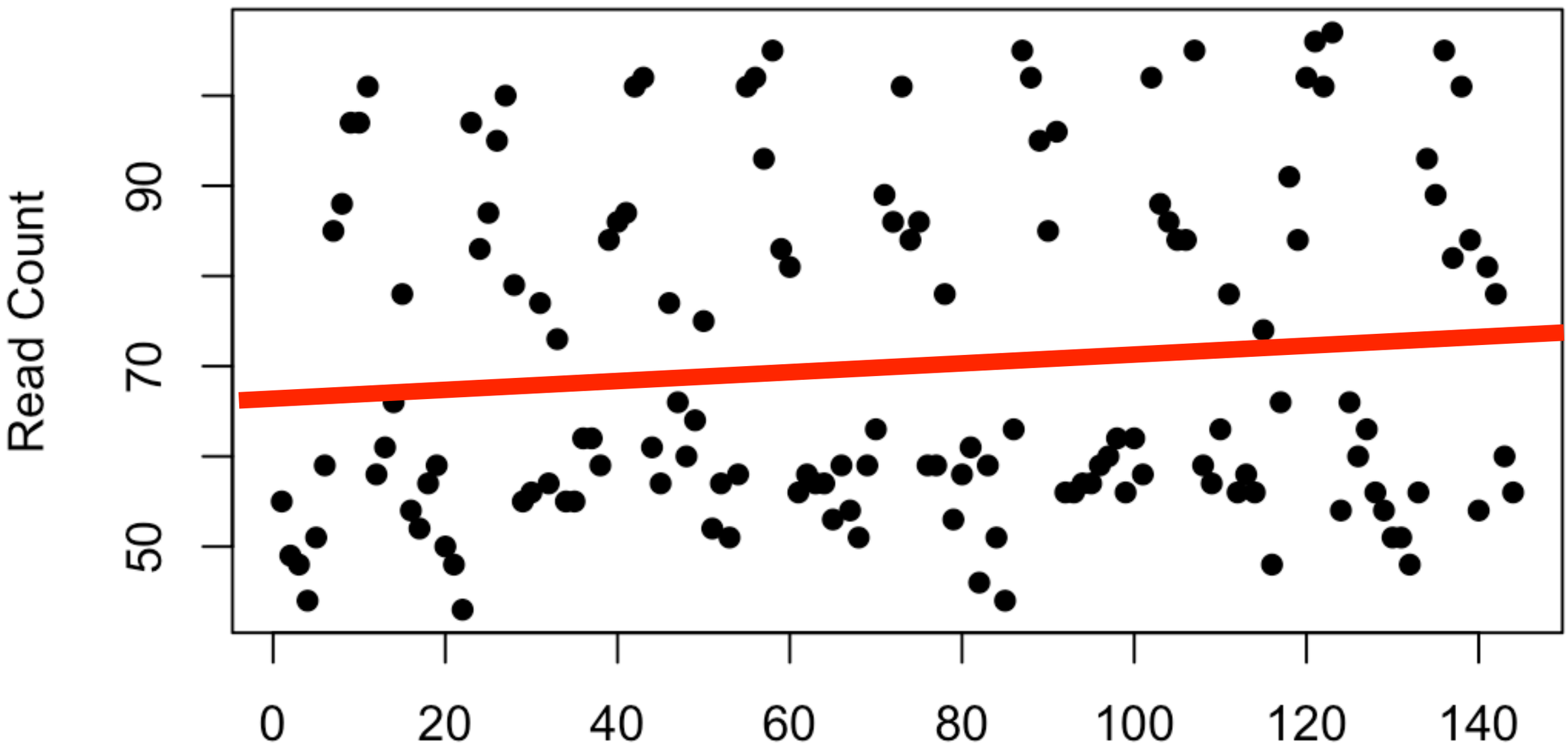


Our experimental data in a scatterplot, but now there's a line on it



Why do you *feel* that way?

Our experimental data in a scatterplot, but now there's a line on it



Is this better? Why?

