# Research Design 2

Quantitative Analysis
Week 3

# Review

- Last week we talked about the importance of having a solid research design (instead of diving straight in and working with data and statistics), and introduced the **PPDAC Cycle**:

  **Problem → Plan → Data → Analysis → Conclusion**

- We discussed the factors that are required for good social science research:

  **(1) Inferential** **(2) Replicable**

  **(3) Falsifiable** **(4) Rules-based**

  **(5) Uncertain** **(6) Public**

# Review (2)

- Next, we talked about the **fundamental problem of causal inference.**
  - We can only ever observe **one** of the possible outcomes of a treatment.
  - For each case *i*, given treatment $T_i$, we can only observe one of the possible outcome variables $Y_i(T)$ - in the simplest example, where the treatment is either given or not, this means we can observe either $Y_i(0)$ or $Y_i(1)$, but we can never see them both.
  - The outcomes we cannot observe are called **counterfactuals**.
  - We try to overcome this problem by designing research that allows us to estimate those counterfactuals – by looking at many cases and finding ways to **control** any other differences between those cases (**pretreatment variables**, or **confounders**).

# Review (3)

- Finally, we discussed the two major categories of research design which we can use to overcome this problem:
- **Randomised Controlled Trials (RCTs)**, in which the treatment is assigned randomly to a large group of people, so there is no *systematic* difference in the pretreatment variables.
    - This has high **internal validity** (the causal effect is very clear and easy to see), but can have low **external validity** (the experiment is an "unnatural" setting that might not reflect what happens in the real world).
- **Observational Research**, where we take measurements from the real world and use statistical analysis to uncover causal relationships.
    - **Internal validity** can be low as we have no control over the treatments, creating various kinds of bias; but **external validity** is high.

# This week…

This week, we're going to discuss arguably the most important part of research planning.

1. **Devising a Research Question**
2. **Establishing Hypotheses**
3. **Operationalising your Hypotheses**

# Research Questions

# Choosing a Research Question

- There is no rule about how to choose a research topic – this is determined by your own interests.
- However, choosing a research question within the field of that topic demands a more systematic approach.
- A research question should be:
- **Important**: this is not a value judgement; it simply means that it's relevant within its own field.
- **Original**: meaning that it contributes something new to the existing literature about the topic.

# The Literature Review

- You cannot assess the **importance** and **originality** of your research question without having an up-to-date knowledge of the literature that already exists in your field.
- Academic papers almost always start with a **literature review** summarising the most important existing work in the field.
- This serves three important functions:
  - It proves the authors' expertise on other work in their field;
  - It makes clear what work by other researchers they are building upon or responding to;
  - It allows readers who may not be experts on the topic to understand the major debates or issues that exist in this field.

# How to do a literature review

- There are two main types of paper you must identify in a literature review.

- **Foundational Works** are the papers which **define** the field you're studying.
  - Some of these may be quite old (for example, papers about conspiracy theories often cite a 1956 paper by Richard Hofstader called "The Paranoid Style in American Politics"), and they are often connected by a narrative – an argument between two different groups of researchers, for example.
- **Recent Works** are papers from the past few years that reflect the "state of the art".
  - Reading these should tell you what discussions or issues are ongoing in your chosen field right now.

# Useful Tools

- Use online tools extensively:
  - **Google Scholar** is the default for most people.
  - **Web of Science** and **Scopus** can give more reliable, but less extensive, results.
  - **CiNii** for Japanese-language papers.
  - There are other databases of research for specific fields or regions; check the Waseda Library website to see if any may be useful to you.

- When you find good papers in your field, **check their bibliography / references**. You'll usually find many other papers cited which will also be relevant to your research.

# AI Bibliography Tools

- A very recent development is the creation of AI tools which use machine learning analysis to automatically identify literature that may be relevant to your research.
- These are very promising, but generally aren't ready to be used in real research yet.
- By all means, use digital literature review tools (**not** ChatGPT – we'll talk about that in a moment) to suggest papers to you, but make sure you read them yourself before you cite them; these tools often make serious mistakes!

# Connected Papers

One digital literature review tool you may find useful is **Connected Papers**.

