Socio-Informatics 348

Intro to Computational Social Science

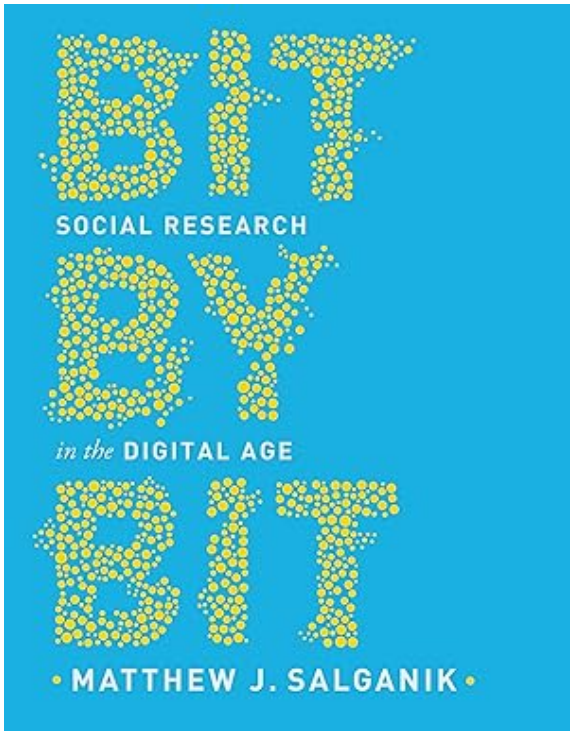
Dr Lisa Martin

Department of Information Science
Stellenbosch University

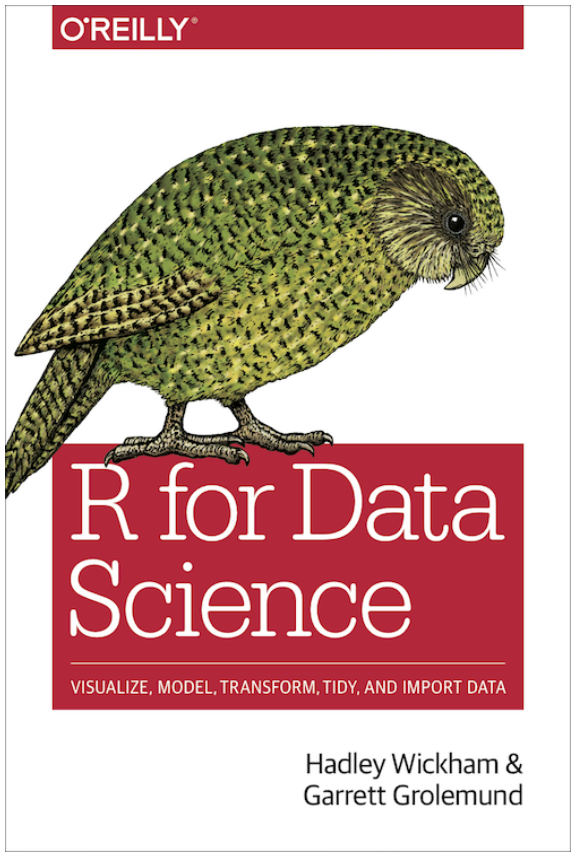
Some Admin

- Class Rep
- Practical timeslots (Group 2)

Today's Reading



Bit by Bit, Chapter 2



R for Data Science, Intro

What is Computational Social Science?

- The use of digital data and computation to study human behaviour
- Intersects social science theory with data science methods
- Examples: social media analysis, digital trace data, text mining, network analysis

Digital data

- Collecting data about behaviour was expensive, and therefore relatively rare...
- New ways to collect and analyse existing sources – digitisation
- Creation of **big data** (focus of this chapter)
  - digital trace data (online sources)
  - Government administrative records (census, tax records)
  - Corporate data (sales, customer interactions)
- Big data is part of a (pre-existing) broader category: observational data.
- No intervention

Big Data

3 Vs Defining Big Data

- Volume: vast amounts of data generated
- Variety: diverse data formats (text, images, video)
- Velocity: data generated at high speed

In general: Created for other purposes, not specifically for research

- Twitter vs a General Social Survey for studying public opinion
- Speed and scale vs careful sampling and measurement
- Typically needs repurposing before it can be used

Repurposing

- "Found data" vs "collected data"
- Important to try to understand the original purpose of the data
- How does the data differ from your ideal?