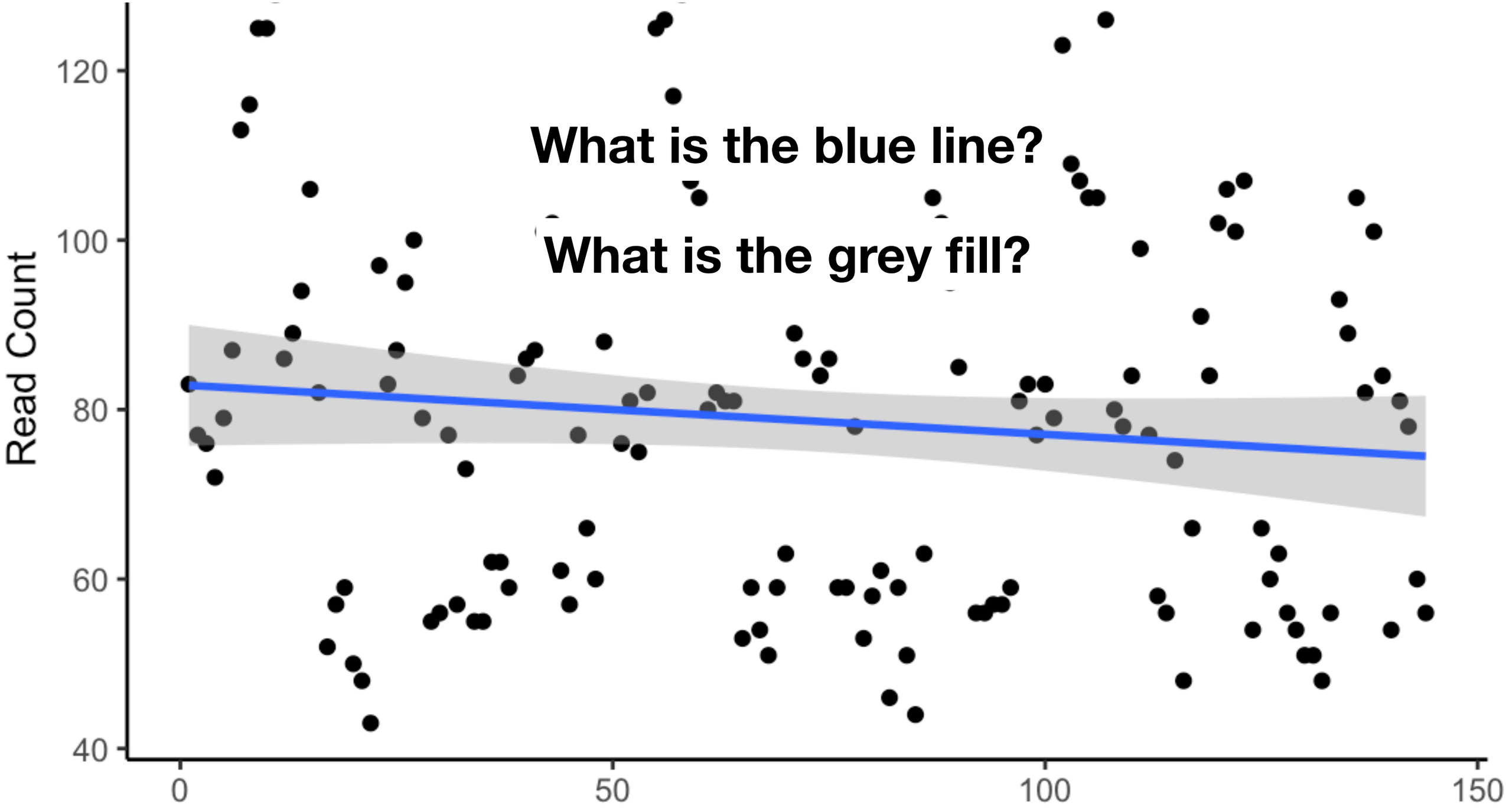
Proof we know math

- A ‘best fit’ line on a scatter plot is showing us the tendency of the dots…where their middle is, and what their slope is. It is the *average* of all those dots.

ReadCount ~ 1

What is the blue line?

