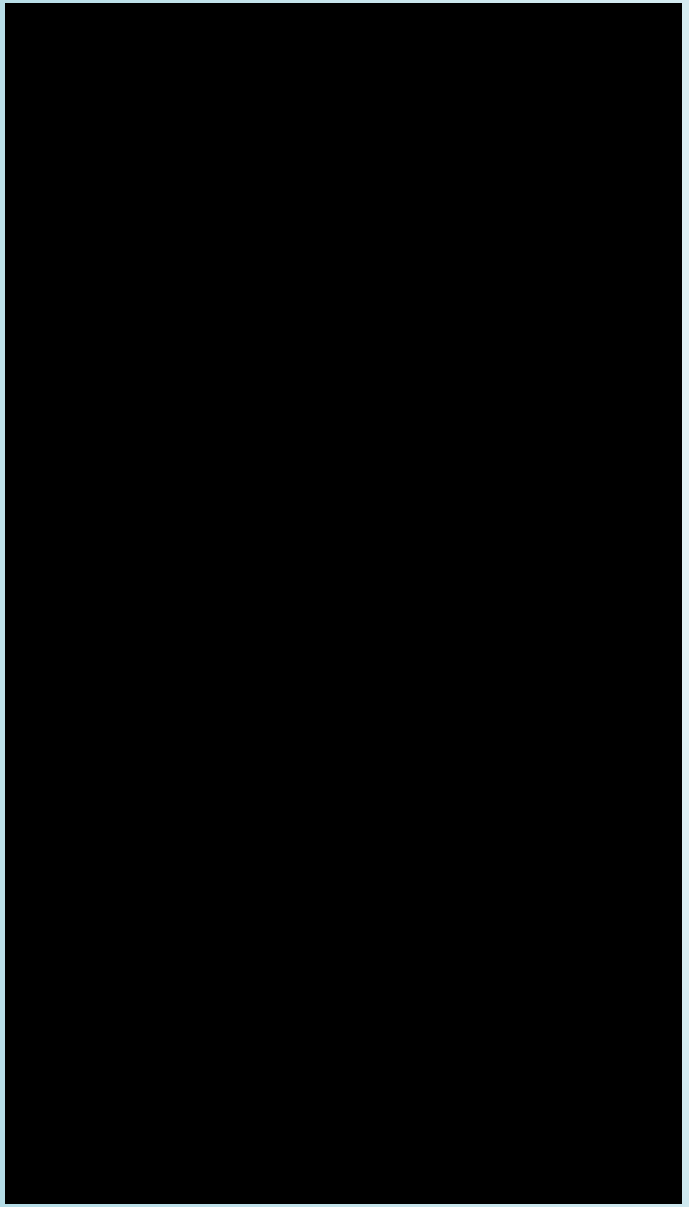
CAUSES OF COVARIATION

Michel Nivard

FLAMINGO'S



CAUSES OF COVARIATION

Today will cover ways to model the *genetic* covariance, and correlation, between two, or more traits.

This hour will cover:

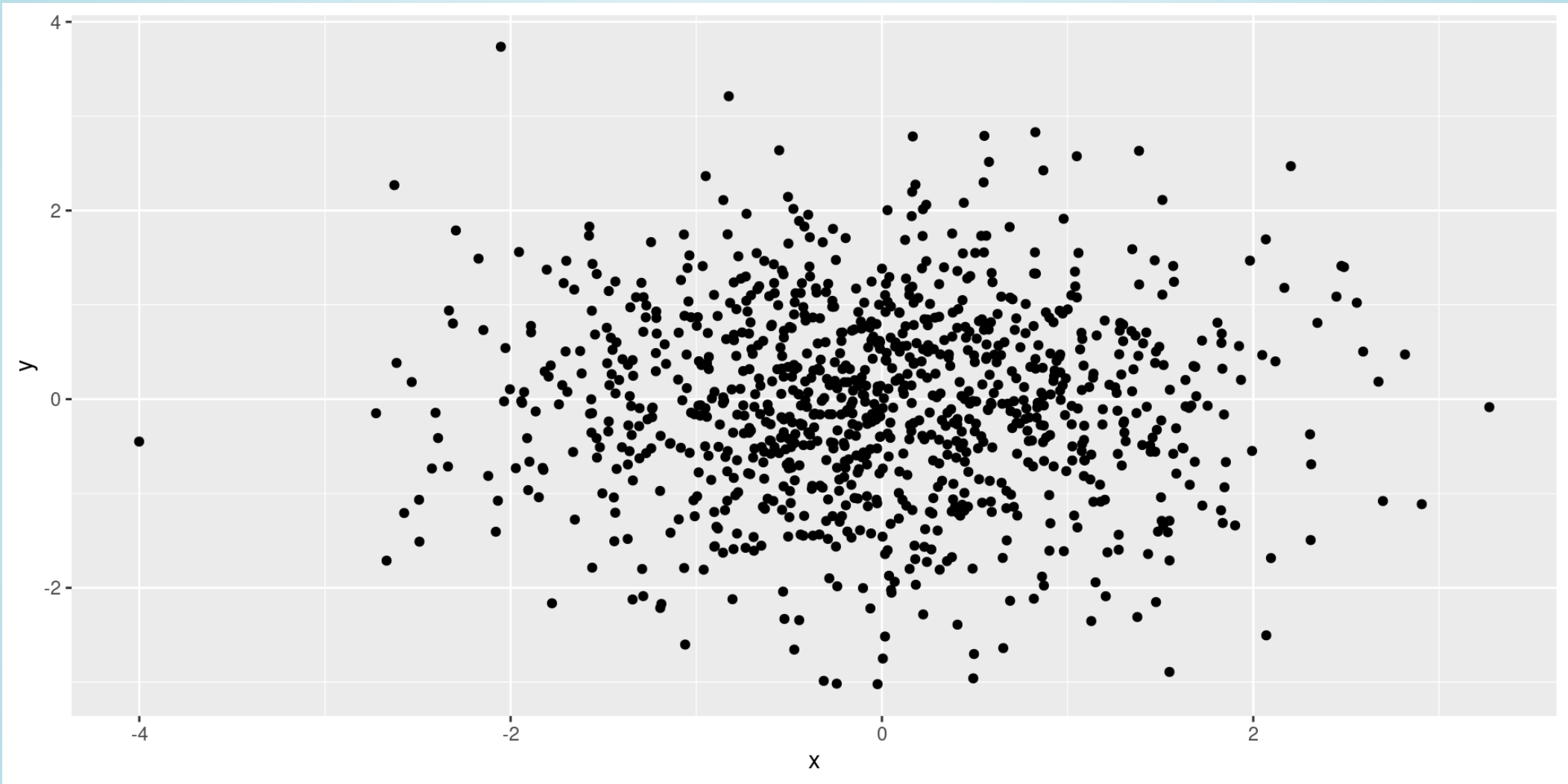
- What co correlation is (and isn't)
- How to relate what you want to know, to a statistical result

WHAT IS A CORRELATION?