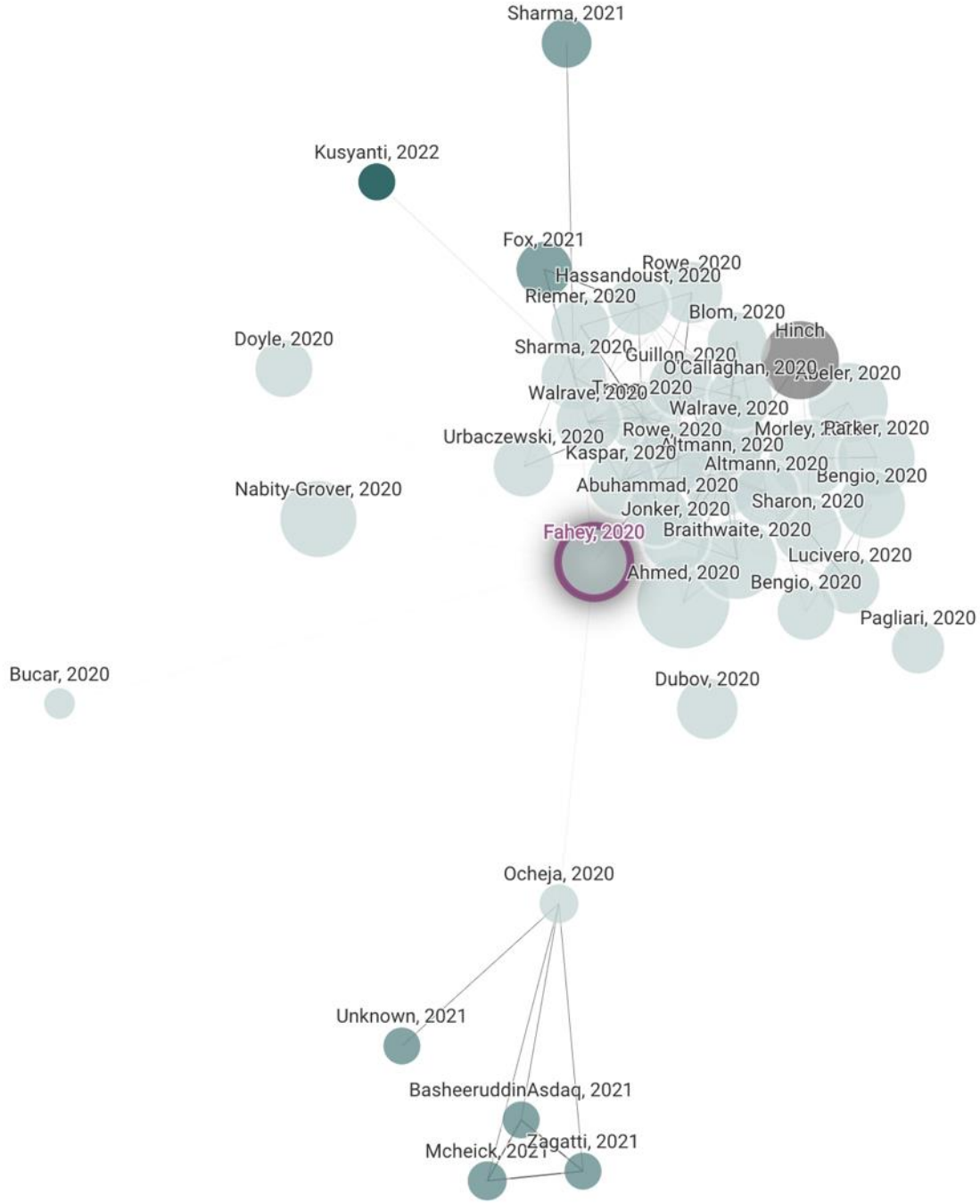
It analyses the network of citations that connects papers and can point you to new areas of academic literature on a topic, as well as helping to uncover the most important foundational literature.