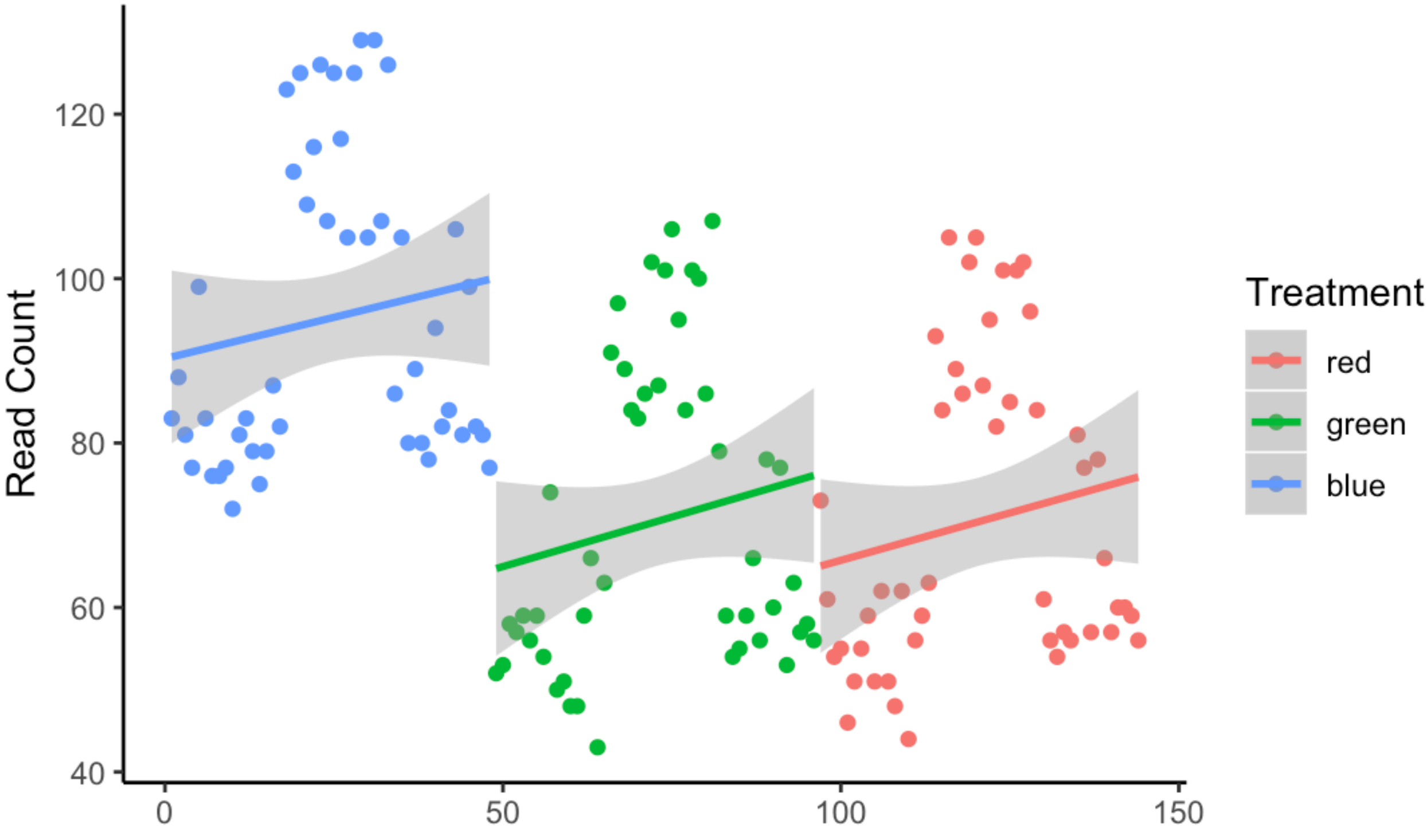
What is the grey fill?



