Imperial College
Business School

Imperial means
Intelligent Business

Welcome to Quantitative Methods

by Ralf Martin (r.martin@imperial.ac.uk)

Course objective

- Equip you with basic data analysis tools used in economics and business.
- Give you practical skills including
  - Data handling
  - Programming
- Enable you to inform decisions with data

Why should you study for this course?

- Data is everywhere and as an consultant or analyst you need to be able to draw meaningful conclusions from data
- Even if you have people who will do the analysis for you it is vital to understand how conclusions are derived and how results can be interpreted.
- You most likely will have to deal with developers in your career. To do that effectively you need a basic grasp of coding.

Your lecturer

PhD in Economics from the London School of Economics

Name: Ralf Martin
Office: CAGB 487
Citizenship: EU

Pioneer in the UK of analyzing government business census data

Pioneer in using business census data to evaluate climate policy

Dad

Windsurfer

Your teaching assistant

Petra Sarapatkova
petra.sarapatkova09@imperial.ac.uk

Assessment

5% Participation (on the course forum after every lecture)

Always remember: Genuine Questions are never dumb

70% Final Exam

25% Final Group Coursework:

- Find question or issue that can be informed by data
- Find data (at 50 observations)
- Conduct analysis providing relevant evidence
- Discuss findings appropriately

Will provide further instructions soon

Basically we will check if your hands are dirty

The plan

Week	Topic	
Week 1	Introduction	
Week 2	Rrrr	Zoom tutorial
Week 3	Visions	
Week 4	Testing times	Zoom tutorial
Week 5	Multivariate Regressions	Hand in Group Coursework outline
	Break	
Week 6	Econometrics for Dummies	Zoom tutorial
Week 7	Instrumental Variables	
Week 8	Learning like a machine	Zoom tutorial
Week 9	Time for series	
Week 10	Loose Ends	Zoom tutorial

Guest lecturer (tbc)

Yves-Alexandre de Montjoye

www.demontjoye.com

Will tell you everything you ever wanted to know about machine learning

Introduction – Data Stories

Converting data into something beautiful

Introduction

- In Business and Economic decisions making we (ideally) need to base decisions on evidence

- Sometimes case studies & anecdotes

- More often: vast amounts of data
  - Consumer purchase decisions
  - Data for many different countries over long periods of time
  - Behaviour of workers: job search, unemployment, wages
  - Data on business outcomes

- To make sense of this there is (traditionally) a sub discipline called

Econometrics = Statistics + Economics

- Closely related: Biometrics, Data Science, Epidemiology

A secret:

What makes a good data analyst or econometrician?

- Mathematical skills?
- Programming skills?

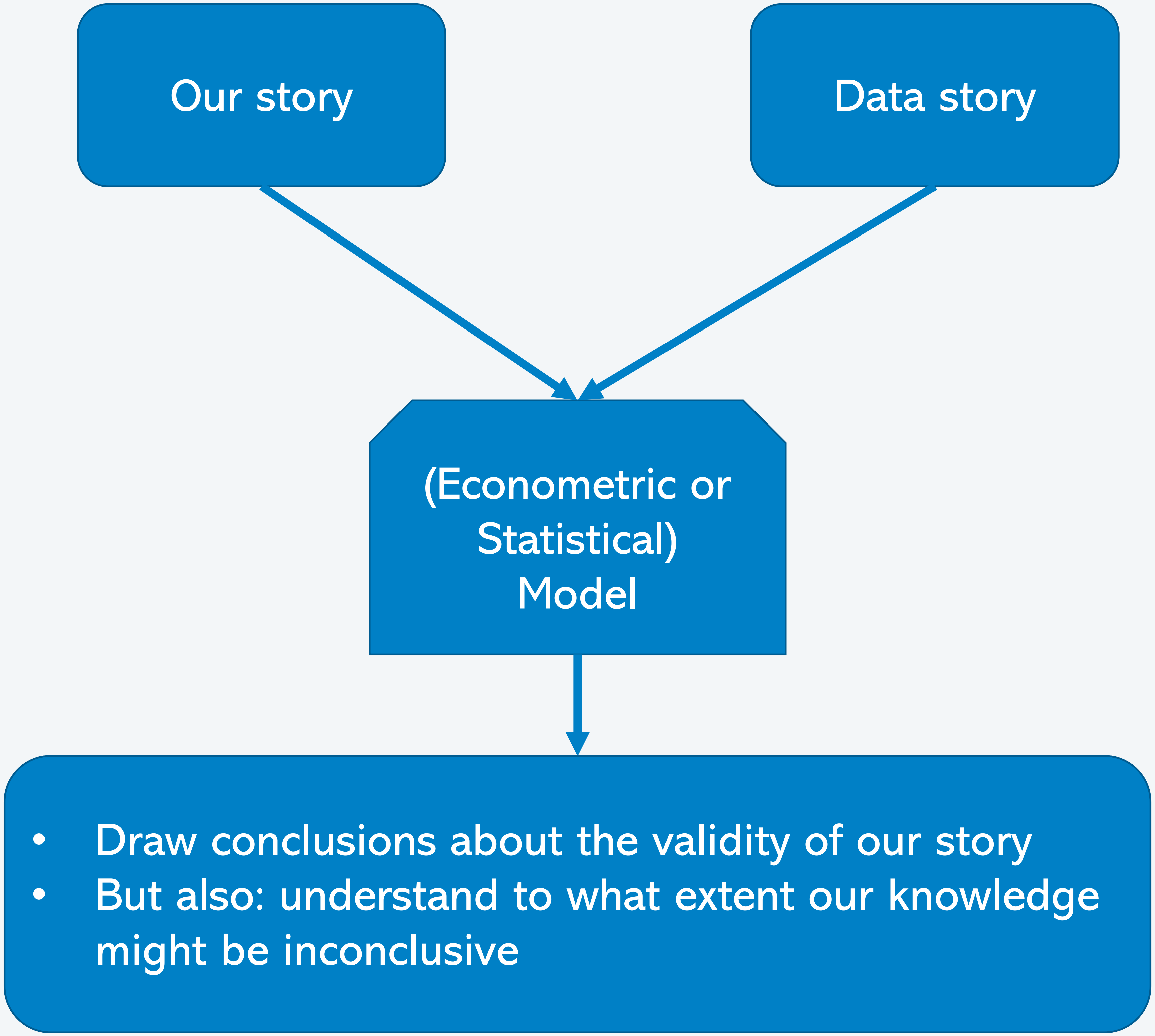
These are helpful, however a key ingredient is also to be an imaginative story teller.

This will help you to

1. Have a clear idea of the story you want to explore with data
2. Have an understanding of the story of the data

What are the potential drivers that could have produced your data?

