

BIG DATA MADE SMALL

SQL


➔ through python

➔ via DuckDb 



Alessandro Tang-Andersen Martinello

Head of Data & Analytics @Realkredit Danmark

 @alemartinello.bluesky.social



You will learn to yield great power...

...but you might have to

- Park for a bit your pandas/tidyverse, typically learned in school
- embrace SQL, typically learned in industry jobs



Who am I and why am I here?



What is SQL and why is it useful?



A gentle introduction to SQL

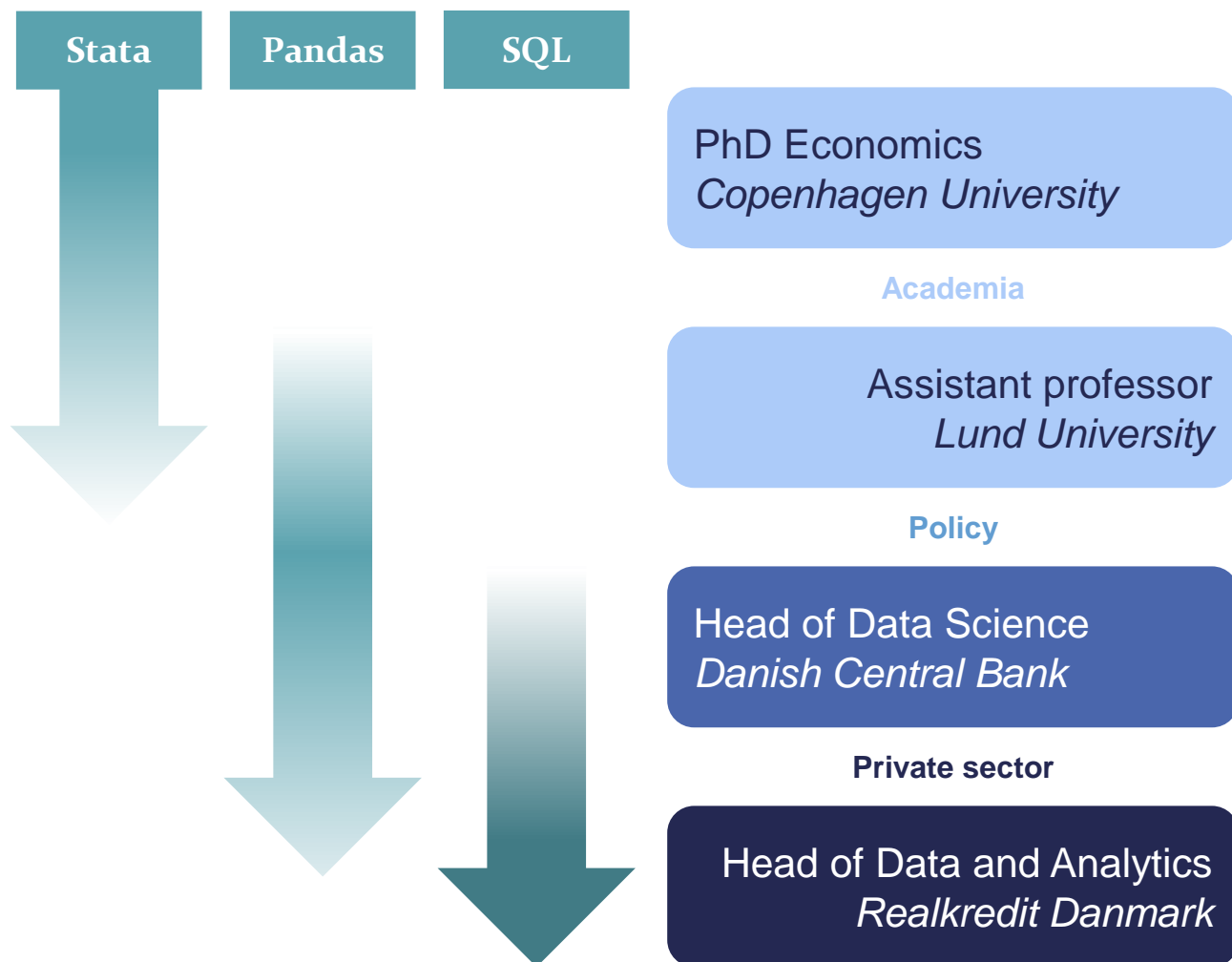


Let's try it out! In a comfy notebook no less!