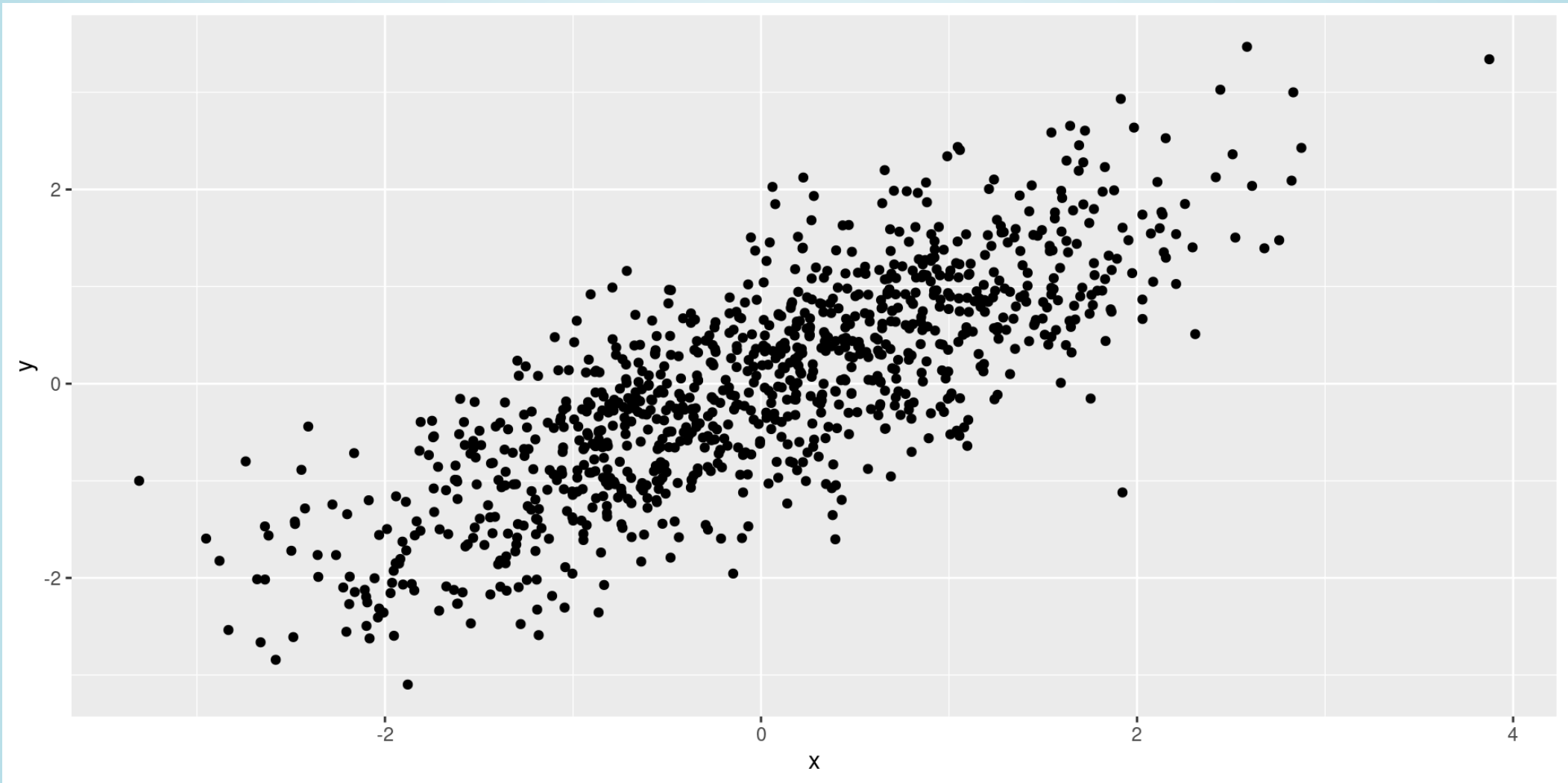
- a quantification of the degree to which two variables are linearly related
- correlation implies dependence
- dependence DOES NOT imply correlation