10 Common Characteristics of Big Data

1. Big

- Researchers have **many** more data points than they did in the past
- However, this does not mean that the data is necessarily better
- Large number of observations **should serve a purpose**
- Chetty et al. (2014) - Heterogeneity:
  - Tax records of 40 million people (+ a worthwhile question)
  - Estimating the heterogeneity in intergenerational mobility across regions in the US
  - Large sample size allowed for a more detailed breakdown of mobility patterns

10 Common Characteristics of Big Data

1. Big

- Similarly, large datasets allow researchers to detect small differences
  - Important for cases where small changes can have large/meaningful impacts
- However, beware of systematic errors:
  - Systematic noise can lead to false positives
  - Example: Back et al. (2010) - Pager data, skewed by messages from a bot

10 Common Characteristics of Big Data

2. Always-on

- Constantly collecting data
- Allows researchers to capture unexpected events, rather than surveying after the fact
- Able to produce real-time estimates
  - Real-time social media data can be used aid emergency response teams
  - Real-time estimates of economic activity

10 Common Characteristics of Big Data

3. Nonreactive

- When people know they are being observed, they may change their behaviour
- With many big data sources, participants are generally not aware that data collection is taking place
  - Ethical concerns?
- This can lead to more natural behaviour, but researchers should still be aware of potential biases

10 Common Characteristics of Big Data

4. Incomplete

- May not contain all the information needed for your specific research question
- Common feature in data not collected for research purposes
- Some information can be added via imputation or record linkage

10 Common Characteristics of Big Data

5. Inaccessible

- Many big data sources are often controlled by government or corporates
- Legal, business or ethical barriers to sharing the data
- Researchers may sometimes gain access to data through partnerships, but this may downsides:
  - Unable to share and have your work verified or replicated
  - Limited questions you can ask / Preserving company reputation
  - People may think that the partnership influenced your results

10 Common Characteristics of Big Data

6. Nonrepresentative

- Even if a big data source is large, it may not be representative of the population you are interested in
- Be aware of the source of the data
- Example: Social media data may not represent the general population

10 Common Characteristics of Big Data

7. Drifting

- Difficult to use big data to study long-term trends
  - Population drift - system users change over time
  - Behavioural drift - use patterns may change over time
  - System drift - the system itself may change over time

10 Common Characteristics of Big Data

8. Algorithmically Confounded

- Though big data sources are typically *nonreactive*, they should not be considered "naturally occurring"
- System algorithms are often built to induce behaviour or to use social theories
- Ugander et al. (2011) - Facebook's "magic" social number
- Ensure that the theory itself was not baked into the system - - Facebook's "People You May Know"

10 Common Characteristics of Big Data

9. Dirty

- Because so much data is collected automatically, it is often messy or noisy
- Can be difficult to clean and prepare for analysis
- May contain noise or errors that can skew results
- Again: Back et al. (2010) - Pager data, identical bot messages repeated throughout the dataset with incorrectly interpretted keyword

10 Common Characteristics of Big Data

10. Sensitive

- Lack of consent/awareness of data collection can lead to privacy concerns
- Especially true to government or corporate data
- Ethical considerations are crucial when working with big data
- More of this in Chapter 6

Three Main Research Strategies

Three main strategies for studying big data:

- Counting things
- Forecasting things
- Approximating experiments
- ... this is not an exhaustive list

Three Main Research Strategies

Counting

- Simple, but powerful with a good question
- What is good? – Meaurable impact or feeds into an important decision/policy
- Again, Chetty et al. (2014) – Gave direction to future research and policy implications of intergenerational mobility

Three Main Research Strategies

Forecasting

- Predicting future events can be tricky
- 'Always-on' data allows us to do **Nowcasting**
  - Predicting the present state of the world using real-time data
  - Useful for decision-making that relies on timely and accurate information
- Beware: Google Flu Trends
  - Performance could be achieved with a simple baseline model
  - *drift* – 2009 Swine flu outbreak changed search patterns from responsive to pre-emptive/fear
  - *algorithmic confounding* – Google search algorithm changed, leading to different search patterns