You need to identify a paper that's very relevant to your research question. Connected Papers will then construct the citation network for that paper.

www.connectedpapers.com

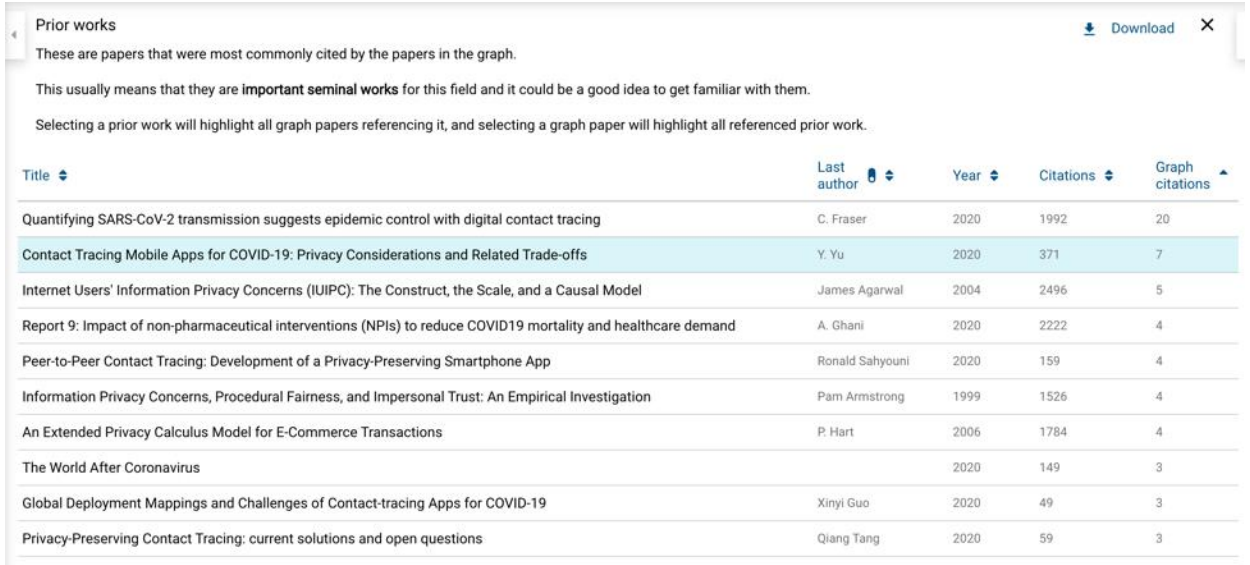
# Citation Networks

- For example, here is the citation network around one of my papers from 2020.
- You can see that there's a big cluster of other papers which all tend to cite each other – this is probably the most "mainstream" scientific discourse on this topic.
- Other smaller clusters might be side discussions, or only tangentially related to the main topic.



# Prior and Derivative Works

- Connected Papers will also show you lists of prior works (which most papers on this topic tend to cite – i.e., **foundational works**), and derivative works (which cite a lot of papers on this topic, and are probably the most **recent works**)