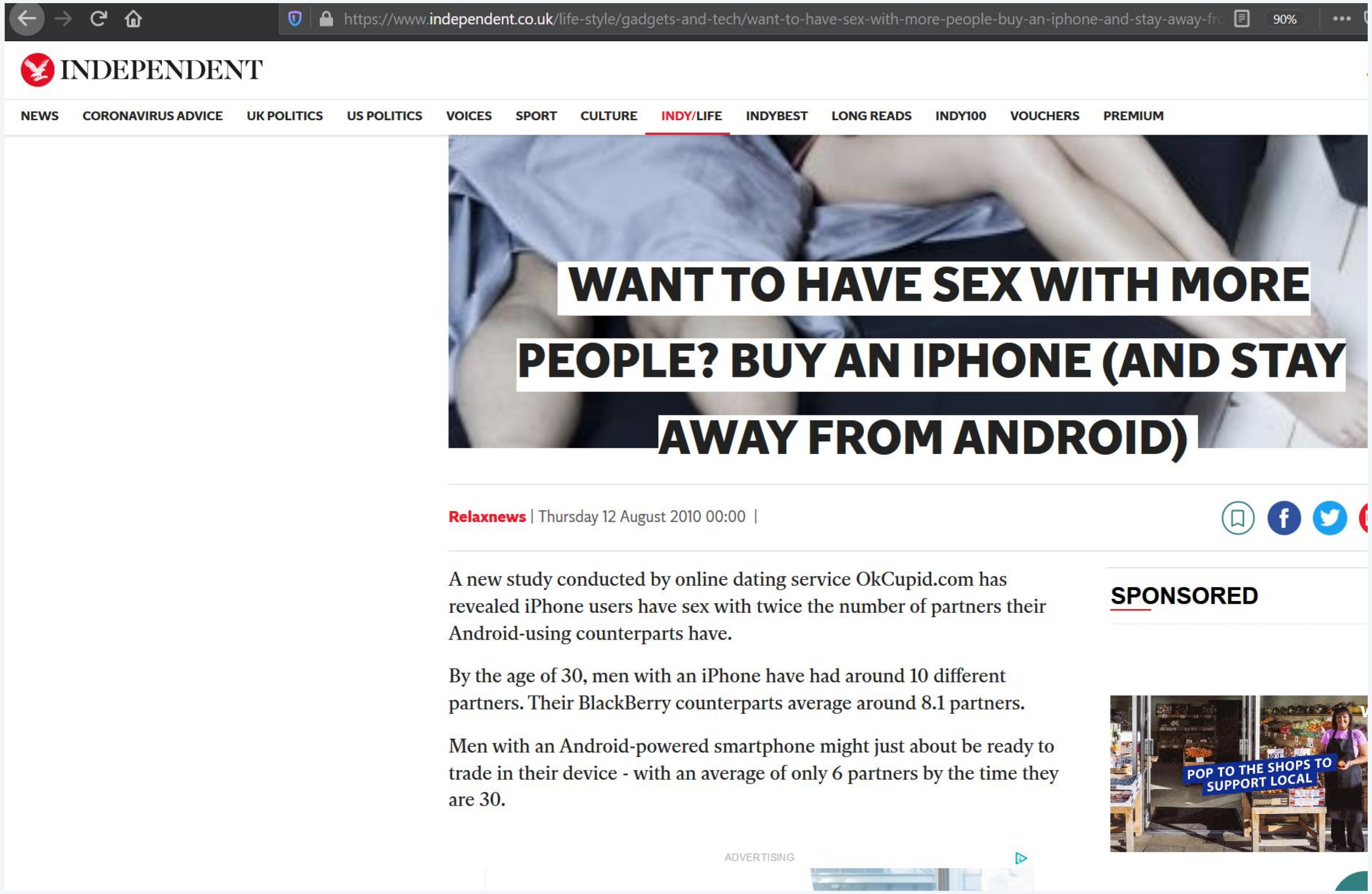
An important tool



What is a Model?

A verbal or mathematical description of relationships between things we can (potentially) measure

For example:



https://www.independent.co.uk/life-style/gadgets-and-tech/want-to-have-sex-with-more-people-buy-an-iphone-and-stay-away-fr

INDEPENDENT

NEWS | CORONAVIRUS ADVICE | UK POLITICS | US POLITICS | VOICES | SPORT | CULTURE | INDY/LIFE | INDYBEST | LONG READS | INDY100 | VOUCHERS | PREMIUM

WANT TO HAVE SEX WITH MORE PEOPLE? BUY AN IPHONE (AND STAY AWAY FROM ANDROID)

Relaxnews | Thursday 12 August 2010 00:00 |

A new study conducted by online dating service OkCupid.com has revealed iPhone users have sex with twice the number of partners their Android-using counterparts have.

By the age of 30, men with an iPhone have had around 10 different partners. Their BlackBerry counterparts average around 8.1 partners.

Men with an Android-powered smartphone might just about be ready to trade in their device - with an average of only 6 partners by the time they are 30.

ADVERTISING

SPONSORED

POP TO THE SHOPS TO SUPPORT LOCAL

Sex with 10 vs 6 people

What's the Independent's (our) story?

It's about **causal** (not casual) relationships

Say it with arrows (your first econometric model):

$$\text{iphone} \xrightarrow{+} \text{Amount of Sex}$$

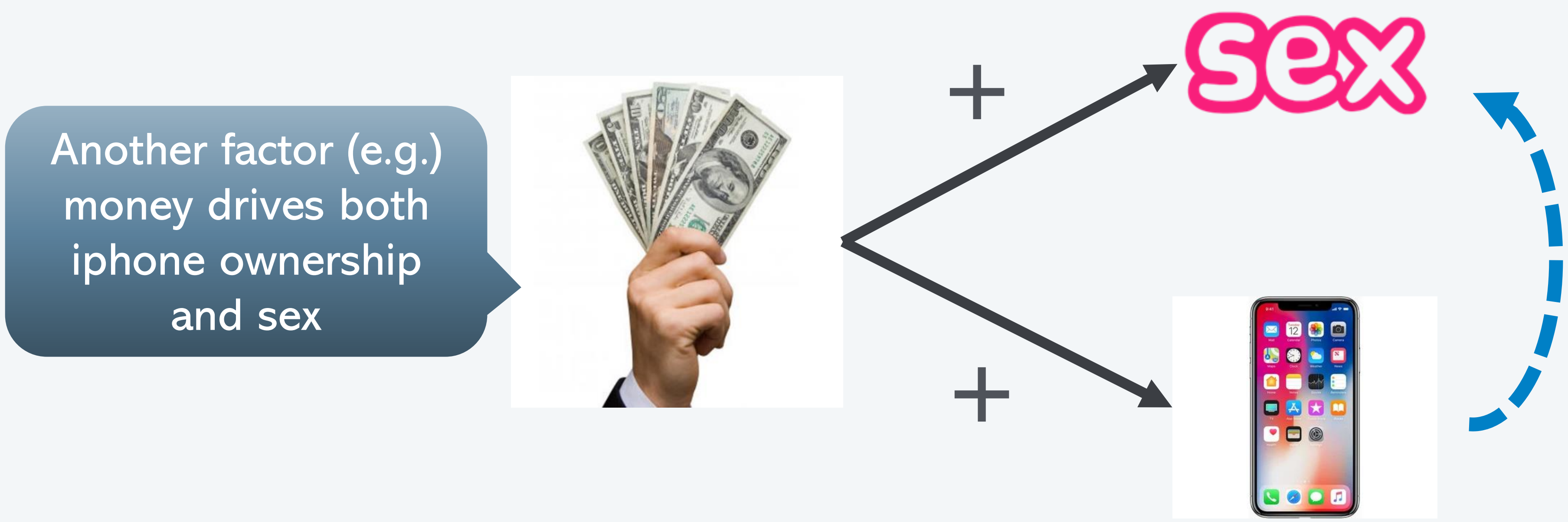
Using arrows we can express causation: iphone causes more (as indicated by the +) sex amount

sex

It also important to think of potential reasons for such a relationship; e.g. iphone signals wealth, style, taste ….? (Really?)