EXAMPLES OF DEPENDENCE VS CORRELATION

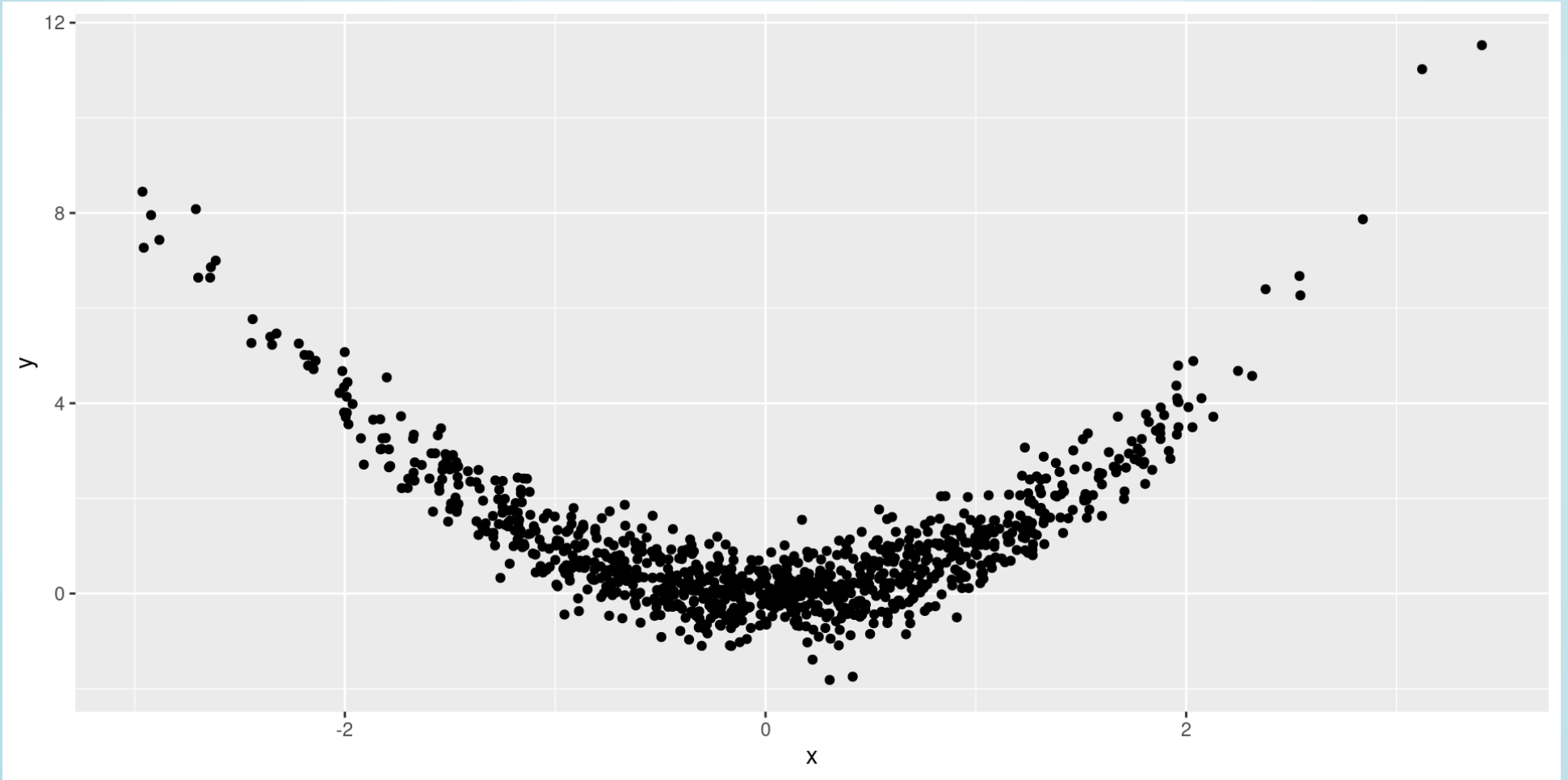
UNCORRELATED



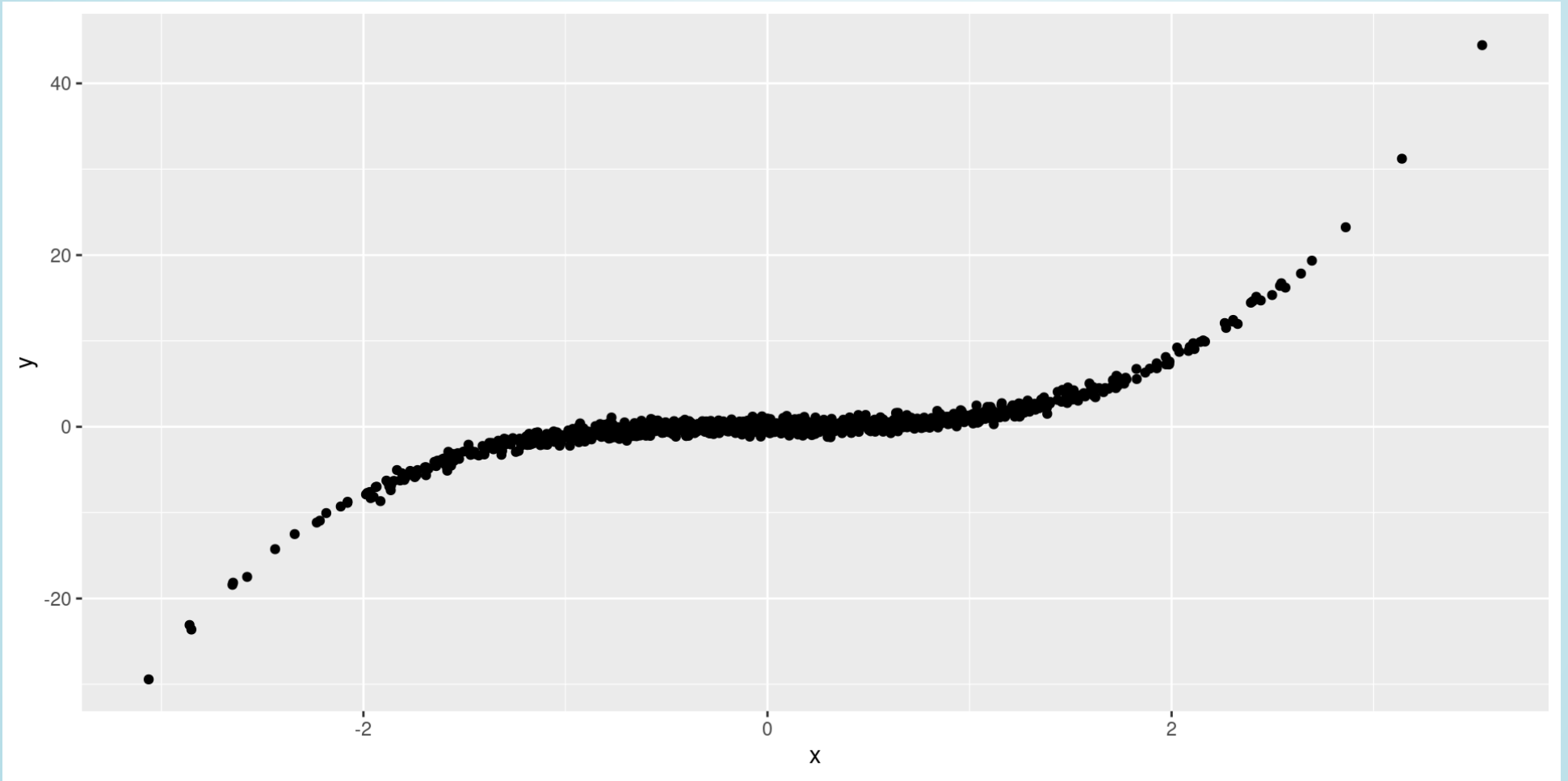
CORRELATED



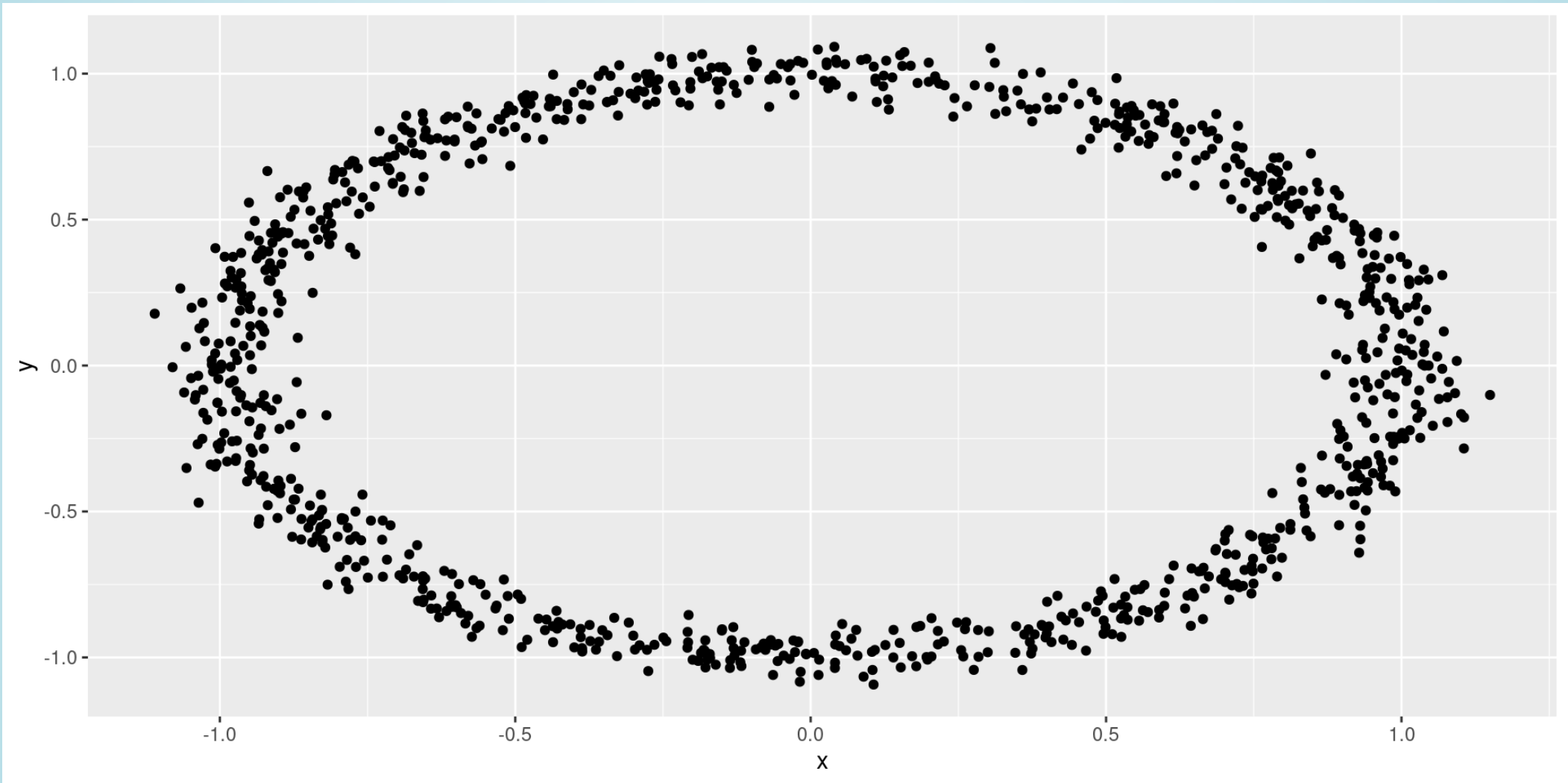
FUNCTIONALLY RELATED, BUT IS IT CORRELATED?