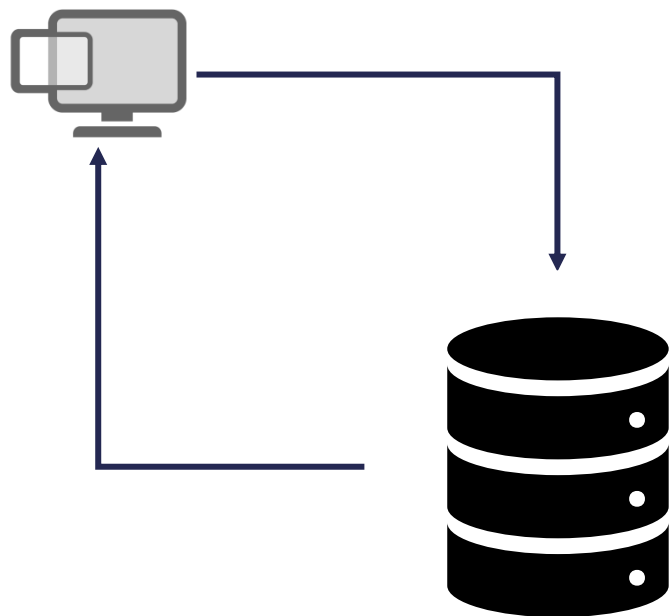
Who am I and why am I here?

My preferred data tools evolved throughout my career



Large corporations have relied on relational databases for their data for 40 years. For good reasons. And SQL is the language allowing you to query that data.

Your workstation



SQL server

Using databases for data flows has multiple perks

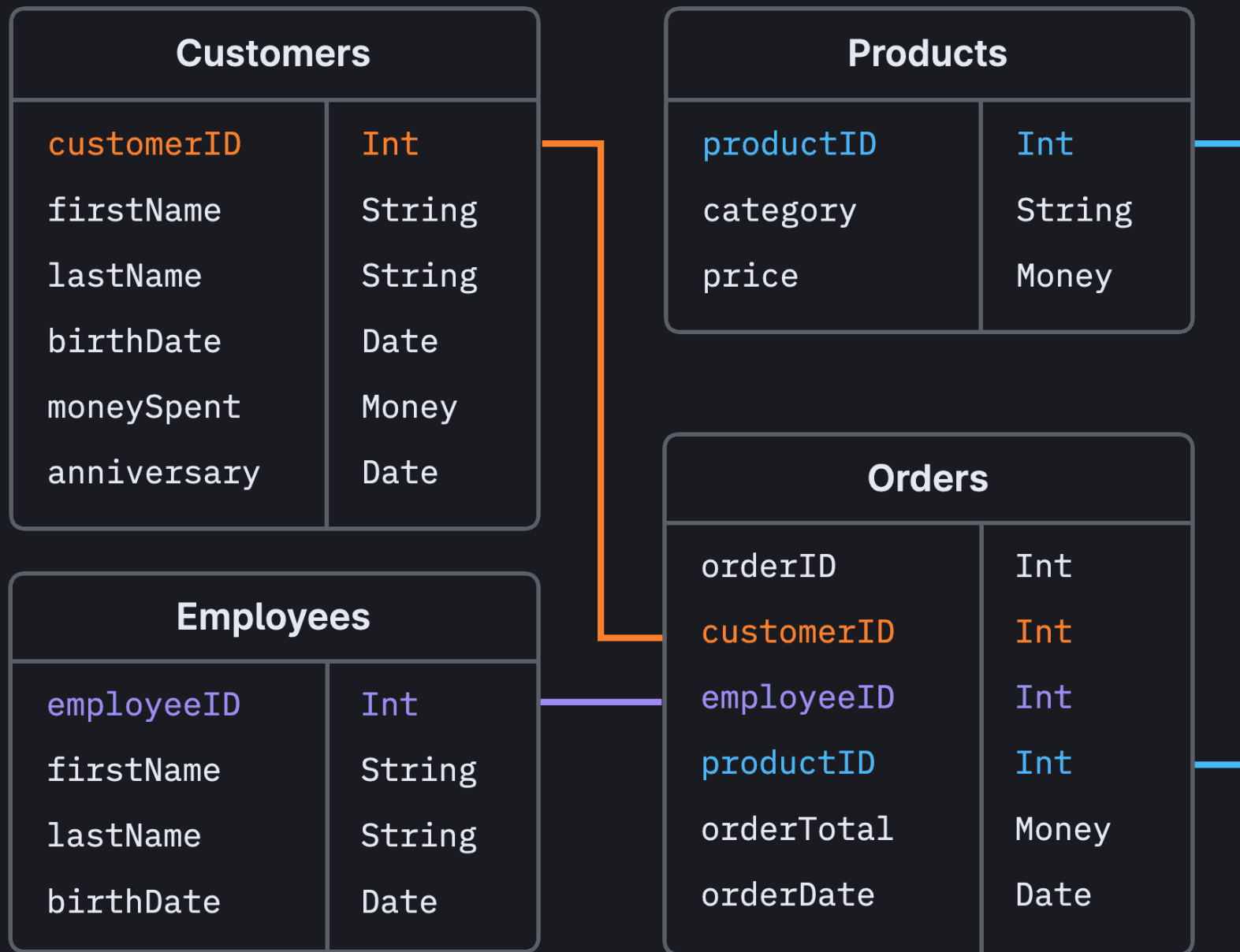
Efficient storage and streamlined data lineage: The raw data is the same for everyone in an organization. There are (supposedly) no multiple copies floating around.