ReadCount ~ Treatment

Sample	Treatment	Place	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18

ReadCount ~ Treatment



Proof we know math

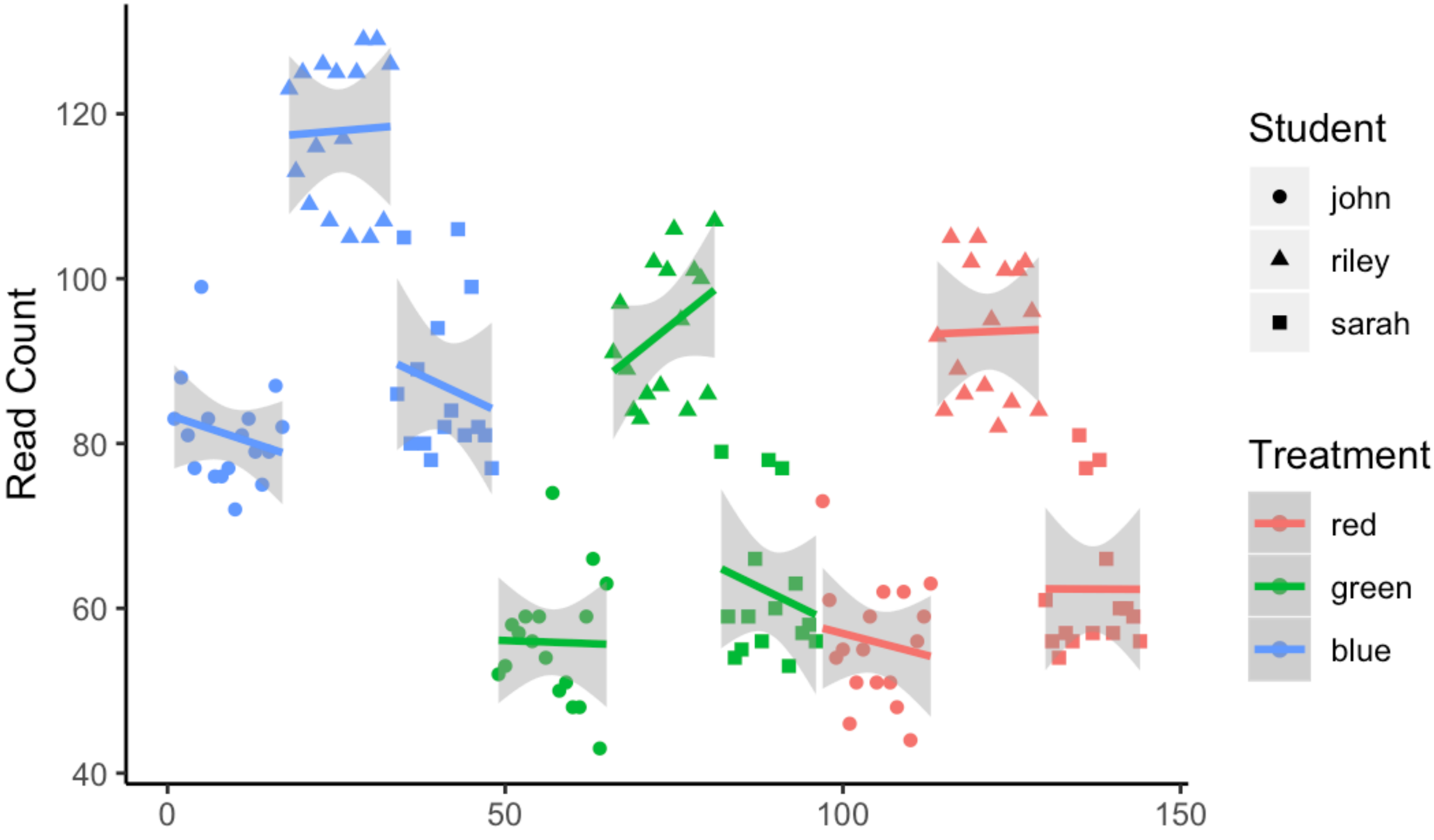
- A ‘best fit’ line on a scatter plot is showing us the tendency of the dots…where their middle is, and what their slope is. It is the *average* of all those dots.

- Mathematical modeling is really just estimating averages for the right data subsets.