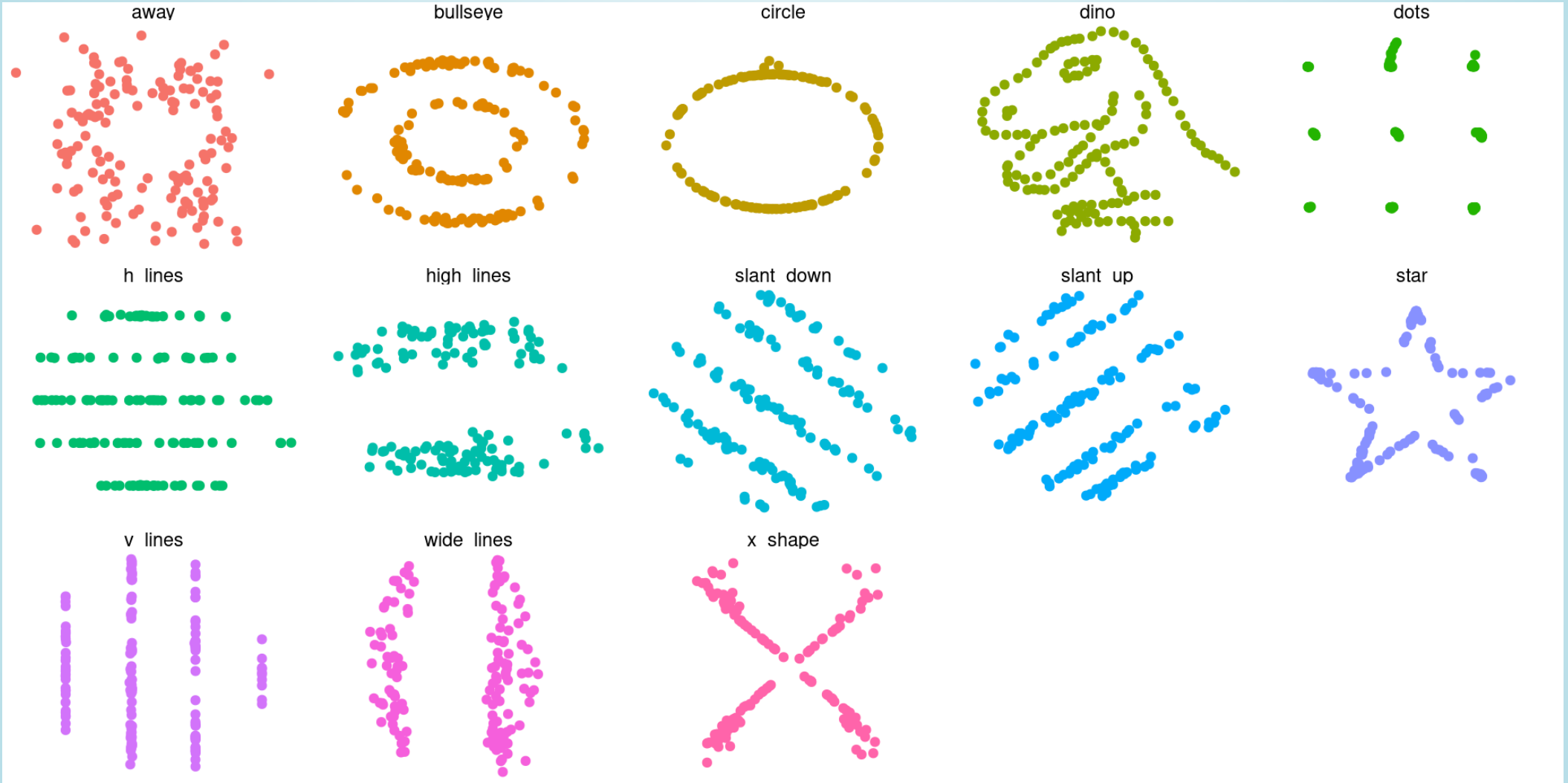
FUNCTIONALLY RELATED, BUT IS IT CORRELATED?



DEPENDENT, LIKELY UNCORRELATED...

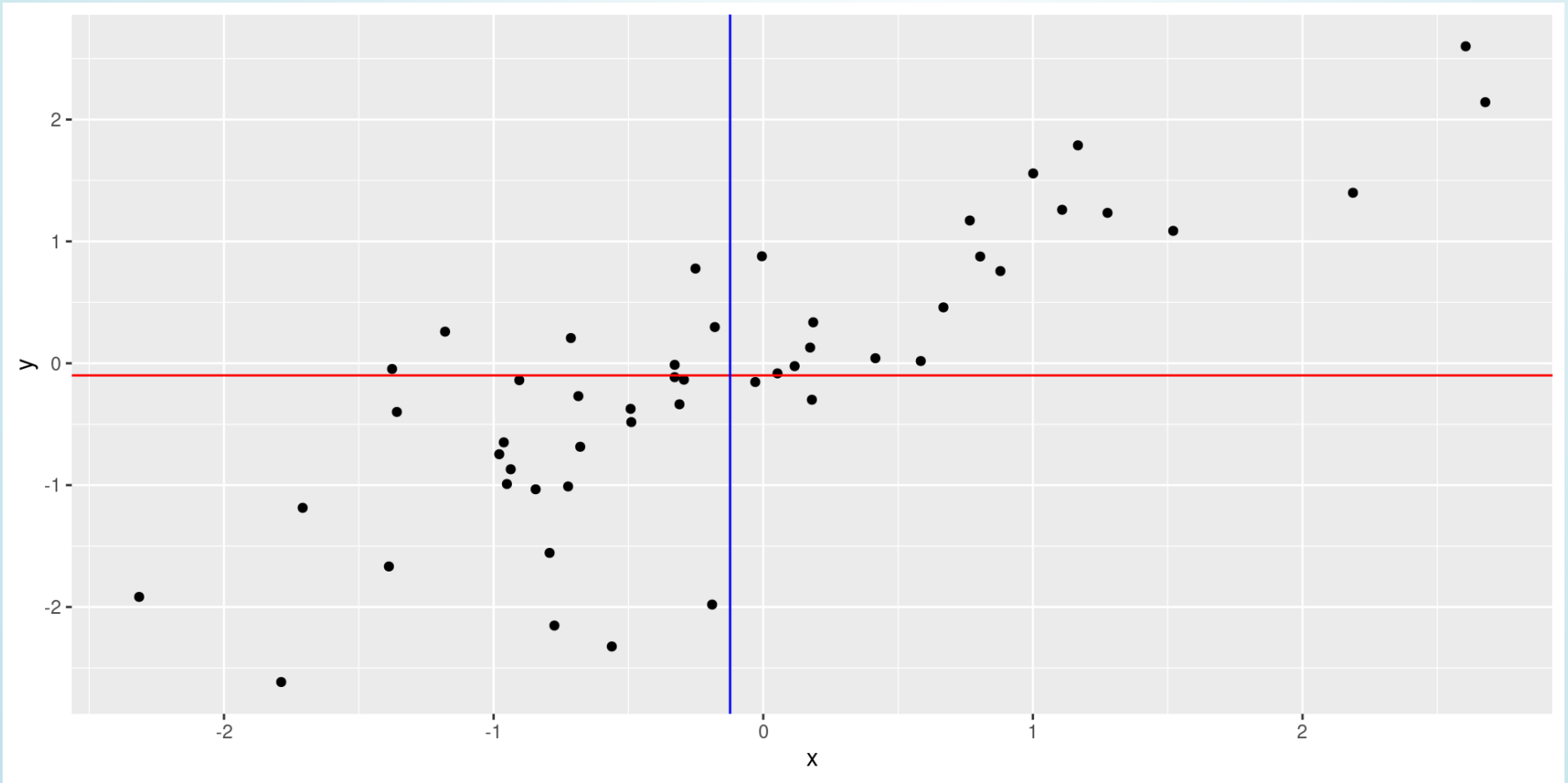


SCATTERPLOTS GO BRRR



A COMMON ESTIMATOR OF COVARIANCE

$$cov_{x,y} = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) * (y_i - \bar{y})}{N - 1}$$