What could be the story of the data?

Your suggestions please?

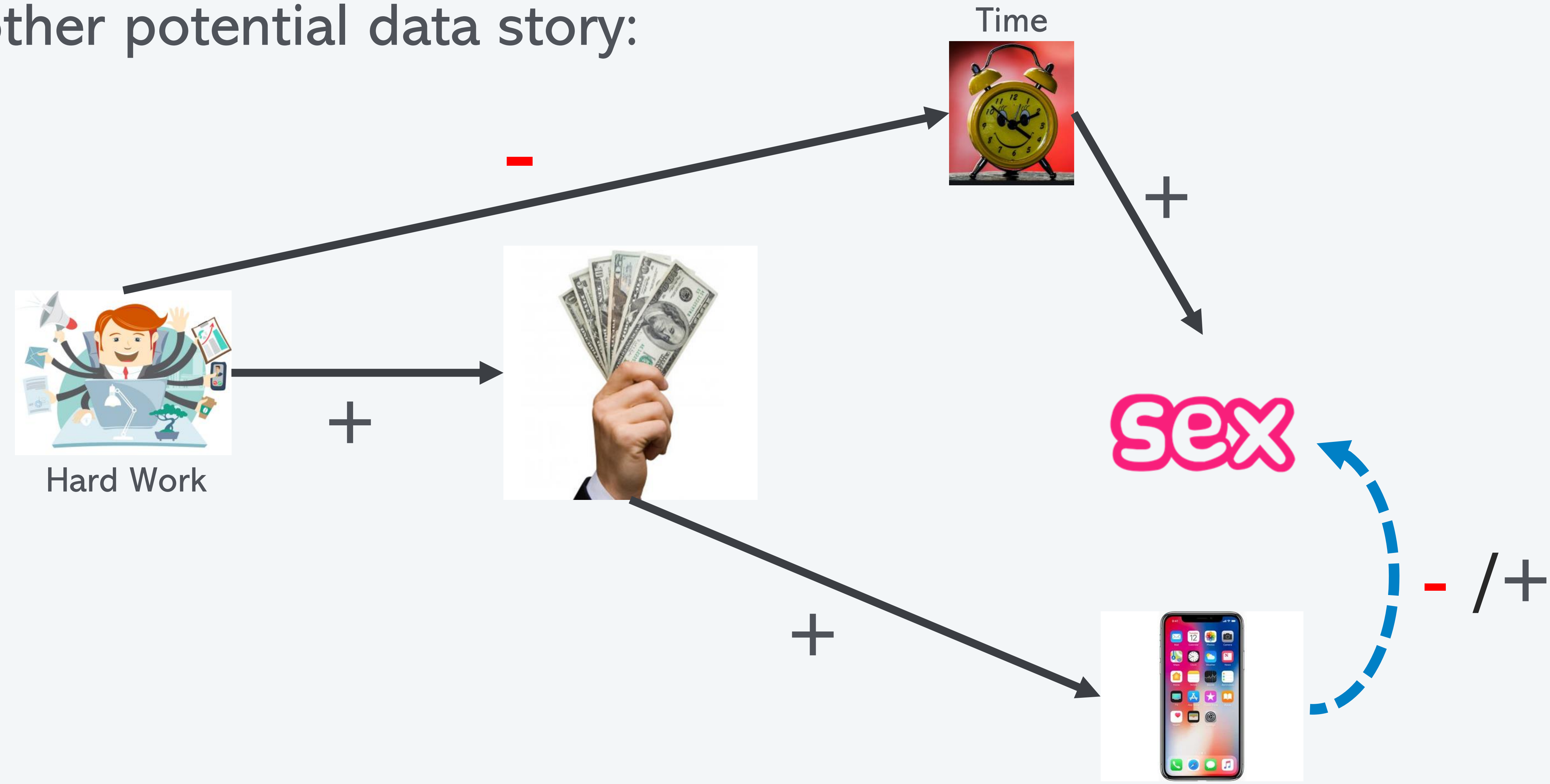


Note:

- Because Money is positively related to both sex and iphone ownerhsip we would see a positive correlation between the two in the data even if there is no causal effect from iphone to sex
- If there is a positive causal effect from iphone to sex, the money effect would make it seems stronger (we see the combined effect of money and iphone in the data)

Upward bias: Our estimated effect is larger than the true one

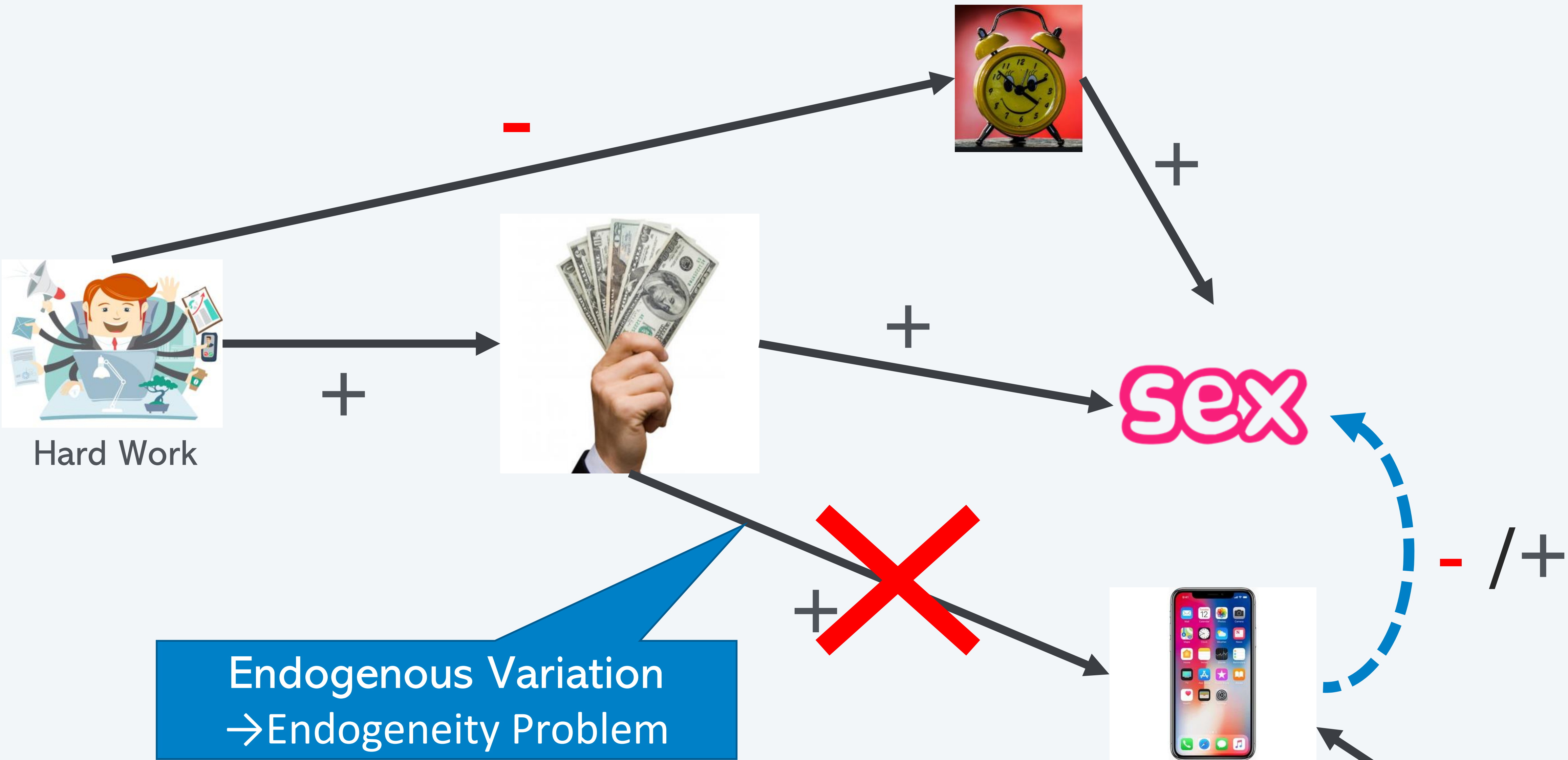
Another potential data story:



- (Good) Sex takes time
- To make money (some of us) have to work hard, which leaves us less time
- Hence, it is possible that people with more money have less time and less sex.
- If they also buy more iphones we could expect to find a dataset where iphone usage is negatively correlated with sex
- If we find a dataset with a positive correlation this would imply that the actual causal effect is even larger

Downard bias: Our estimated effect is smaller than the true one

What is the problem and what could be a solution?



- **Key Problem:** The explanatory variable of interest (iphone) is driven by factors (e.g. money) that exert their own causal effect on the outcome variable of interest (sex)

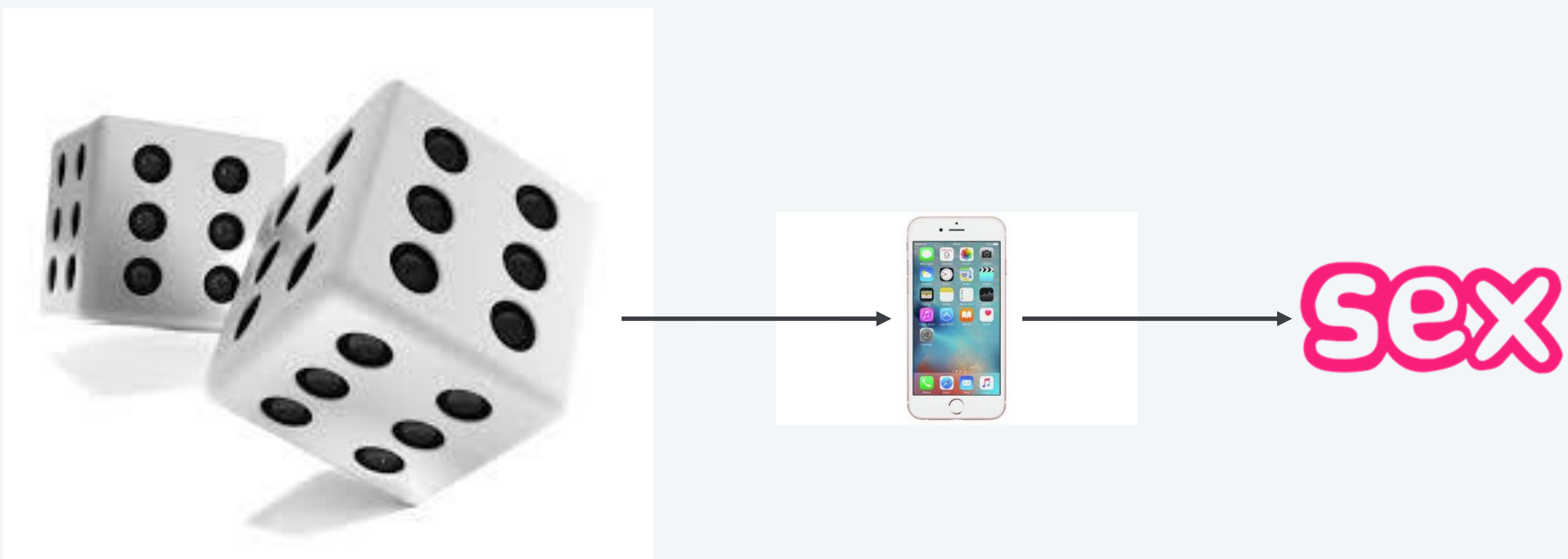
- **Hence, to get an <u>unbiased</u> estimate of the causal effect of iphones we need a dataset where variation in iphone ownership is not driven by such factors**

- **How?**

Instead we need Exogenous Variation

Randomized Control Trials (RCT)

- How about giving people iphone's at random to a "treatment group"?
- Ensures that iphone ownership is not drivey by money or other factors



The data story is aligned with "our" story so our conclusions are informative about the story of interest.

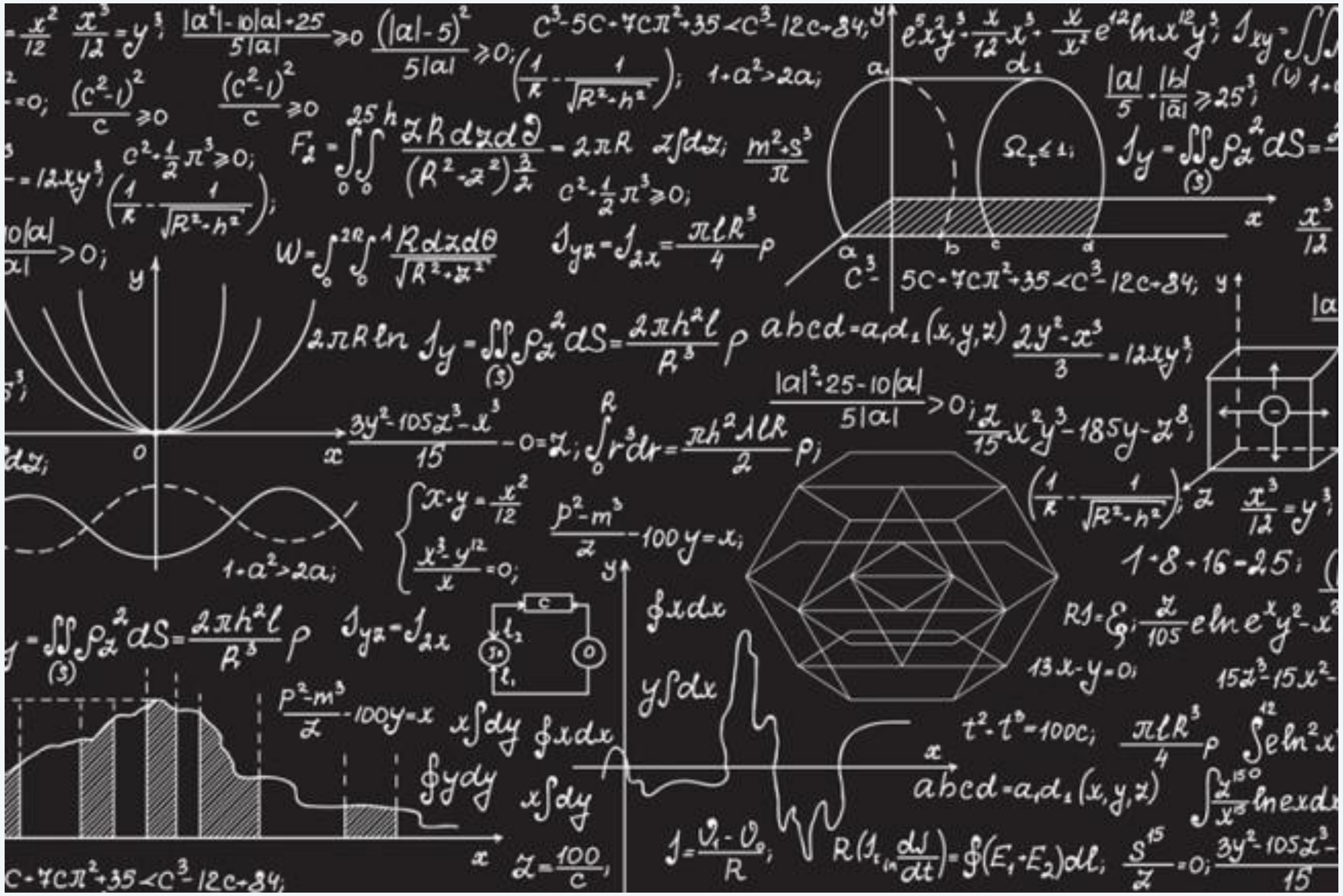
Randomized Control Trials (RCT) increasingly common in economic/business research