Prior works

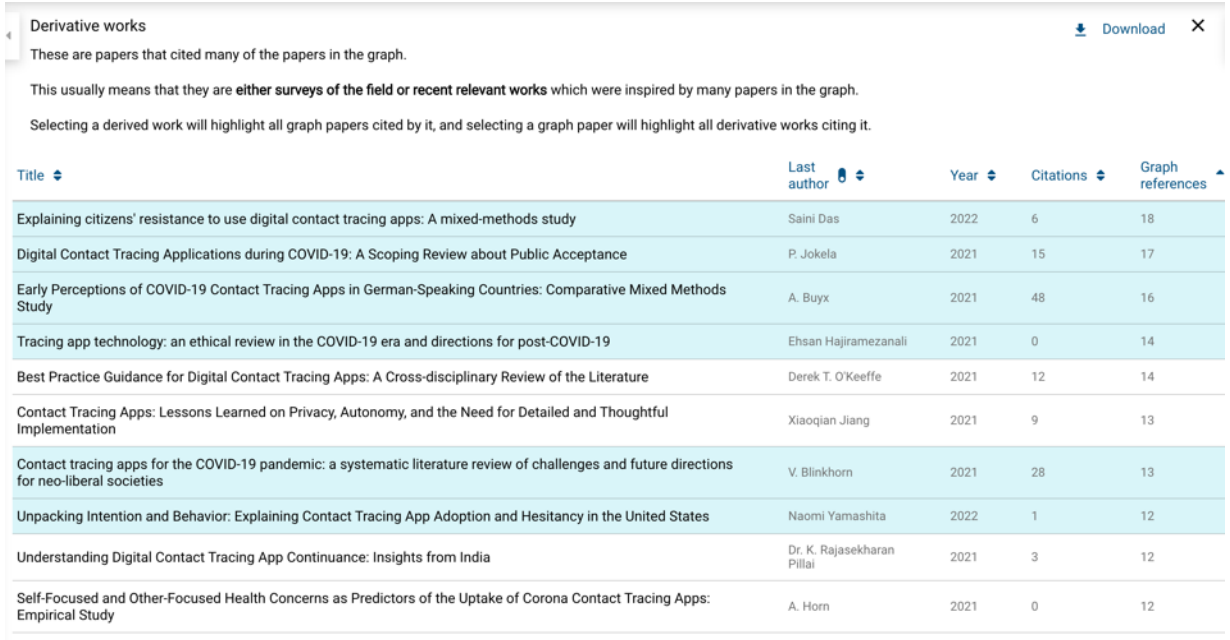
Download

These are papers that were most commonly cited by the papers in the graph.

This usually means that they are **important seminal works** for this field and it could be a good idea to get familiar with them.

Selecting a prior work will highlight all graph papers referencing it, and selecting a graph paper will highlight all referenced prior work.

| Title | Last author | Year | Citations | Graph citations |
|---|---|---|---|---|
| Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing | C. Fraser | 2020 | 1992 | 20 |
| Contact Tracing Mobile Apps for COVID-19: Privacy Considerations and Related Trade-offs | Y. Yu | 2020 | 371 | 7 |
| Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model | James Agarwal | 2004 | 2496 | 5 |
| Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand | A. Ghani | 2020 | 2222 | 4 |
| Peer-to-Peer Contact Tracing: Development of a Privacy-Preserving Smartphone App | Ronald Sahyouni | 2020 | 159 | 4 |
| Information Privacy Concerns, Procedural Fairness, and Impersonal Trust: An Empirical Investigation | Pam Armstrong | 1999 | 1526 | 4 |
| An Extended Privacy Calculus Model for E-Commerce Transactions | P. Hart | 2006 | 1784 | 4 |
| The World After Coronavirus | | 2020 | 149 | 3 |
| Global Deployment Mappings and Challenges of Contact-tracing Apps for COVID-19 | Xinyi Guo | 2020 | 49 | 3 |
| Privacy-Preserving Contact Tracing: current solutions and open questions | Qiang Tang | 2020 | 59 | 3 |



Derivative works

Download

These are papers that cited many of the papers in the graph.

This usually means that they are **either surveys of the field or recent relevant works** which were inspired by many papers in the graph.

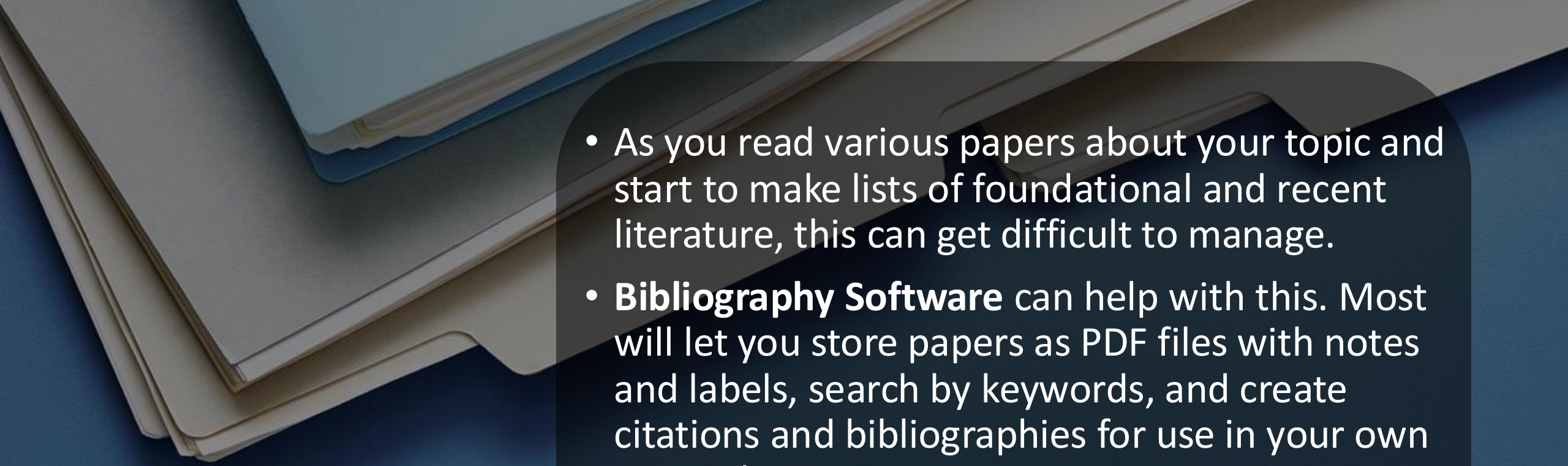
Selecting a derived work will highlight all graph papers cited by it, and selecting a graph paper will highlight all derivative works citing it.