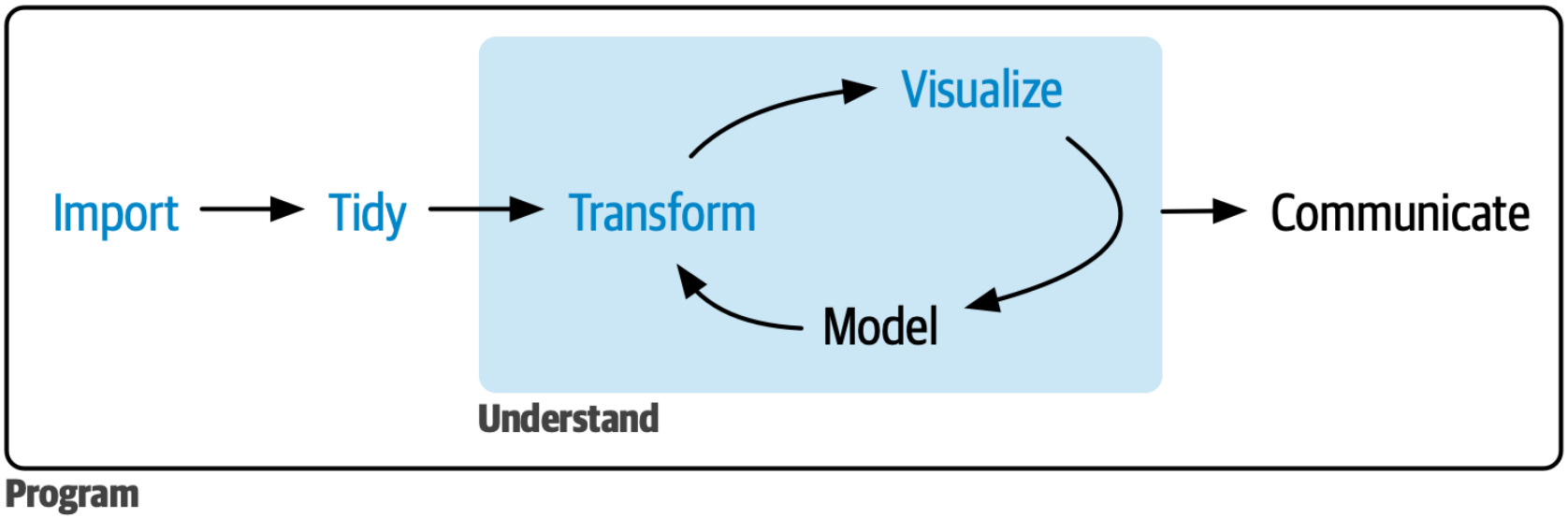
Three Main Research Strategies

Approximating experiments

- Estimating causal effects from non-experimental data when true experiments are infeasible
- Natural experiments: Leveraging naturally occurring random (or quasi-random) assignments of treatment
- Matching: Statistically adjusting non-experimental data by finding similar pairs with and without treatment

So... how do we **actually** work with this data?...