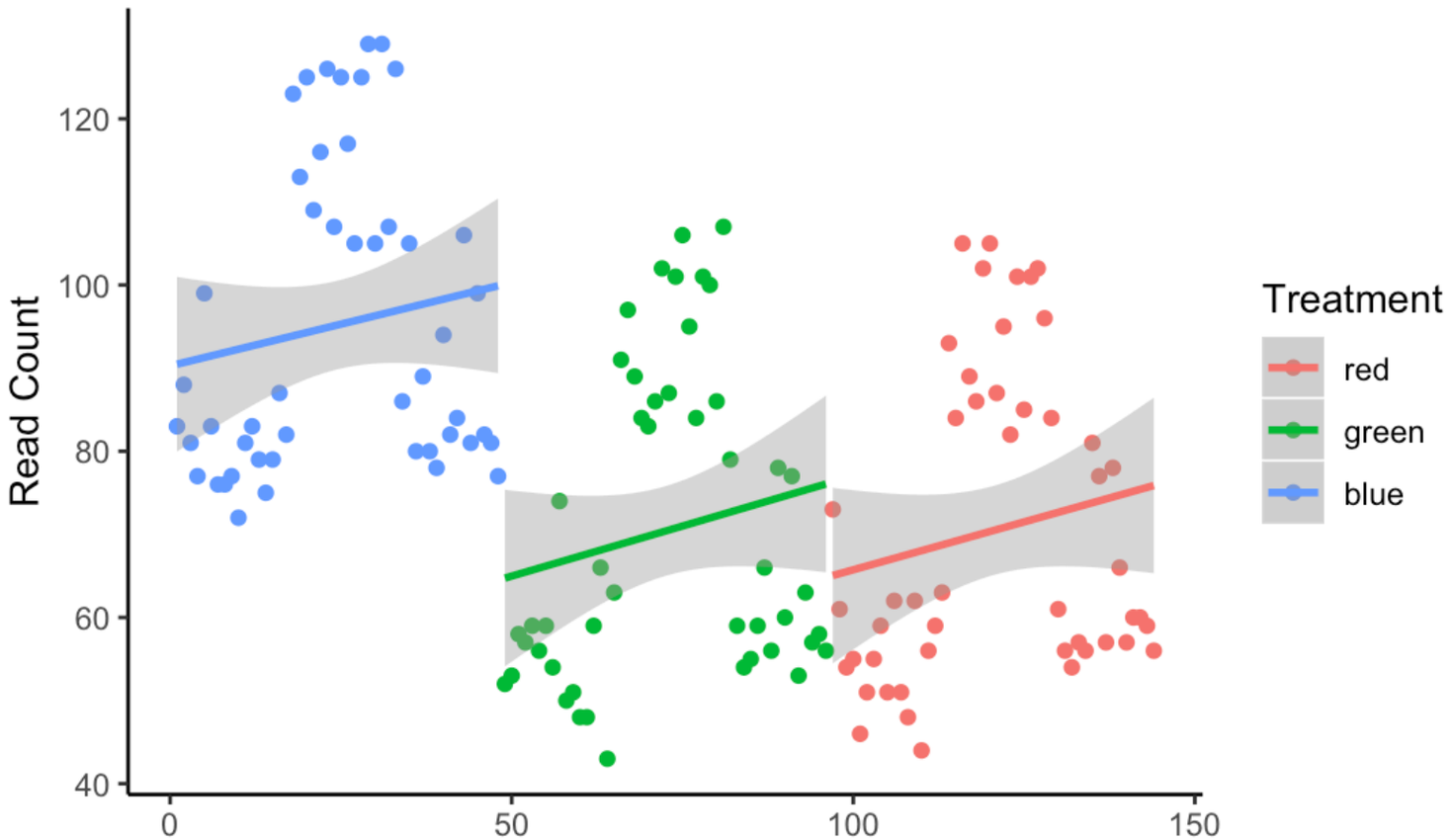
ReadCount ~ Treatment + Student

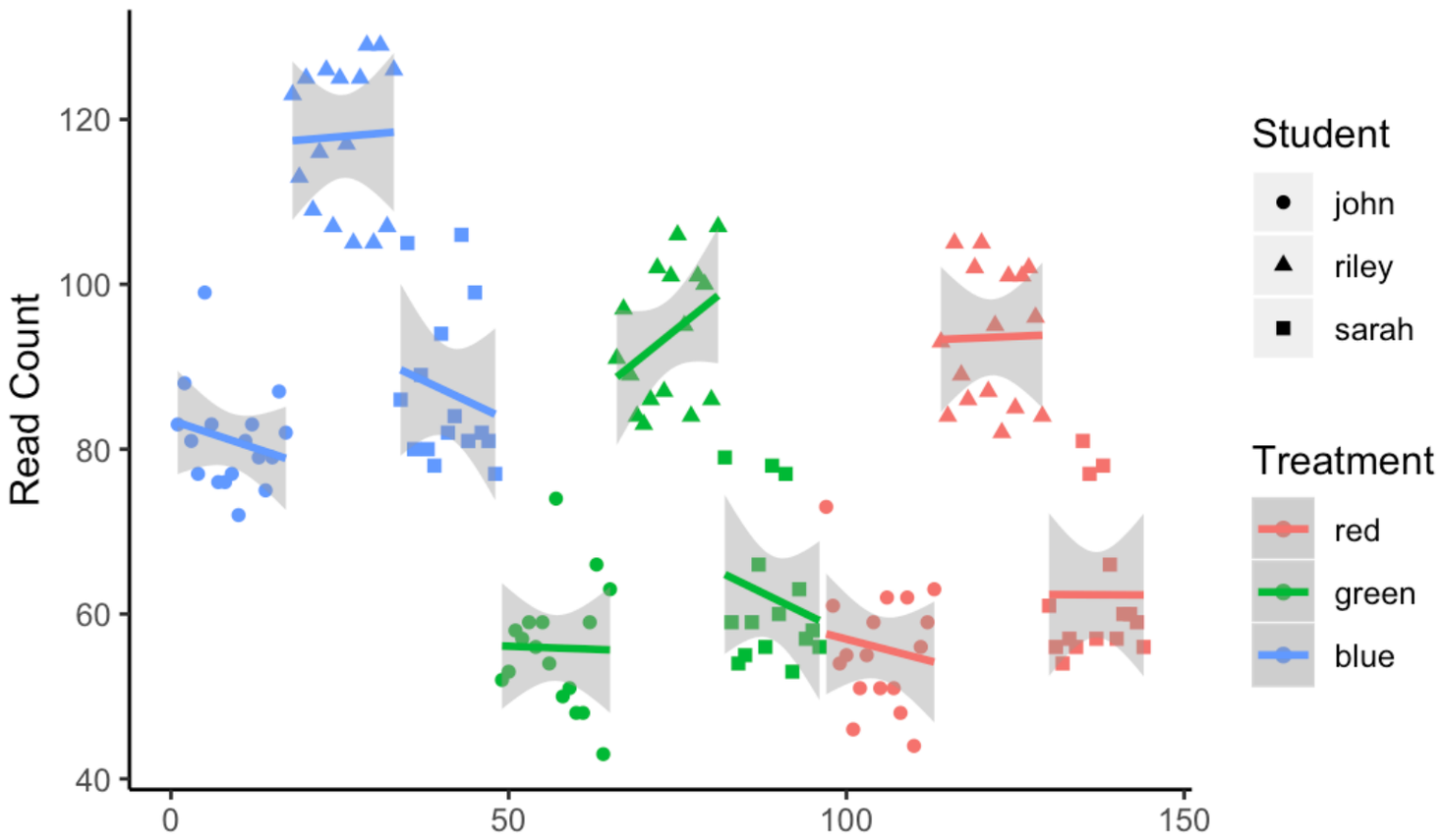
Sample	Treatment	Place	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18