- Giving firms management practices
- Energy consumption feedback
- Fuel consumption feed to Airline pilots
- Giving students in Hong Kong incentives to protest
- Giving men in Saudi Arabia Information to induce them to let their wives work

But RCTs are not always feasible. What then?

Frome arrows to formulas

- Using Arrows and +/- signs to describe relationships of variables is one valid modelling approach **(also known as Directed Acyclic Graphs, DAG)**
- However, to come up with precise quantification of these relationships we need to start modelling with mathematical formulas

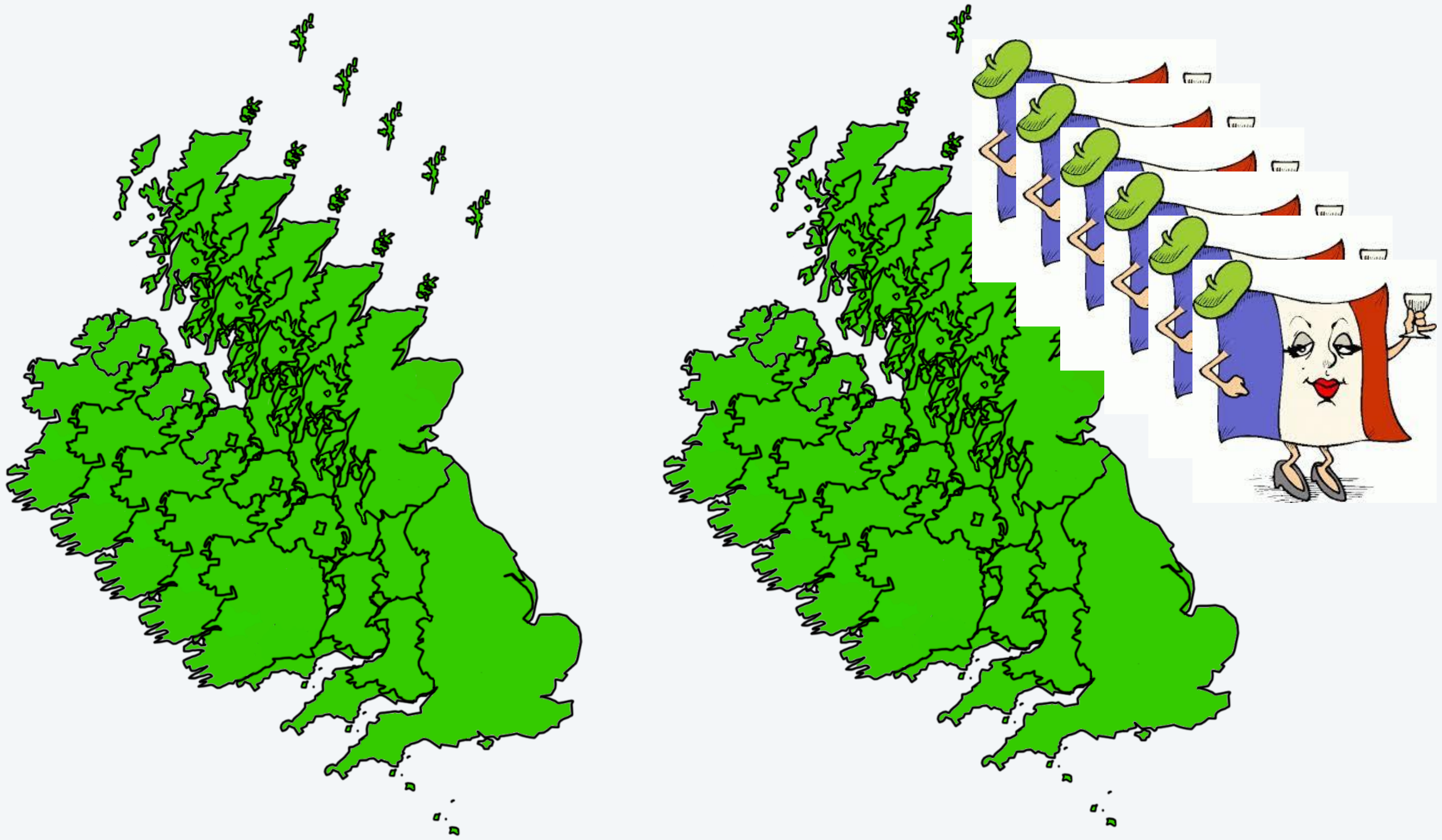


Another example: UK xenophobia in the UK

Since at least 2010 the UK government together with large parts of the media engaged in a virulent campaign of xenophobia and institutional discrimination against people without UK passports living in the UK. This involved blaming foreign born residents for all manner of things including crime.

Story: Foreigners cause crime

What data?



- Hard to do RCTs here
- Key to data is to have **variation**
- Two basic ways

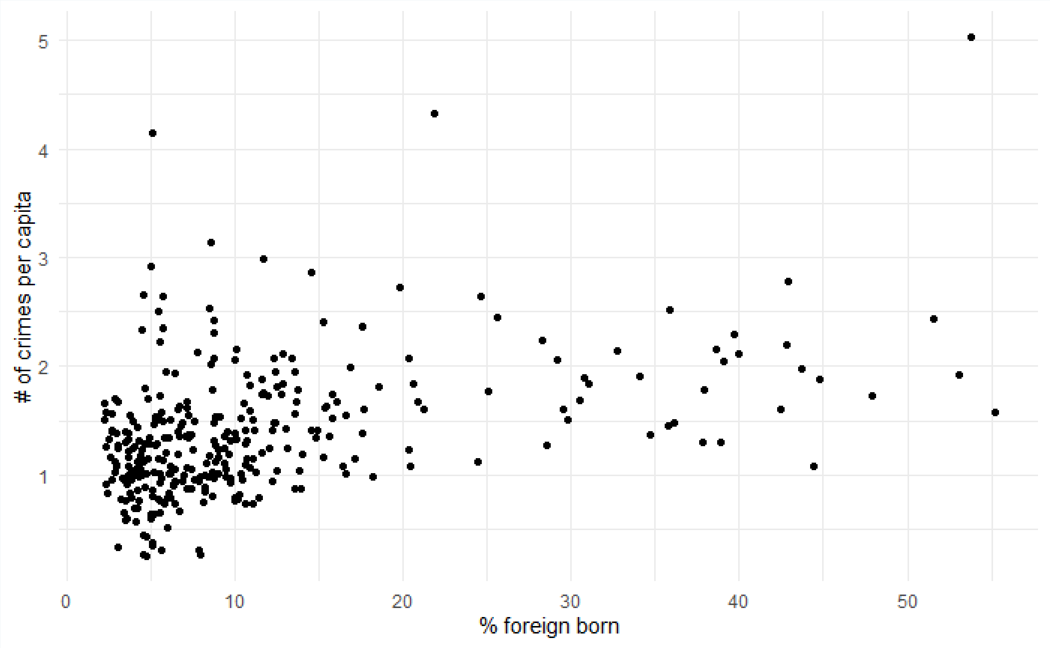
Things need to change from one data point to the next