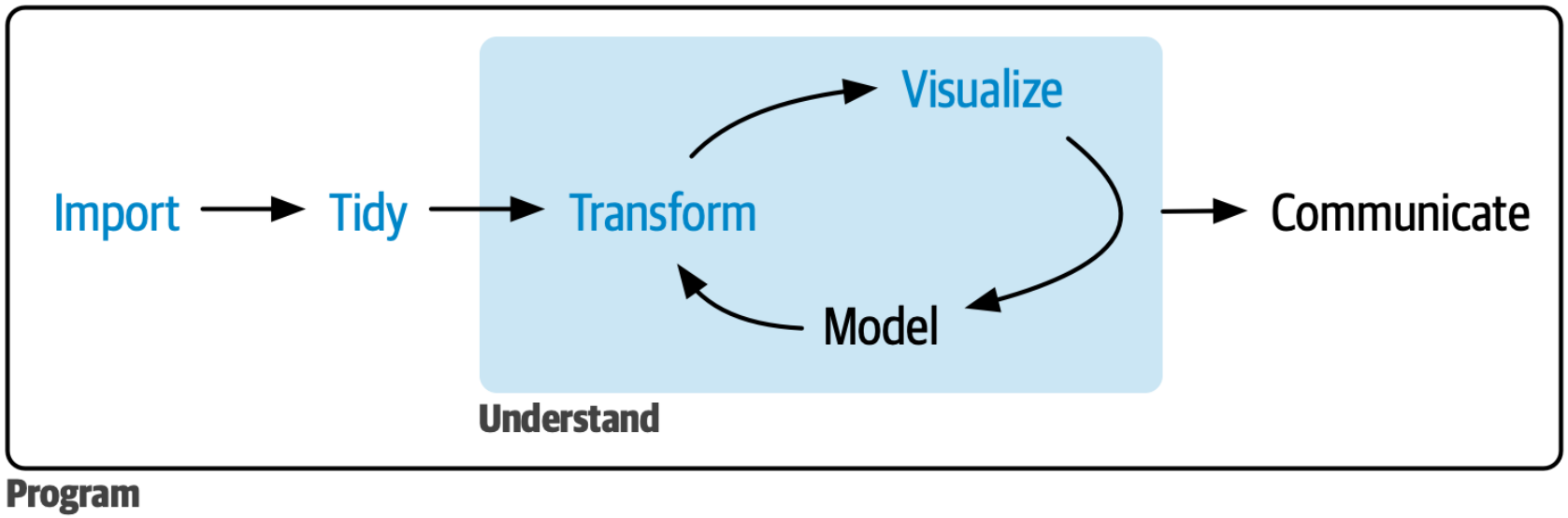
Typical Data Science Project Model



Source:R4DS, Wickham et al. (2023)

- **Import** – Getting your data into R

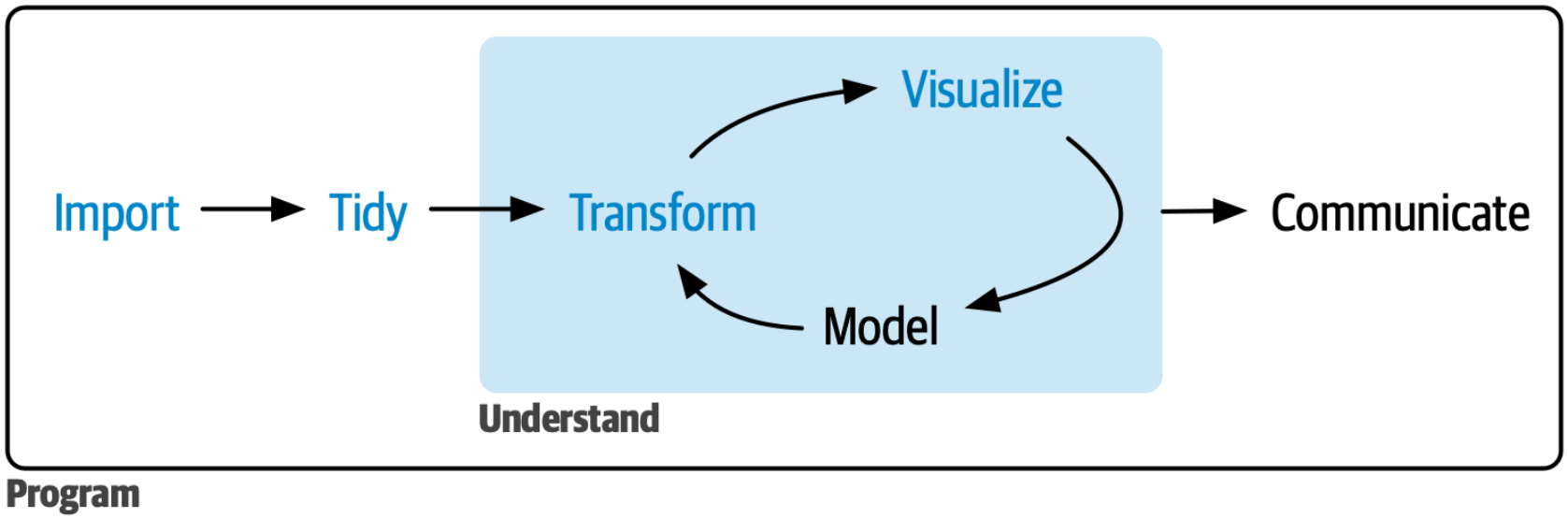
Typical Data Science Project Model



Source:R4DS, Wickham et al. (2023)