ReadCount ~ Treatment + Student



What is happening to my *data*?

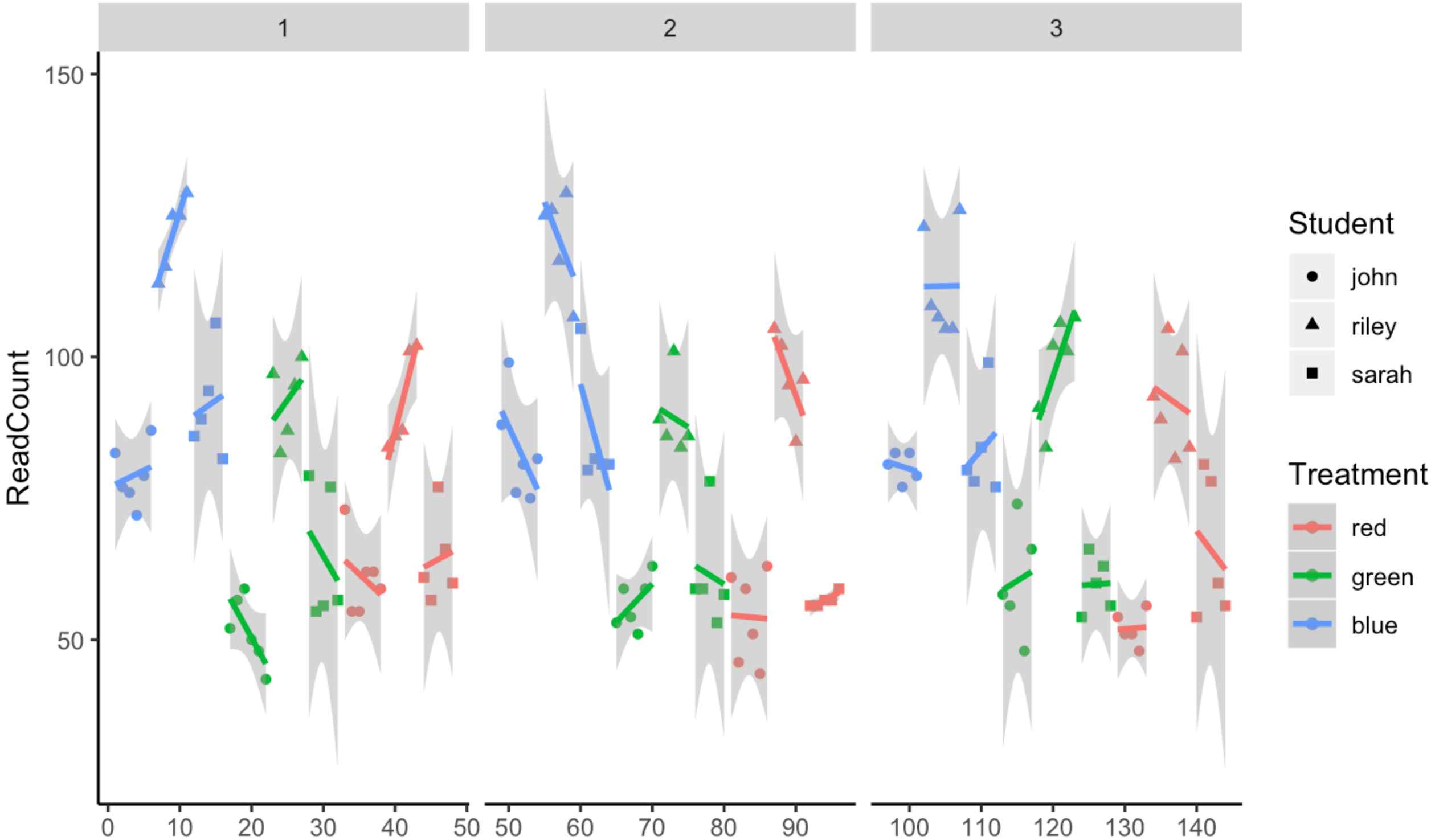

Read Count
120
100
80
60
40
0
50
100
150
Treatment
red
green
blue


Read Count
120
100
80
60
40
0
50
100
150
Student
john
riley
sarah
Treatment
red
green
blue

ReadCount ~ Treatment + Place + Student

Sample	Treatment	Place	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18

ReadCount ~ Treatment + Place + Student



What does this mean for your experimental design?