Data protection: It's possible to control who has the right of accessing which data. I.e. you might want to exclude from specific users the right of accessing PEP (politically exposed person) data, or specific tables.

Data versioning becomes code versioning: Instead of having a bunch of heavy data files lying around (*workdata_v1*, *workdata_v2*, *workdata_submission*) you would only keep a versioned (git) history of the SQL code creating your work data.



What is SQL and why is it useful?



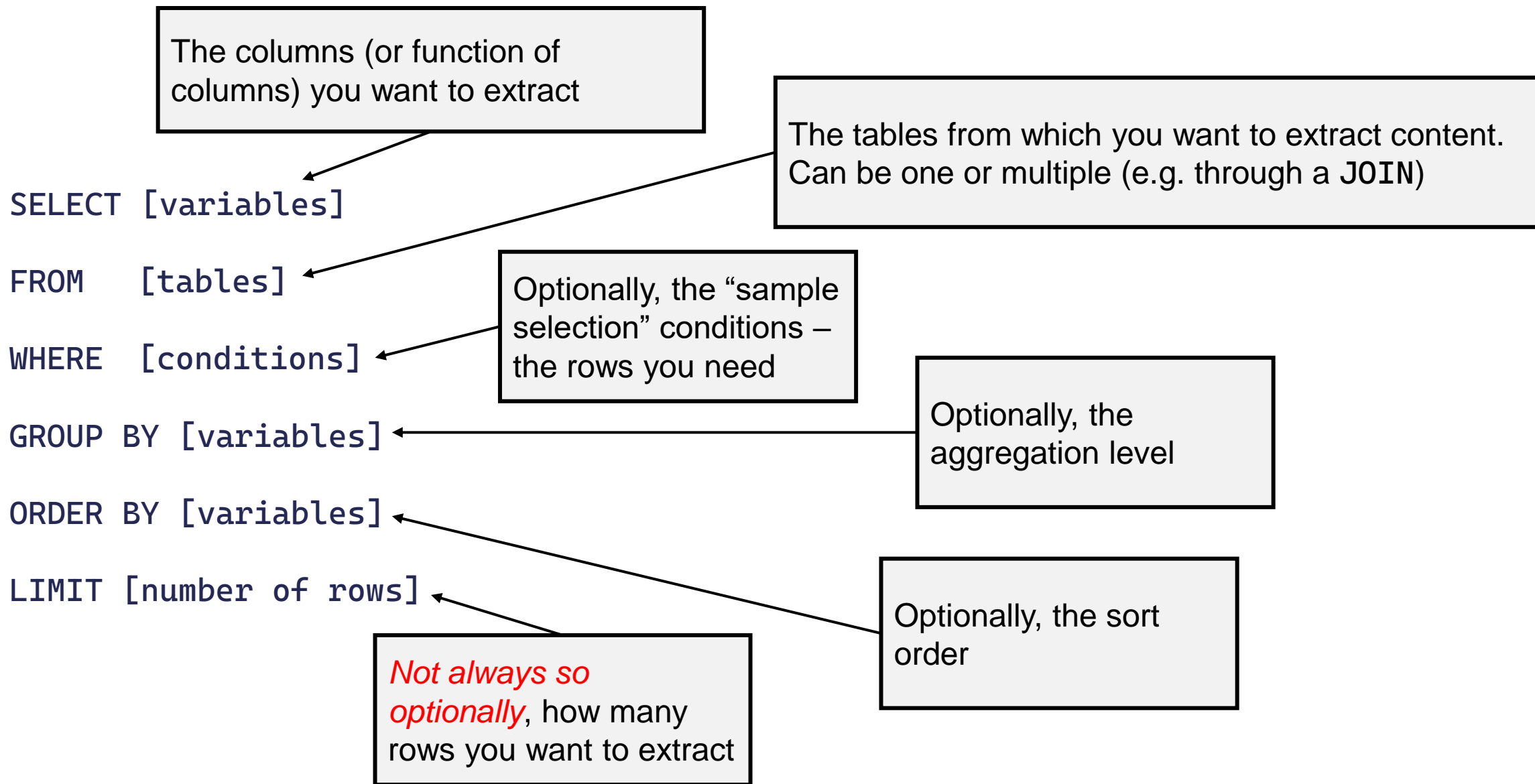
A relational database is an efficient way to store data.