| Title | Last author | Year | Citations | Graph references |
|---|---|---|---|---|
| Explaining citizens' resistance to use digital contact tracing apps: A mixed-methods study | Saini Das | 2022 | 6 | 18 |
| Digital Contact Tracing Applications during COVID-19: A Scoping Review about Public Acceptance | P. Jokela | 2021 | 15 | 17 |
| Early Perceptions of COVID-19 Contact Tracing Apps in German-Speaking Countries: Comparative Mixed Methods Study | A. Buyx | 2021 | 48 | 16 |
| Tracing app technology: an ethical review in the COVID-19 era and directions for post-COVID-19 | Ehsan Hajiramezanali | 2021 | 0 | 14 |
| Best Practice Guidance for Digital Contact Tracing Apps: A Cross-disciplinary Review of the Literature | Derek T. O'Keeffe | 2021 | 12 | 14 |
| Contact Tracing Apps: Lessons Learned on Privacy, Autonomy, and the Need for Detailed and Thoughtful Implementation | Xiaoqian Jiang | 2021 | 9 | 13 |
| Contact tracing apps for the COVID-19 pandemic: a systematic literature review of challenges and future directions for neo-liberal societies | V. Blinkhorn | 2021 | 28 | 13 |
| Unpacking Intention and Behavior: Explaining Contact Tracing App Adoption and Hesitancy in the United States | Naomi Yamashita | 2022 | 1 | 12 |
| Understanding Digital Contact Tracing App Continuance: Insights from India | Dr. K. Rajasekharan Pillai | 2021 | 3 | 12 |
| Self-Focused and Other-Focused Health Concerns as Predictors of the Uptake of Corona Contact Tracing Apps: Empirical Study | A. Horn | 2021 | 0 | 12 |

# What about ChatGPT?

- ChatGPT is good at a lot of things: **literature review is not one of them!**
- You have probably heard of “AI hallucinations” – the tendency of AI tools to invent, or hallucinate, facts and information that doesn’t exist.
- It does this **extensively** with academic papers. ChatGPT will invent papers that don’t exist and claim they are from journals that don’t exist, with plausible sounding titles. Often it will say they were authored by real researchers!

- However: ChatGPT **can** be useful for summarising papers rapidly, to help you decide whether it’s worth taking the time to read them in full.



# Manage your Bibliography

- As you read various papers about your topic and start to make lists of foundational and recent literature, this can get difficult to manage.
- **Bibliography Software** can help with this. Most will let you store papers as PDF files with notes and labels, search by keywords, and create citations and bibliographies for use in your own research papers.
- **Zotero** is a good, free option which integrates directly into your web browser (one click to save the paper you're reading to your library) and with Microsoft Word (for easily inserting citations).

# What next?

- So you've read the important papers in your field and you're up to date on the latest research.
- If you started with a clear question in mind, you should now know whether it's been answered before, or what other researchers think about it.
- If not, by now you should hopefully have noticed some questions that haven't yet been answered by other researchers – or perhaps ones you think have been answered wrongly.