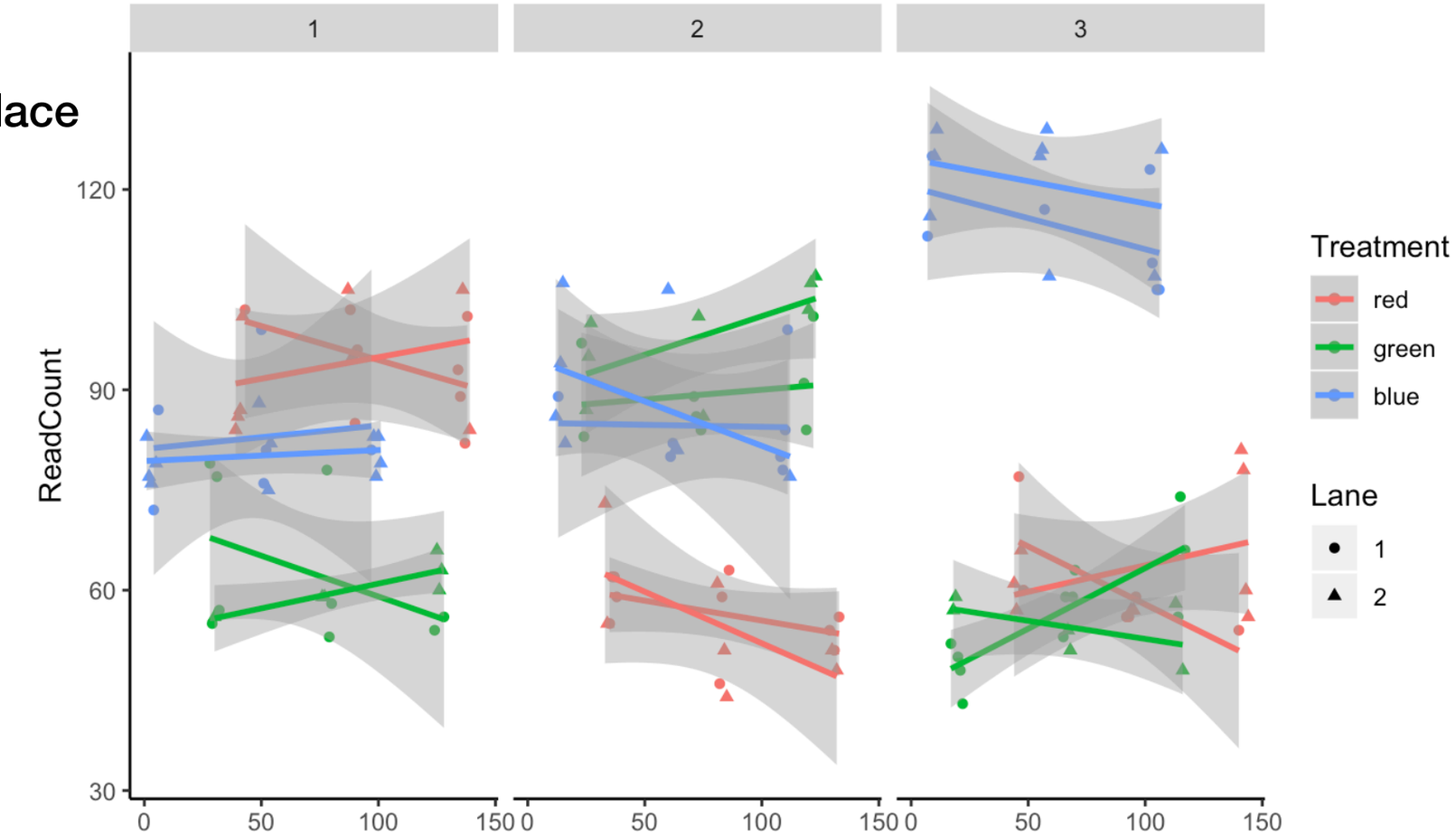
Proof we know math

- A ‘best fit’ line on a scatter plot is showing us the tendency of the dots…where their middle is, and what their slope is. It is the *average* of all those dots.