- **Tidy** – Storing your data in a consistent form

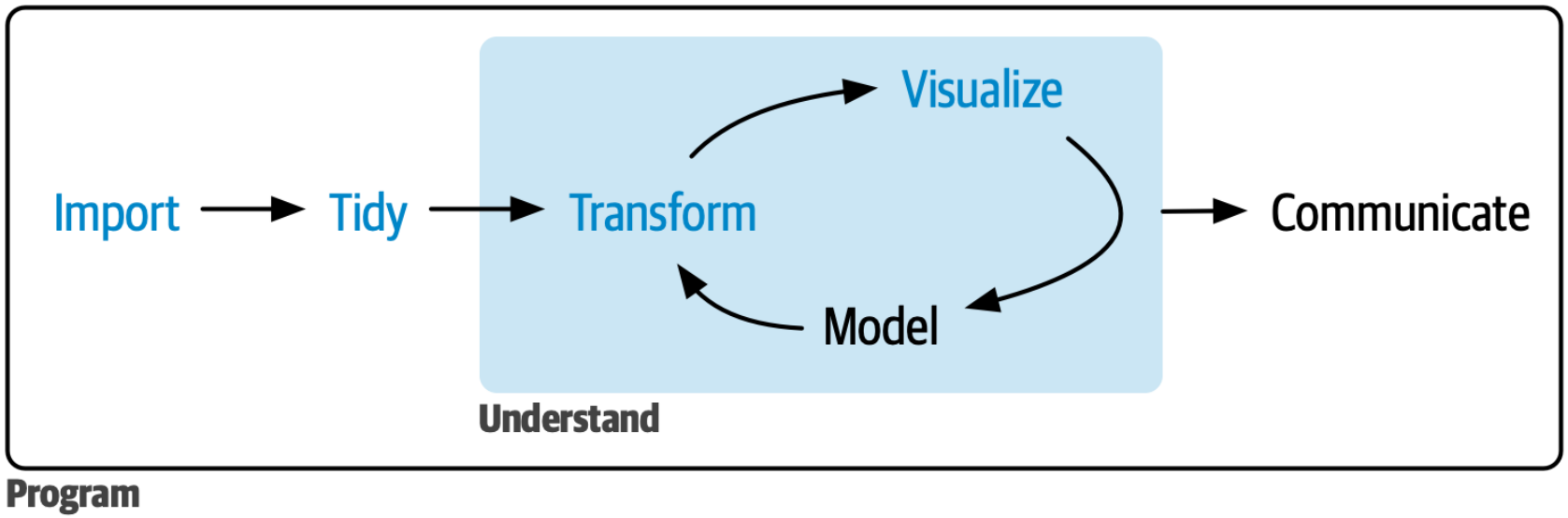
Typical Data Science Project Model



Source:R4DS, Wickham et al. (2023)

- **Transform** – Narrowing in on observations of interest and creating new variables from ones already existing