2
1
0
-1
-2
y
-2
-1
0
1
2
x

A COMMON ESTIMATOR OF COVARIANCE

$$var_x = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) * (x_i - \bar{x})}{N - 1}$$

$$var_x = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{N - 1}$$

A COMMON ESTIMATOR OF CORELATIONS

$$cor_{x,y} = \frac{cov_{x,y}}{\sqrt{var_x * var_y}}$$

TWO DEFINITION OF GENETIC CORRELATION...

$$p1 = a1 + c1 + e1$$

$$p2 = a2 + c2 + e2$$

$$r_g = cor(a1, a2)$$

TWO DEFINITION OF GENETIC CORRELATION...

$$p1_i = \sum_{j=1}^{m} (b1_j * snp_j) + e_i$$

$$p2_i = \sum_{j=1}^{m} (b2_j * snp_j) + e_i$$

$$r_g = cor(b_1, b_2)$$

LETS PLAY A GAME!

correlation game

FROM RESEARCH QUESTION, TO STATISTICAL OUTPUT

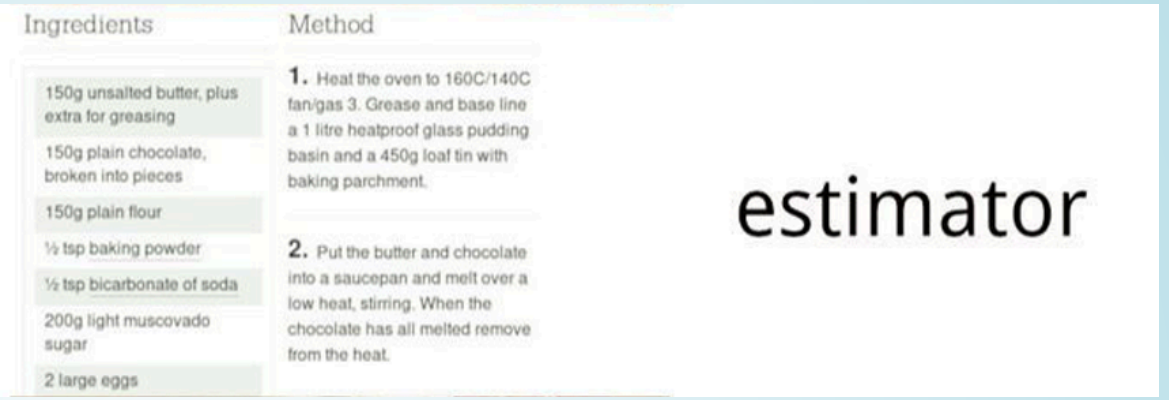
- How to relate what you want to know, to a statistical result?

WHAT IS IT THAT YOU WANT TO KNOW?

"are risk for depression and BMI genetically correlated?"

HOW WILL YOU GO AND FIND OUT?

“we will apply LD score regression to two sets of GWAS summary data form two different consortia, that studied BMI and MDD”



WHAT DID YOU FIND?

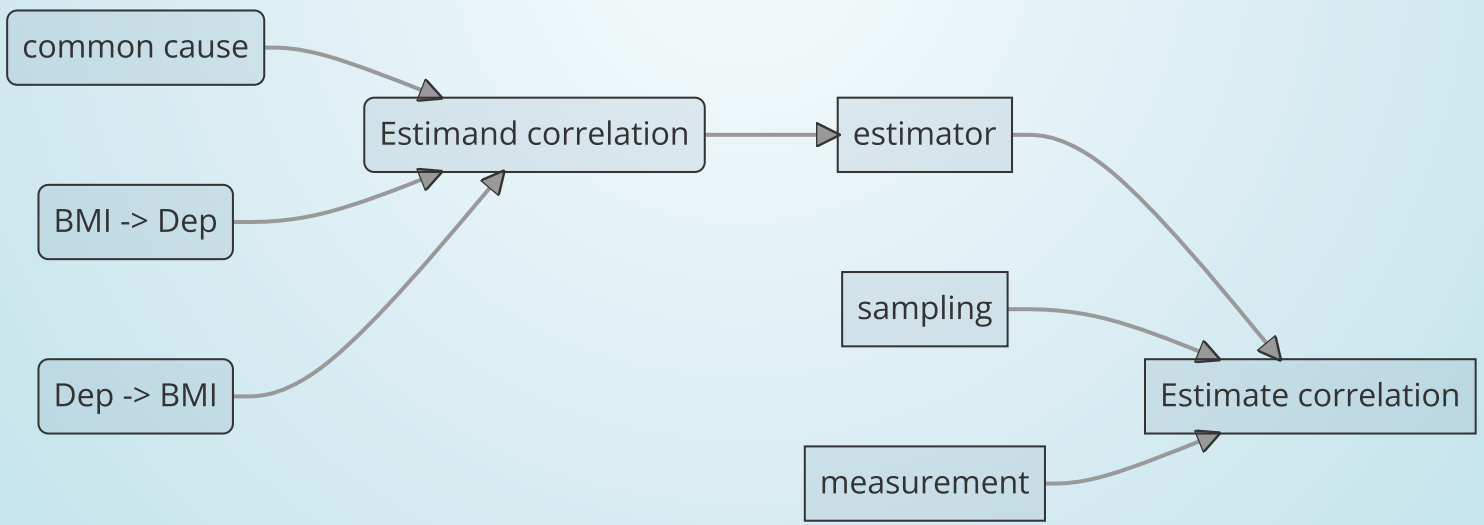
“The estimate of the genetic correlation between the PGC MDD, and GIANT BMI GWASs, Using LD score regression is 0.09”

LETS GO OVER THIS STEP BY STEP

An **estimand** is a quantity that is to be estimated in a statistical analysis. The term is used to distinguish the target of inference (**estimand**) from the method used to obtain an approximation of this target (i.e., **the estimator**) and the specific value obtained from a given method and dataset (i.e., **the estimate**).