  - Between similar units: **Cross Sectional Data**
  - Same unit over time: **Time Series Data**

- For the story at hand: Let's look at local authorities (about 350) across the UK; i.e. Cross Sectional

Crime and foreigners

Whenever we first look at data it is a good idea to make ourselves aware what is being measured exactly and what kind of units it is in



Correlation of 0.433

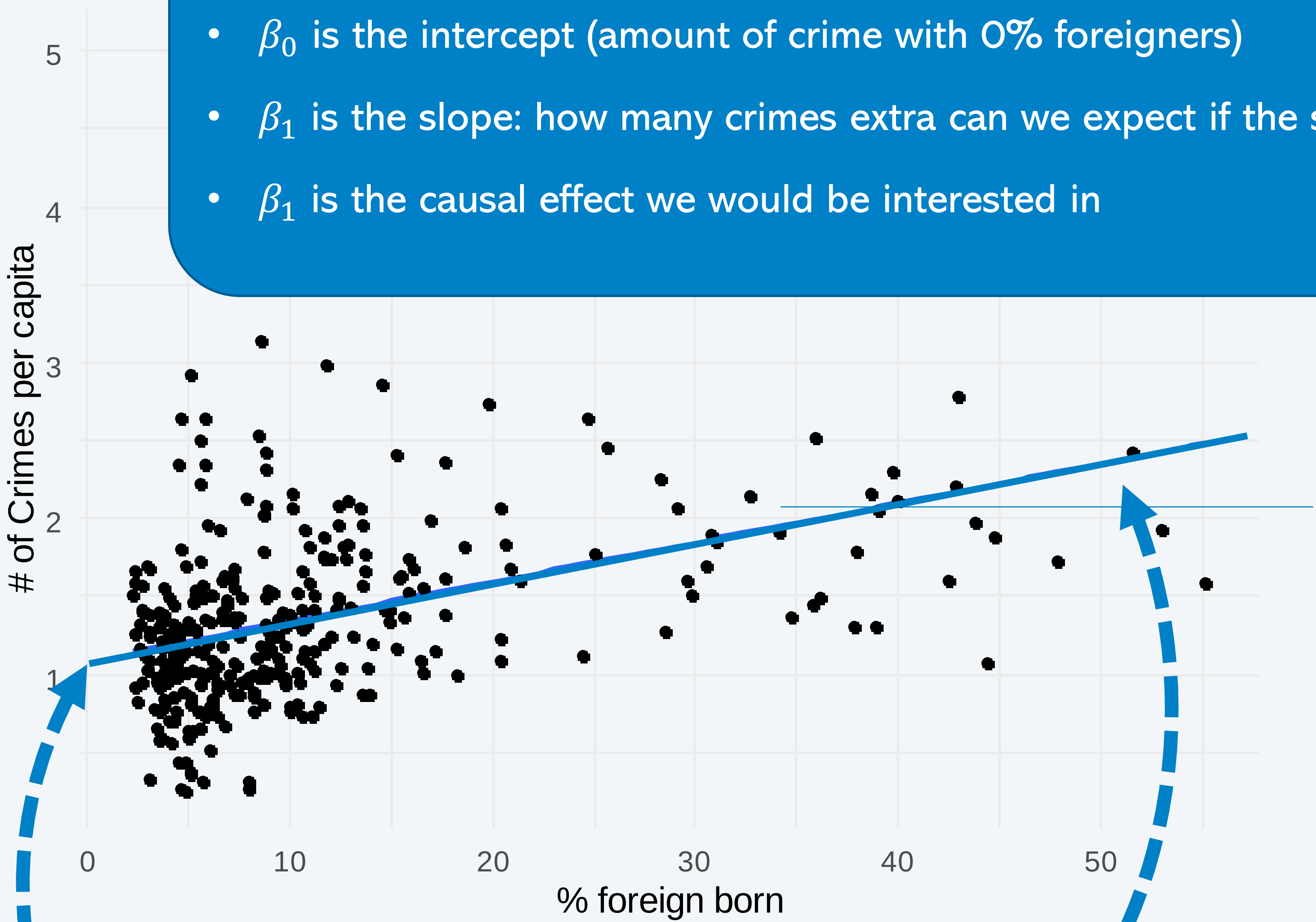
So what would you say? Should we blame foreigners for crime?