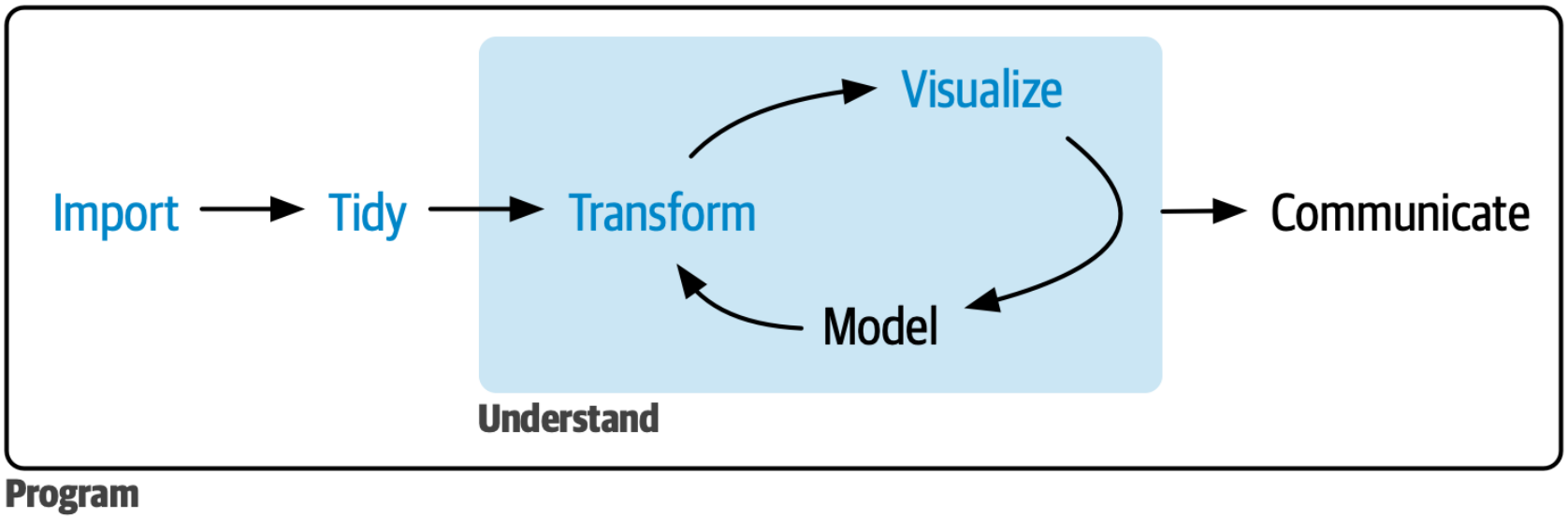
Typical Data Science Project Model



Source:R4DS, Wickham et al. (2023)

- **Visualise** – Visually representing your data

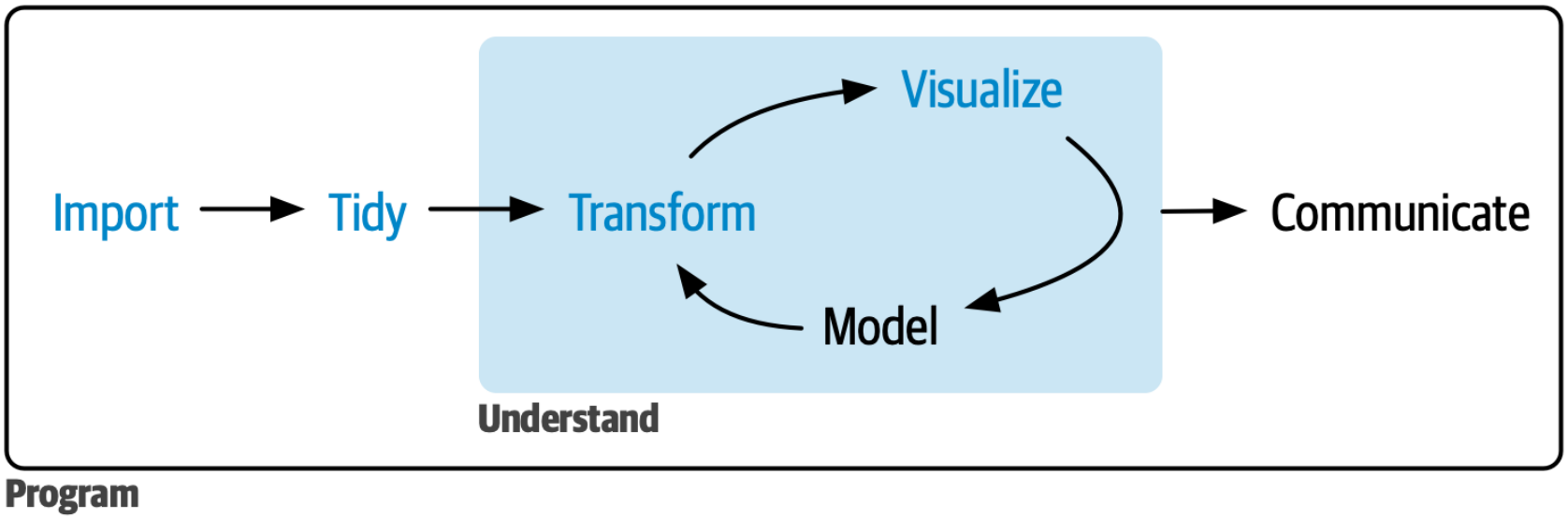
Typical Data Science Project Model



Source:R4DS, Wickham et al. (2023)

- **Model** – Using your data to answer the question at hand

Typical Data Science Project Model



Source:R4DS, Wickham et al. (2023)

- **Communicate** – Relaying your findings to others

Key Takeaways

Big Data

- Many new sources of data, but can be tricky to work with
- 10 Common characteristics
- 3 Main research strategies
- Big data sources can help researchers who can ask interesting and important questions!

Data Science Project Model

- Import, Tidy, Transform, Visualise, Model, Communicate

Next Lecture

- R, RStudio and Quarto – First steps!