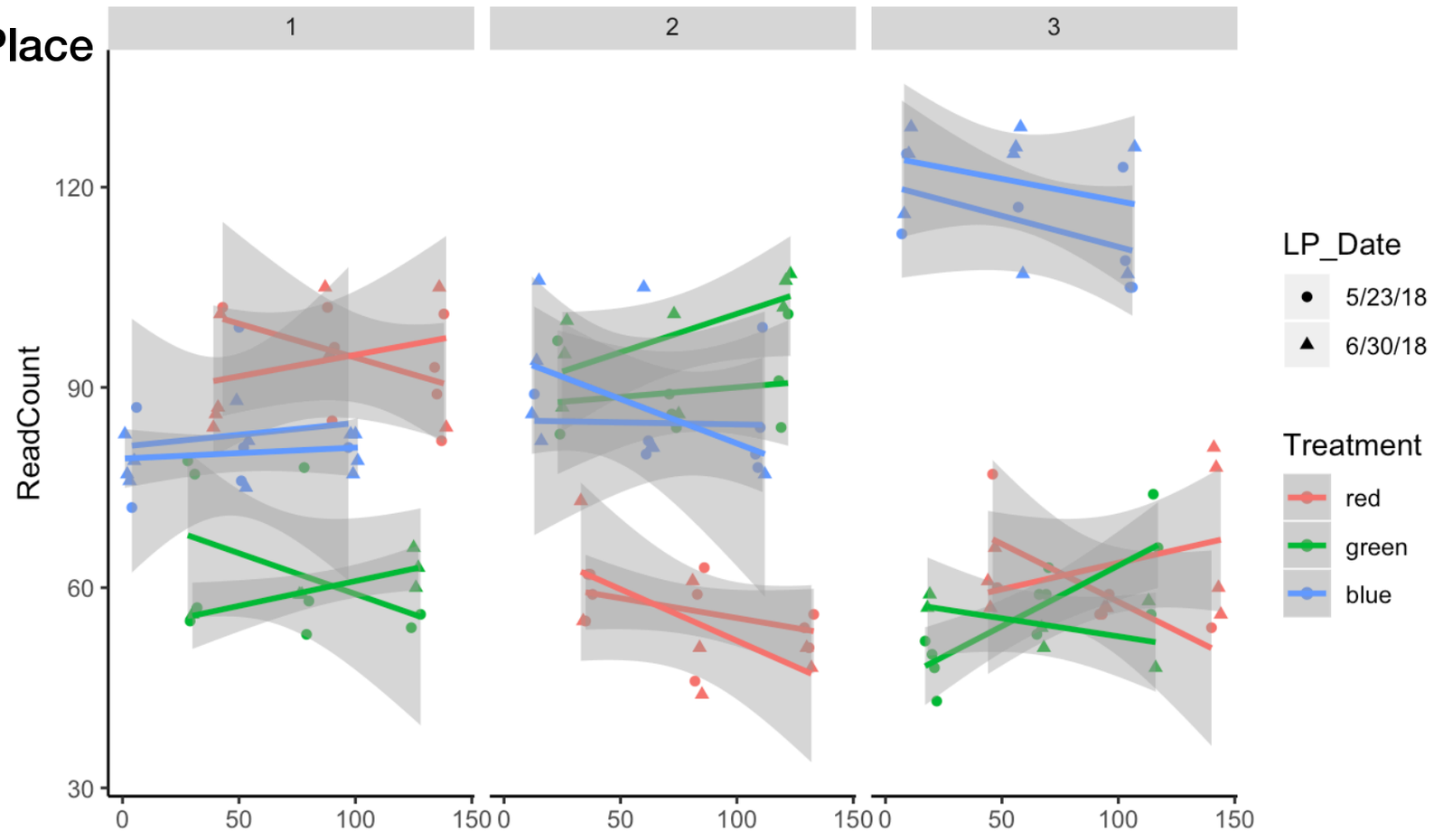
- Mathematical modeling is really just estimating averages for the right data subsets.

- A power analysis is figuring out how much data I need, to ask the question I care about, given the variables I’ll need to subset by. If I run out of points before I run out of subsets, I can’t answer my question.

ReadCount ~ Treatment + Lane + Place



ReadCount ~ Treatment + LP_Date + Place



Lane and LP_Date are perfectly confounded!

Sample	Treatment	Week	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18

Proof we know math

- A ‘best fit’ line on a scatter plot is showing us the tendency of the dots…where their middle is, and what their slope is. It is the *average* of all those dots.

- Mathematical modeling is really just estimating averages for the right data subsets.

- A power analysis is figuring out how much data I need, to ask the question I care about, given the variables I’ll need to subset by. If I run out of points before I run out of subsets, I can’t answer my question.

- My variables are confounded if they’re both averaging over the same points.

How many variables is ‘enough’?

Sample	Treatment	Week	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18

Which model should we use?