Adding a trendline = Modelling the data

- Adding a trendline = modelling the foreigner crime impact as <u>linear</u>

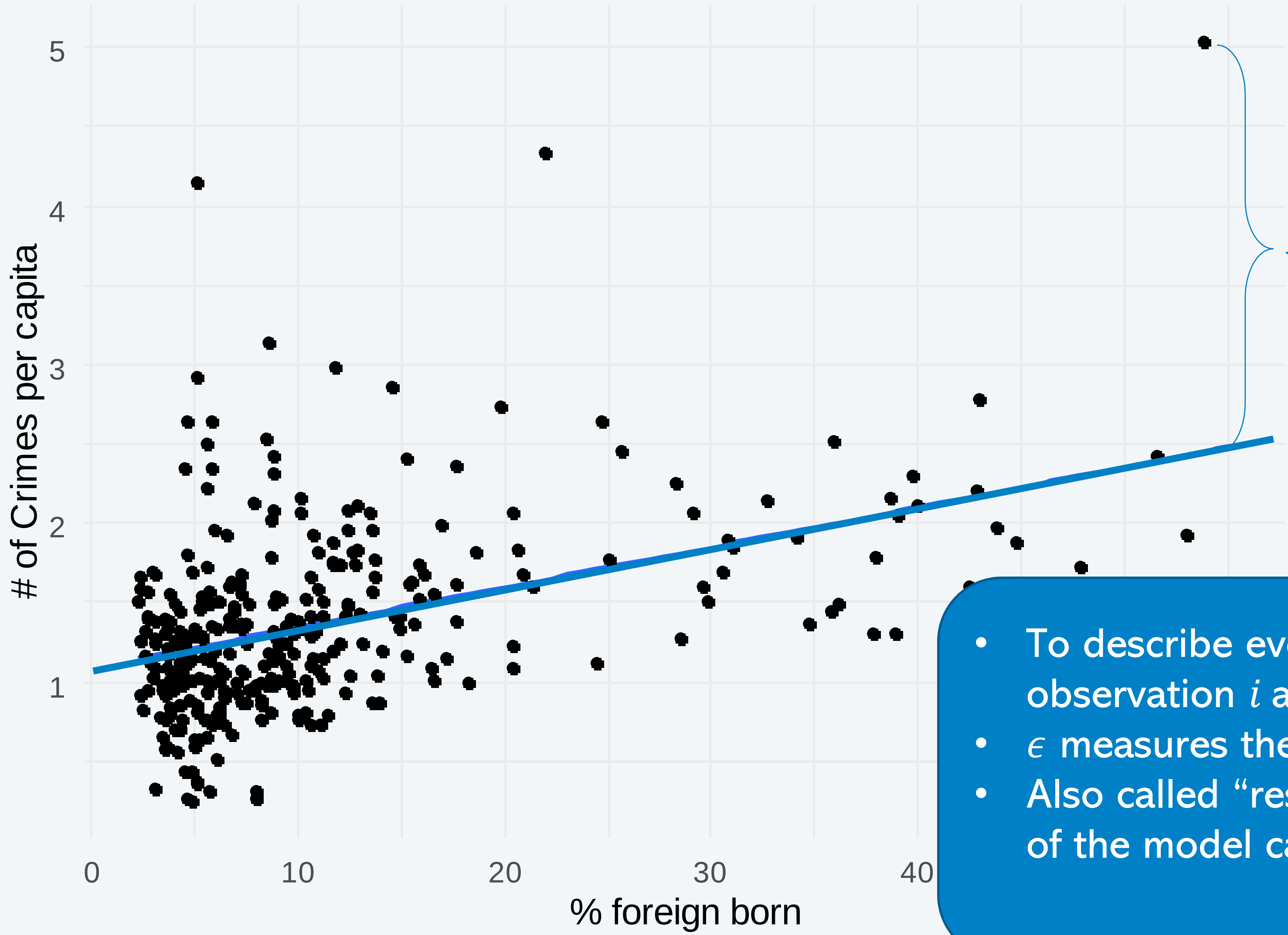
Simplest case

- β_0 is the intercept (amount of crime with 0% foreigners)
- β_1 is the slope: how many crimes extra can we expect if the share of foreigners goes up by 1 percentage point
- β_1 is the causal effect we would be interested in



$$Crime = \beta_0 + \beta_1 foreign$$

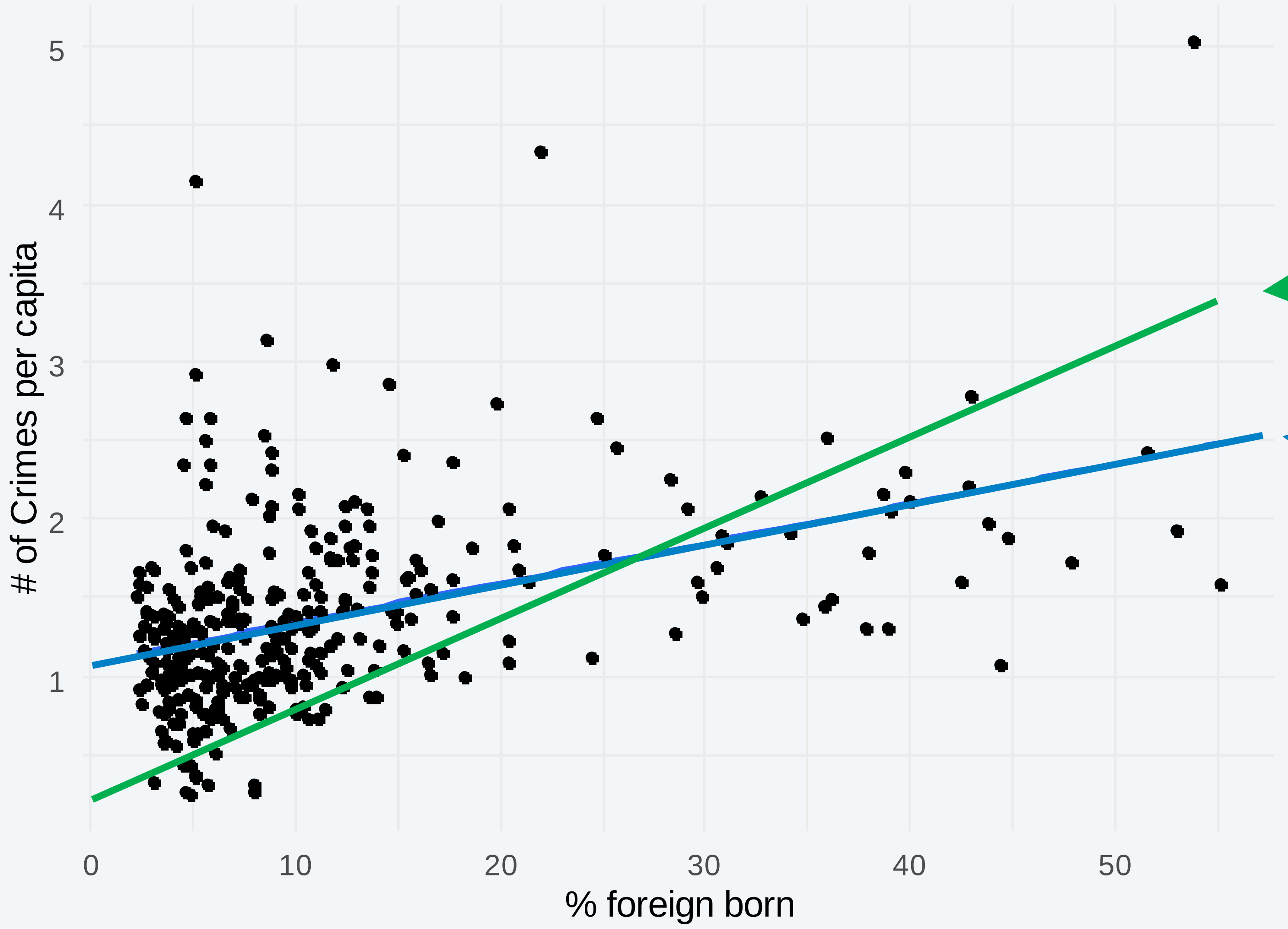
To really model the data we need to do a bit more:



- To describe every individual data point we introduce for every observation i an additional variable ϵ
- ϵ measures the gap between an observation and the trendline
- Also called "residual" or "error term". Everything the first term of the model cannot account for.

$$Crime_i = \beta_0 + \beta_1 foreign_i + \epsilon_i$$

And a bit more: True model driving the data vs estimate



$$Crime_i = \hat{\beta}_0 + \hat{\beta}_1 foreign_i + \epsilon_i$$

$$Crime_i = \beta_0 + \beta_1 foreign_i + \epsilon_i$$

Reasons for differences between true model and estimated model

$$Crime_i = \hat{\beta}_0 + \hat{\beta}_1 foreign_i + \hat{\epsilon}_i$$

$$Crime_i = \beta_0 + \beta_1 foreign_i + \epsilon_i$$

1. It's based on a sample of finite data

That means we are wrong but not systematically wrong: i.e. we took another sample of similar data we are <u>unlikely to make the same</u> mistake

Sample Average $\frac{4+5+6+6+3}{5}$=4.8. Actual (True) Average: 3.5

2. There could be confounding factors that make us systematically over- or under- estimate the actual parameters.

The story of the data?

What are factors that could be driving the data (apart from foreigners being criminals)?