(GENETIC) CORRELATION, ESTIMANDS AND ESTIMATE

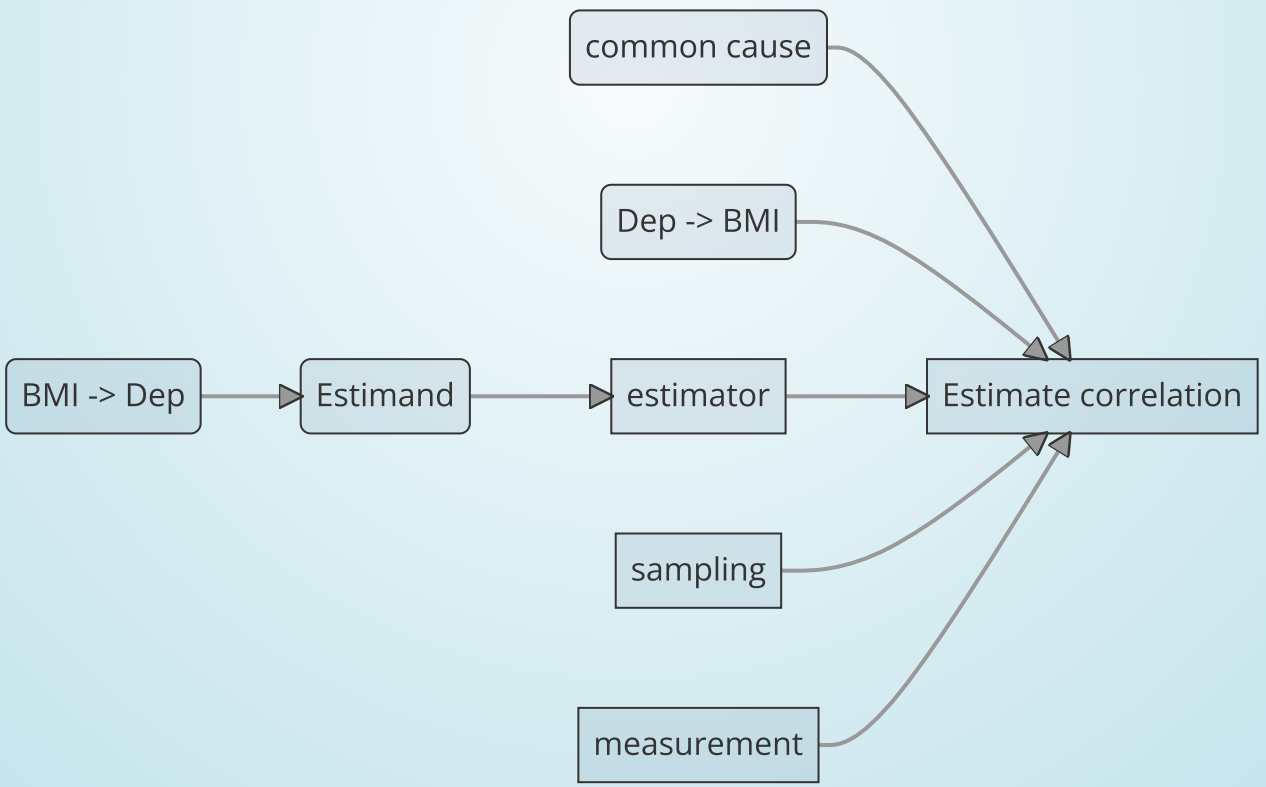
- We almost always want to know about processes that move the estimand
- The diagram below, **depends on your estimand!**



CORRELATION, ESTIMANDS AND ESTIMATE

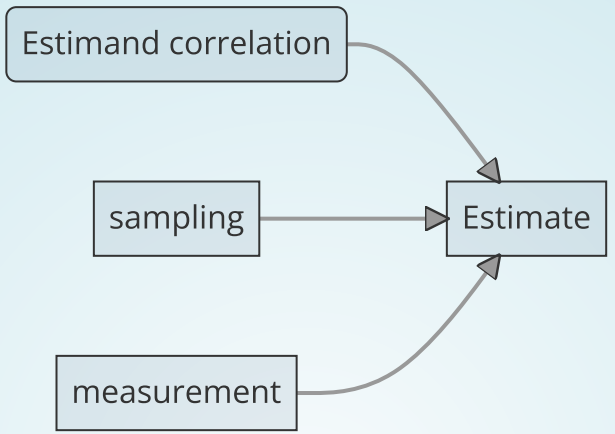
CAUSATION, ESTIMANDS AND ESTIMATE

- If we change the estimand, or estimator the diagram shifts!
- Estimand: "*The causal effect of BMI on Depression*"



CAUSATION, ESTIMANDS AND ESTIMATE

LETS LOOK AT SOME SPECIFIC CASES...



There are some very specific causes of correlation we need to discuss:

- ascertainment (and colider bias)
- measurement (and measurement error)

ASCERTAINMENT & MEASUREMENT

- The people in your study aren't always representative of the population (***sampling***)
- The measurement of your trait is not the same as your trait (***measurement***)
- These aspects of a study can arise by ***design***, or ***unintentionally***

ASCERTAINMENT BY DESIGN

- Over-sample cases in a schizophrenia GWAS (because its rare)
- Target a study at a specific populations with specific health needs
- You will need to adjust your estimator of h^2!!

UNINTENTIONAL ASCERTAINMENT (USUALLY SAMPLING)

- participants who social economic position is fragile might not have the time to spare for a day long lab study at a location that has poor access via public transport
- Elderly people might only respond to email if their 1. online 2. able too
- level of institutional trust may influence people's willingness to consent

UNINTENTIONAL ASCERTAINMENT (SAMPLING)

- Why would I care?
- It will bias all(!!) statistical estimates and inference
- There is a long causal chain between population and sample