But doing anything with the data typically requires joining multiple tables.

You need a way of doing it fast...



Basic SELECT syntax consists of few stacked blocks



We will replicate this research design

We used ~2bn card transactions (clients of a large Danish bank) paired with ~100ml COVID-19 tests to identify:

- Contagious people visiting a supermarket
- People who shopped **in the same supermarket within ~1 hour** of potential infectors
- How these people tested in the following 7 days

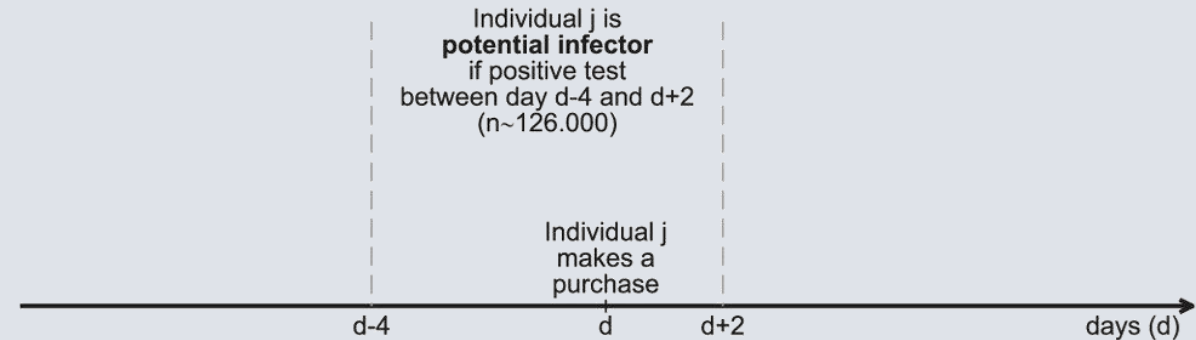
Substantial transmission of SARS-CoV-2 through casual contact in retail stores: Evidence from matched administrative microdata on card payments and testing

Niels Johannesen , Alessandro Tang-Andersen Martinello, Bjørn Bjørnsson Meyer , and Thais Lærkholm Jensen [Authors Info & Affiliations](#)