+

Foreigners → Crime

We would see a spurious + relationship

+ +

Big Cities=High Population

Foreigners could be attracted by bigger cities
Bigger cities have more crime

-

Foreigners → Crime

We would see a spurious - relationship

- +

Unemployment

Foreigners often come to work
Hence they go to areas with less unemployment = more work, which also might have less crime

Confounding factors in the linear model: Consider population

$$Crime_i = \beta_0 + \beta_1 foreign_i + \epsilon_i$$

\+

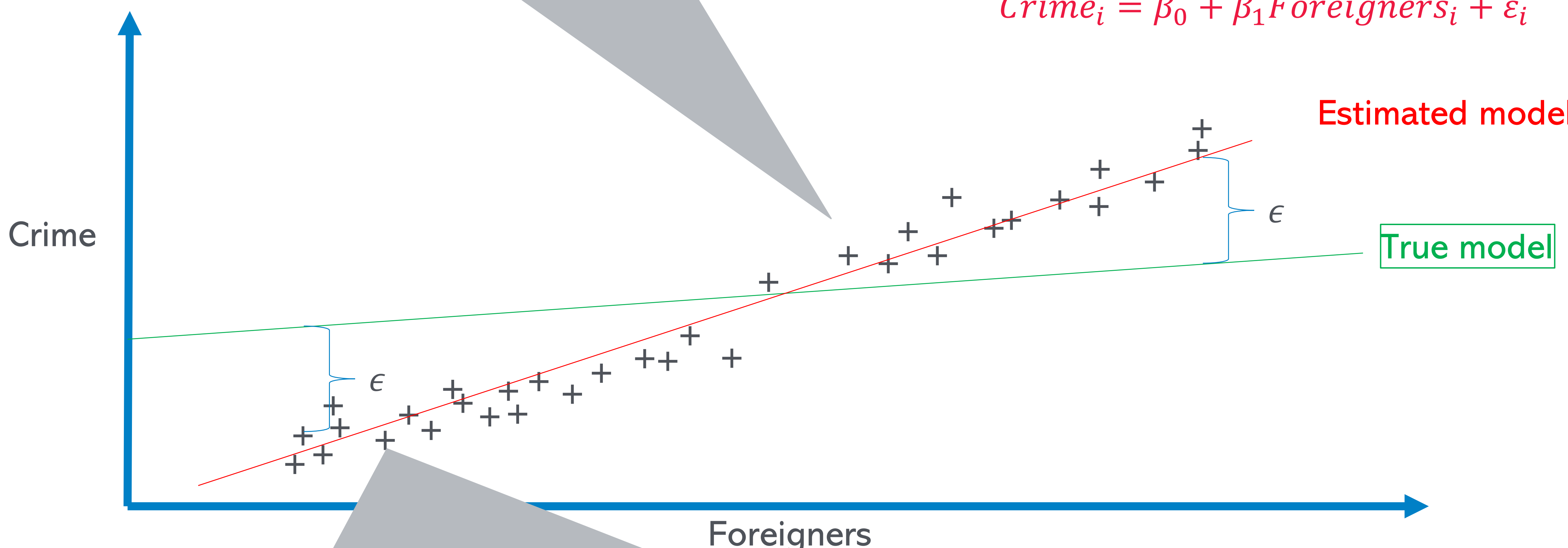
\+

Population

Areas with more foreigners also have more population which means ϵ is big (i.e. positive) when foreign is big (i.e. a positive correlation)

$$Crime_i = \beta_0 + \beta_1 Foreigners_i + \varepsilon_i$$

$$Crime_i = \widehat{\beta_0} + \widehat{\beta_1} Foreigners_i + \widehat{\varepsilon_i}$$



If we estimate the true relationship by a trendline we **overestimate** the strength of the actual relationship; i.e. β_1 **is estimated with upward bias**

Confounding factors in the linear model: Unemployment

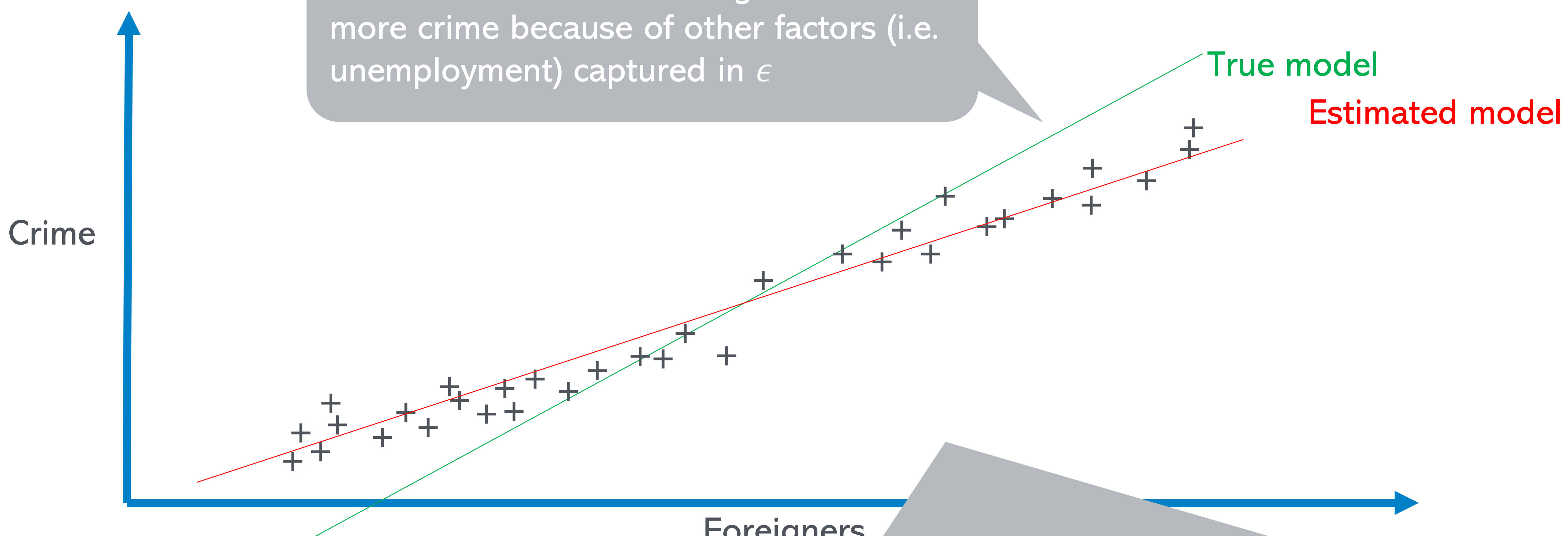
$$Crime_i = \beta_0 + \beta_1 foreign_i + \epsilon_i$$

\+

Unemployment

\-

When there are fewer foreigners there is more crime because of other factors (i.e. unemployment) captured in ϵ



Hence we would underestimate the strength of the true relationship; i.e. **downward bias**

Takeaways

- For good data analysis we have to be aware of:
  - The causal mechanisms we wish to study
  - The causal mechanism that is driving the data (The data story)
- A clear description of these mechanisms (verbally, by arrows or formulas) constitutes the empirical model which helps us to organise the data analysis
- Often the "data story" suggests reasons that could lead to biases in parameters we want to estimate
- There can be upward bias and downward bias
- Make sure you understand the notation of the linear model and how biases might affect it

For next time

- Look at Exercise sheet 1
- Try to install the R and Rstudio software packages (but don't panic if you don't manage)