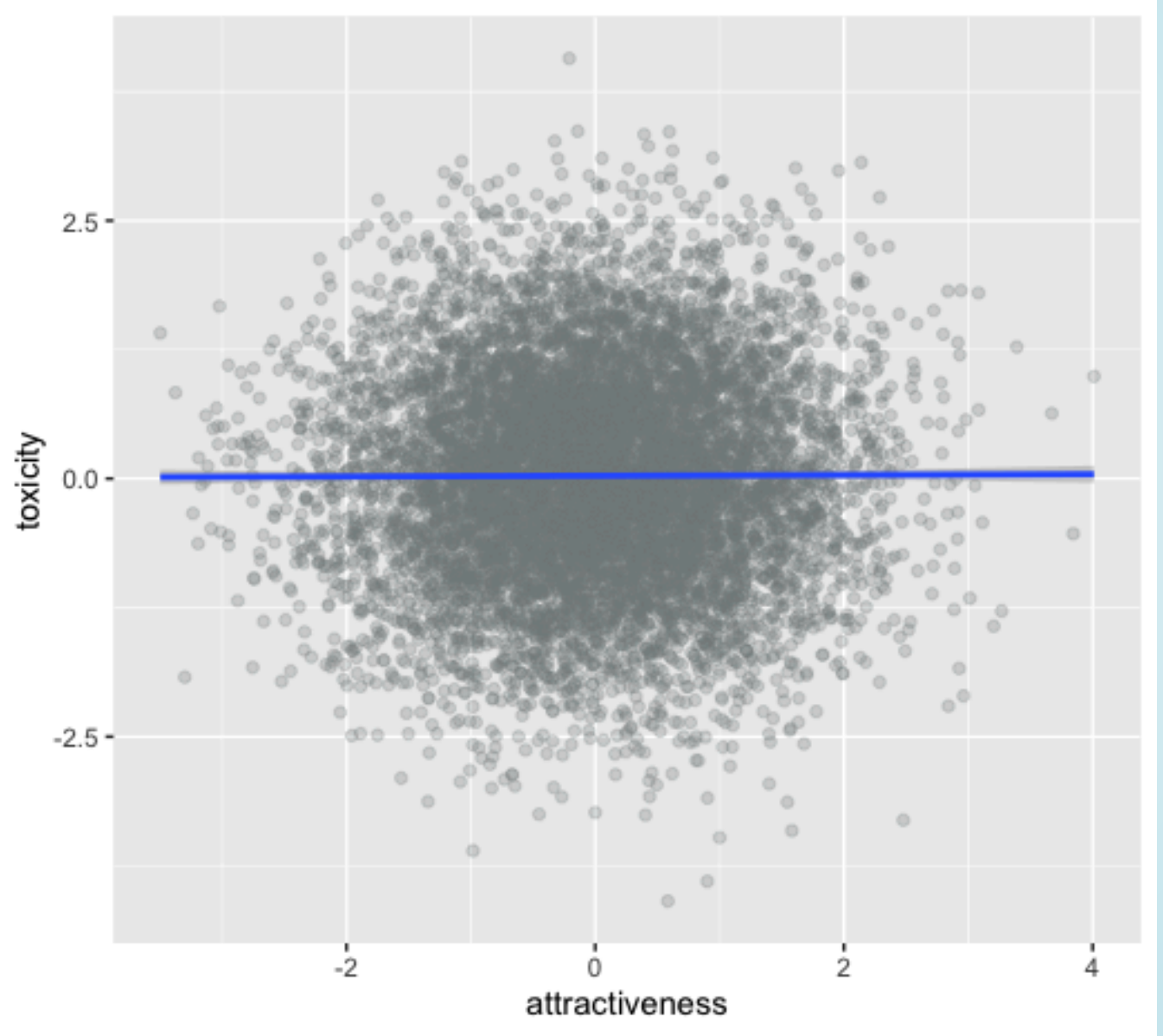
UNINTENTIONAL ASCERTAINMENT (SAMPLING): COLLIDER BIAS

- if: outcome1 -> ascertainment & outcome2 -> ascertainment
- in the ascertained sample outcome1 and outcome2 will correlate!

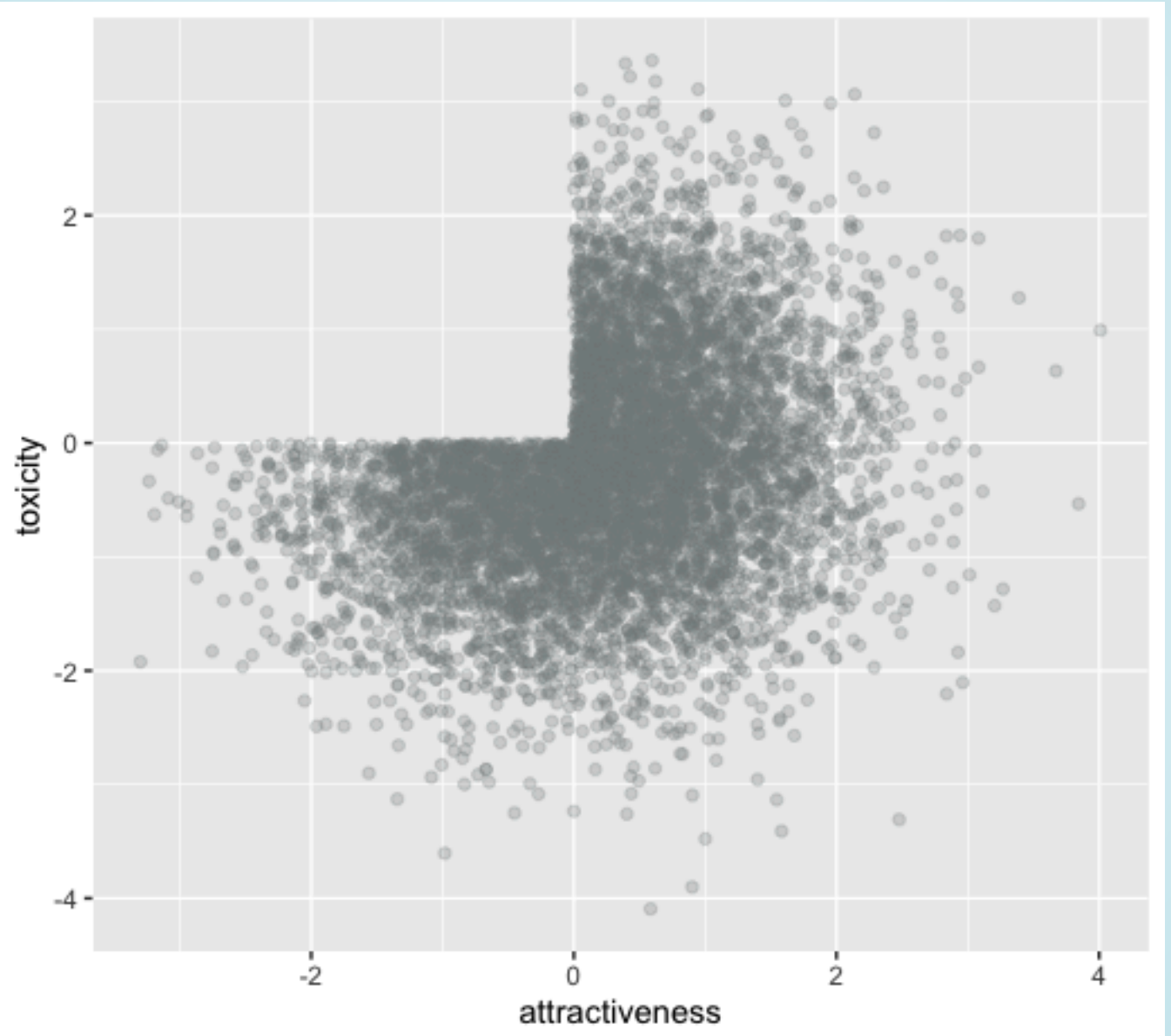
COLLIDER BIAS: DATING EXAMPLE

- Why do people feel their more attractive partners where also more toxic?
- Maybe its true? (maybe it effects my estimand)
- Or is it collider bias? (or it effects my estimate)