# Types of Question

**Descriptive Question**

- "What percentage of people are opposed to policies that increase immigration?"
- *This kind of question has no causal aspect!*

**Relational Questions**

- "Is there a relationship between migrant numbers in a community and local attitudes to immigration?"
- "What demographic factors are associated with negative sentiments towards immigration?"

**Causal Questions**

- "Does interaction with migrants change citizens' attitudes to immigration?"
- "What effects do anti-immigration political candidates' election campaigns have on public sentiment towards migrants?"

# Establishing Hypotheses

Making an educated guess.

# What is a hypothesis?

- A hypothesis is your best guess at the answer to your research question – based on your knowledge and expertise of the subject, and the prior literature you have read.
- It is phrased as a statement which your research project will seek to prove or disprove.
- Sometimes you start from a research question and devising the hypothesis requires further study; often, you'll devise them both together.
- Generally, you should not devise a hypothesis **after** gathering and looking at your data. **Always go from Theory to Data; never from Data to Theory.**

# From Question to Hypothesis

- Descriptive Questions
  - These do not have hypotheses – descriptive research does not seek to prove or disprove anything.
- Relational Questions
  - The hypothesis for a relational question is simply that the relationship exists.
  - Some relational questions are more descriptive ("what factors are associated with X"), and do not have a hypothesis statement.
- Causal Questions
  - The hypothesis for a causal question is a **clear statement of cause and effect**.

# Choosing Research Questions

KKV Chapter 1 suggests this list of ways you might go about choosing a research question and/or hypothesis.

1. Choose a hypothesis seen as important by scholars in the literature but for which no one has completed a systematic study. Find evidence in favor of or opposed to the favored hypothesis.
2. Choose an accepted hypothesis in the literature that we suspect is false (or one we believe has not been adequately confirmed) and investigate whether it is indeed false or whether some other theory is correct.
3. Attempt to resolve or provide further evidence of one side of a controversy in the literature—perhaps demonstrate that the controversy was unfounded from the start.
4. Design research to illuminate or evaluate unquestioned assumptions in the literature.
5. Argue that an important topic has been overlooked in the literature and then proceed to contribute a systematic study to the area.
6. Show that theories or evidence designed for some purpose in one literature could be applied in another literature to solve an existing but apparently unrelated problem.

# Questions and Hypotheses

Pick **one** of these topics, and write:

One **descriptive** research question;

One **causal** question and hypothesis.

- The invasion of Ukraine
- LGBT rights in Japan / your home country
- Policies to tackle inflation
- Voters' faith in U.S. elections

# Operationalising Hypotheses

From Abstract to Empirical

# Operationalisation

- Hypotheses describe relationships between **concepts** – abstract ideas that try to capture something about our reality.
- Examples of abstract concepts you might find in political research include **Justice**, **Freedom**, **Democracy**, **Populism**, **Apathy**, **Trust**…
- Some concepts can be observed directly; e.g., we can observe the concept of **Wealth** by looking at income, or net worth; we can observe the concept of **Education** by looking at diplomas and degrees.
- Most concepts cannot be directly observed in reality. They represent **abstract** ideas, and their definitions may be **subjective**.

- Operationalisation is the process of moving a hypothesis from the **abstract** to the **empirical**, by transforming its concepts into measurable variables.
- Each concept in the hypothesis is given an **operational definition**, which clearly specifies what the concept means (in the context of your research), and how the concept will be measured.
- The more abstract or complex a concept is, the more important it is to establish a clear, detailed, and unambiguous definition.

# Defining Concepts

- Even simple concepts have to be clearly defined.
  - For example, the concept of **individual wealth** might be measured very differently if you define it as *income,* as *net worth*, or as *purchasing power*.
  - The decision of how to measure this variable could totally change your research outcomes!
- If you're using data someone else has compiled, like government statistics, you have to be aware of the definitions they have chosen to use.
  - Seemingly concrete concepts like **unemployment rate** can be measured very differently in various countries – the measurement method may even change in the same country over time.