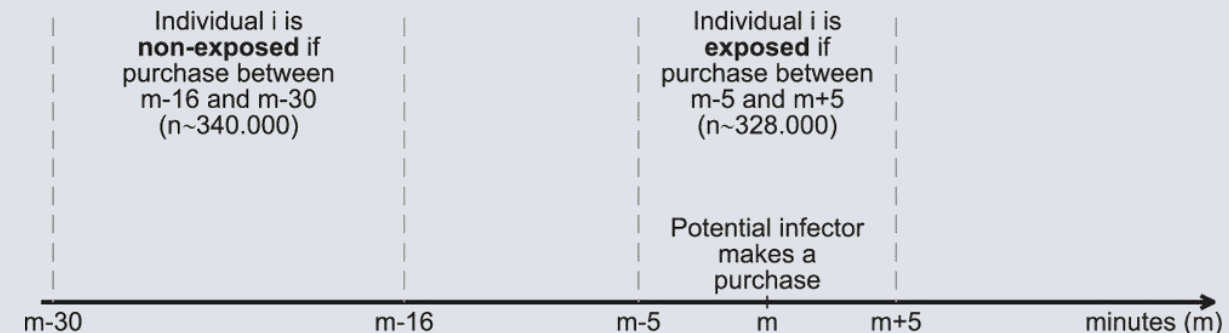
Edited by Jose Scheinkman, Columbia University, New York, NY; received October 10, 2023; accepted March 21, 2024

April 17, 2024 | 121 (17) e2317589121 | <https://doi.org/10.1073/pnas.2317589121>

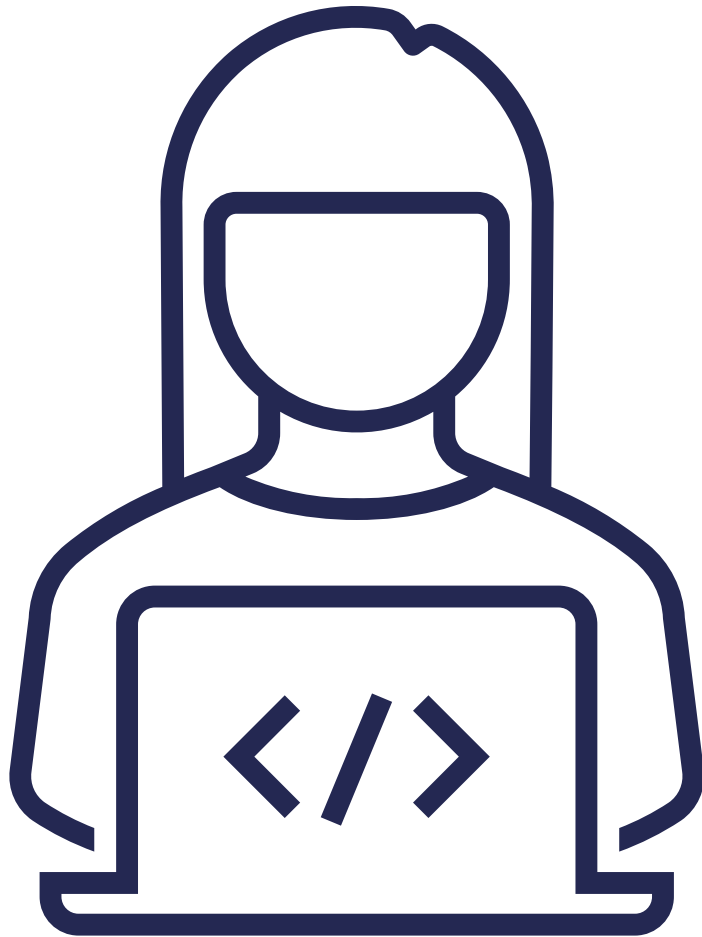
Step #1: Potential infectors



Step #2: Exposed and non-exposed



Let's try it out!



Setup

Steps 1-3 are only required the first time you set up the folder.

1. Install [uv](#), an extremely fast Python package manager, according to [these instructions](#).
2. Git clone this repository to your chosen directory and change directory

```
git clone git@github.com:alemartinello/BigDataMadeSmall.git  
cd BigDataMadeSmall
```



3. Create the main dataset we'll be using during this workshop (simulated). It will take a few minutes to run.

```
uv run create_data.py
```



4. Open jupyter and select the notebook

```
uv run jupyter lab .
```