COLLIDER BIAS: DATING EXAMPLE

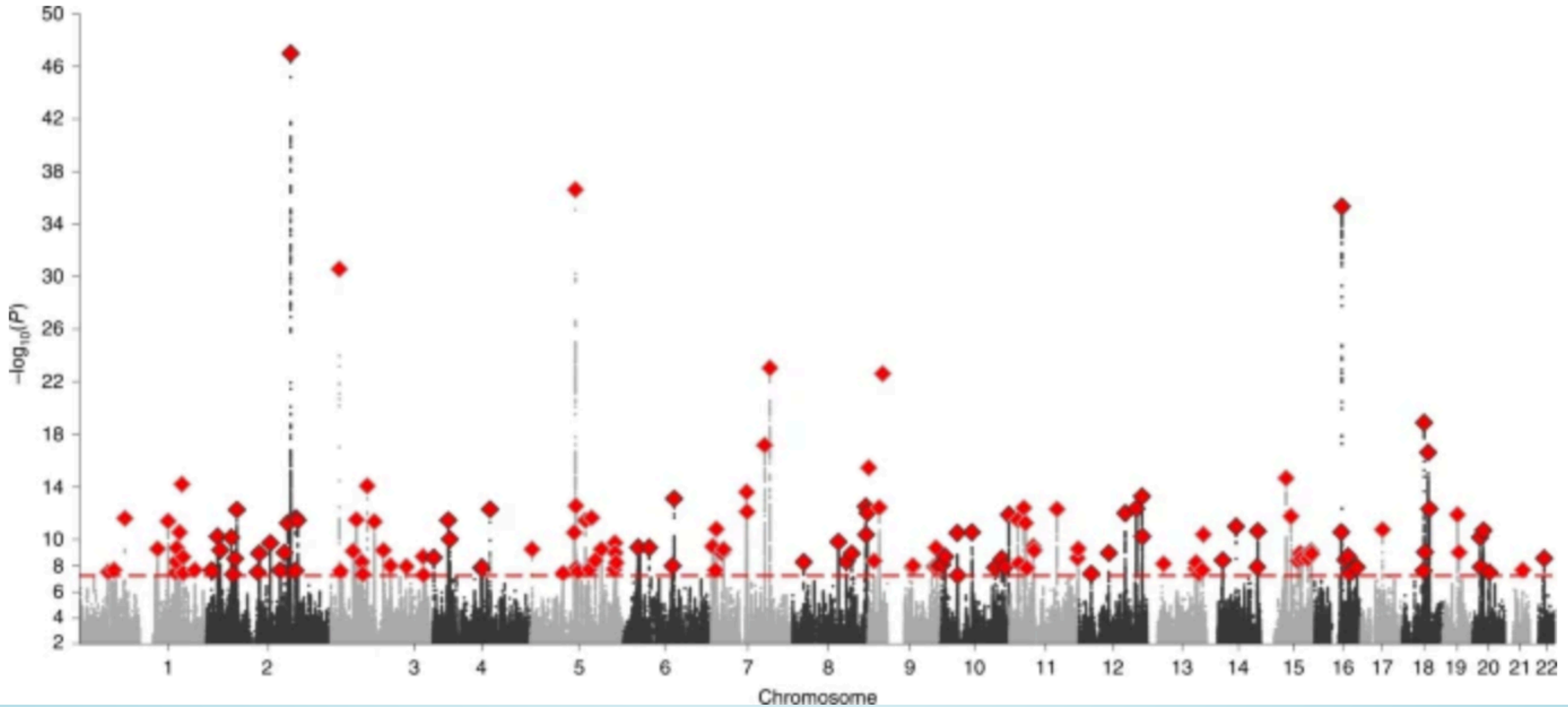


COLLIDER BIAS: DATING EXAMPLE



HOW COMMON IS THIS? SHOULD I CARE?

Fig. 1: Manhattan plot for a GWAS of sex in 2,462,132 participants from 23andMe.

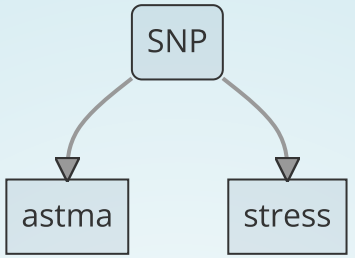


HOW COMMON IS THIS? SHOULD I CARE?

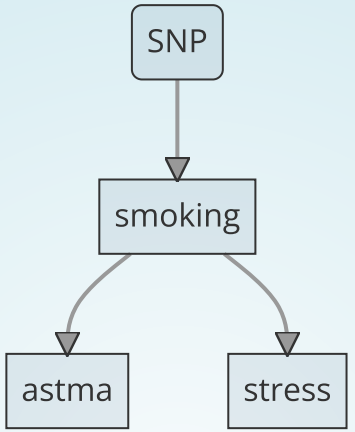
THE CAUSES OF A (GENETIC) CORRELATION THAT WE DO CARE ABOUT?

- (latent) common cause
- causal relation between two traits

A COMMON CAUSE

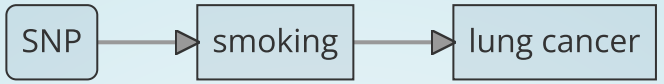


ALSO A COMMON CAUSE



A CAUSAL EFFECT

ALSO A CAUSAL EFFECT



TAKE HOME

- You have to consider what you want to know (estimand) carefully
- This will help you understand what your actually estimate means
- When analyzing the relation bertween two or more traits, consider all the causes of covariation!

GLANCE AT THE REST OF THE DAY:

- **Margot** will discuss estimating genetic correlation between two traits, using twin/family data.
- **Brad** will discuss models for the genetic correlations between more than 2 traits in family data
- **I** will discuss estimators of genetic correlation based on GWAS summary data (LDSC/Genomic SEM)
- **Andrew** will discuss models for the genetic correlations between more than 2 traits based on GWAS summary data (LDSC/Genomic SEM)

BIVARIATE TWIN MODEL WITH MARGOT

$$p1 = a1 + c1 + e1$$

$$p2 = a2 + c2 + e2$$

$$r_g = cor(a1, a2)$$

BIVARIATE TWIN MODEL WITH MARGOT

$$Vp1_{mz} = Va1 + Vc1 + Ve1$$

$$cov(p1_{mz1}, p1_{mz2}) = Va1 + Vc1$$

BIVARIATE TWIN MODEL WITH MARGOT

$$Vp1 = Va1 + Vc1 + Ve1$$

$$Vp2 = Va2 + Vc2 + Ve2$$

$$cov(p1_{mz1}, p2_{mz2}) = Coc(a1, a2) + Cov(c1, c2)$$

BIVARIATE MOLECULAR MODEL WITH ME

- we can do a ry similar thing with GWAS summary data.

$$p1_i = \sum_{j=1}^{m} (b1_j * snp_j) + e_i$$

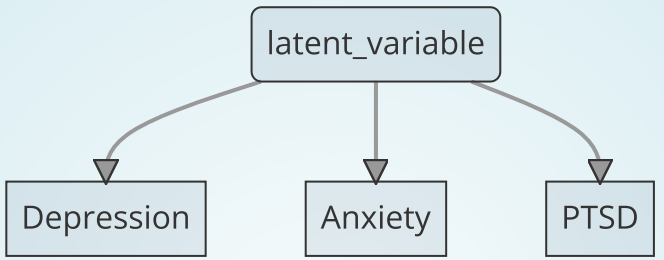
$$p2_i = \sum_{j=1}^{m} (b2_j * snp_j) + e_i$$

$$r_g = cor(b_1, b_2)$$

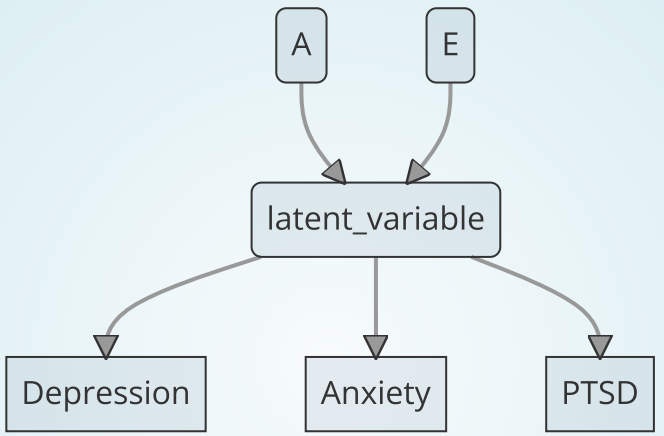
genetic correlations

- The bivariate twin model, and LDSC are complementary estimators of a similar quantity
- Its not an identical quantity(!)

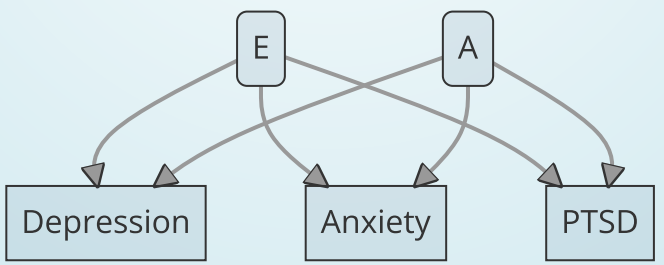
LATENT VARIABLE MODELS WITH BRAD & ANDREW



LATENT VARIABLE MODELS WITH BRAD & ANDREW



Or...



GENETICS IN THE CONTEXT OF GENETIC LATENT VARIABLE MODELING

- Brad and Andrew discuss complimentary estimators of genetic latent variable models
- The methods and code might look very different, various concepts are shared