# Defining Complex Concepts

- For complex concepts, there is often a debate in the prior literature about how they should be defined.
  - Mostly, researchers will pick one side of this debate (and justify this choice).
  - It's also possible to critique all the existing definitions and put forward your own instead, as long as you can clearly justify this decision.
- For example, there is a major debate over how to define **Populism**.
- Most political scientists believe populism has been an important factor in world events in recent years – but there is often disagreement over how to operationalise this abstract concept.

# Defining Populism

- Most researchers agree broadly on what populism is, but not necessarily on how to measure it, or what factors are required to label a politician or voter as "populist".
- **Anti-elite** feelings are a key part – but does "elite" just mean political and governmental elites? Are there instances where it includes media, academia, wealthy / educated people, etc.?

- **Popular sovereignty**, or the "will of the people", is also key – but some definitions claim that populists see "the people" as an exclusive subset of the population ("the real/true people of the nation"), and other definitions ignore this.
- Kirk Hawkins argues that **Manicheanism** (seeing the world in black and white, as a struggle between good and evil) is a key factor. Otake Hideo has argued that populism is inherently **theatrical** in its presentation.

- To use the concept of “populism” in your research, you would need to sort through these definitions and pick the one that is most appropriate and justifiable for you.
- Then, you would need to find a way to **measure** it consistently.
  - This might be a set of **survey questions** (there are many different “batteries” of questions on populism, all of them measuring slightly different versions of the concept), or a **rubric** for evaluating the level of populist factors in a political speech.
  - You may also find an **existing data set** created by other researchers – e.g., there is a data set which labels political parties as populist or not – but make sure they created that data based on the same definition you’re using!
- Populism is an especially contentious concept right now – but you will need to follow a similar process for any complex concept.

# Variables

- Operationalising the concepts in your hypothesis results in the creation of **variables**.
- A variable is a measurement of a characteristic or quality that defines an object – a person, group, country, etc.
- In social science research, many of our concepts are abstract and cannot be directly measured. Other concepts are concrete but may be difficult to measure for practical or ethical reasons.
- In these cases, operationalisation produces **proxy variables** – things which we **can** measure, and hypothesise to be closely correlated with the unmeasurable concept we're interested in.

# Examples of Proxies

- Commonly used proxies include:
  - *Test scores* or *years of education* as as proxy for **intelligence**.
  - *Gross Domestic Product (GDP)* as a proxy for a nation's **quality of life**.
  - *Ability to identify political leaders* as a proxy for **political knowledge.**
  - *Number of reported crimes* as a proxy for **actual crime rate.**

- Proxies are common in all fields of science; e.g., in environmental science, tree rings are used as a proxy for historical climate conditions.
- It's important to be aware when you're using a proxy – and what its limitations may be.
  - "Number of reported crimes" *may* be closely correlated to the actual crime rate – but it is *not* the actual crime rate.

# Variable Types

## Categorical

- **Nominal**
  - Two or more categories, with no specific order – e.g. prefectures, ethnicities.
- **Ordinal**
  - Multiple categories with a ranked order – e.g. education level, degree of preference.
- **Dichotomous / Binary**
  - Categories with only two levels – e.g. Male / Female, Yes / No.

## Continuous / Quantitative

- **Interval**
  - Measured quantities with a numeric value – e.g. temperature.
- **Ratio**
  - Measured numeric quantities with a fixed zero point – e.g. distance, vote counts.

# Other Operationalisation Decisions

**What is your unit / level of analysis?**

- This might be individuals, groups, companies, towns, regions, nations…
- It's generally best practice to measure all your variables at the same **unit level**.
  - Measuring one variable with an individual-level survey, and another with national economic data, makes any comparison between them tricky (although not necessarily impossible).

**Snapshot or Time Series?**

- Snapshot data captures your variables at one point in time – e.g. election results, or a survey conducted all in one go.
- Time series data measures the same variables (usually in the same cases) repeatedly over time – e.g. economic trends, panel surveys conducted in multiple waves.
- The conclusions you can draw from each type of data are different.

# Next Week

You have a reading for next week…

Rather than one full chapter, this will be selected sections from KKV (Designing Social Inquiry) on data collection and sample bias.
This reading will be available on Moodle tonight.