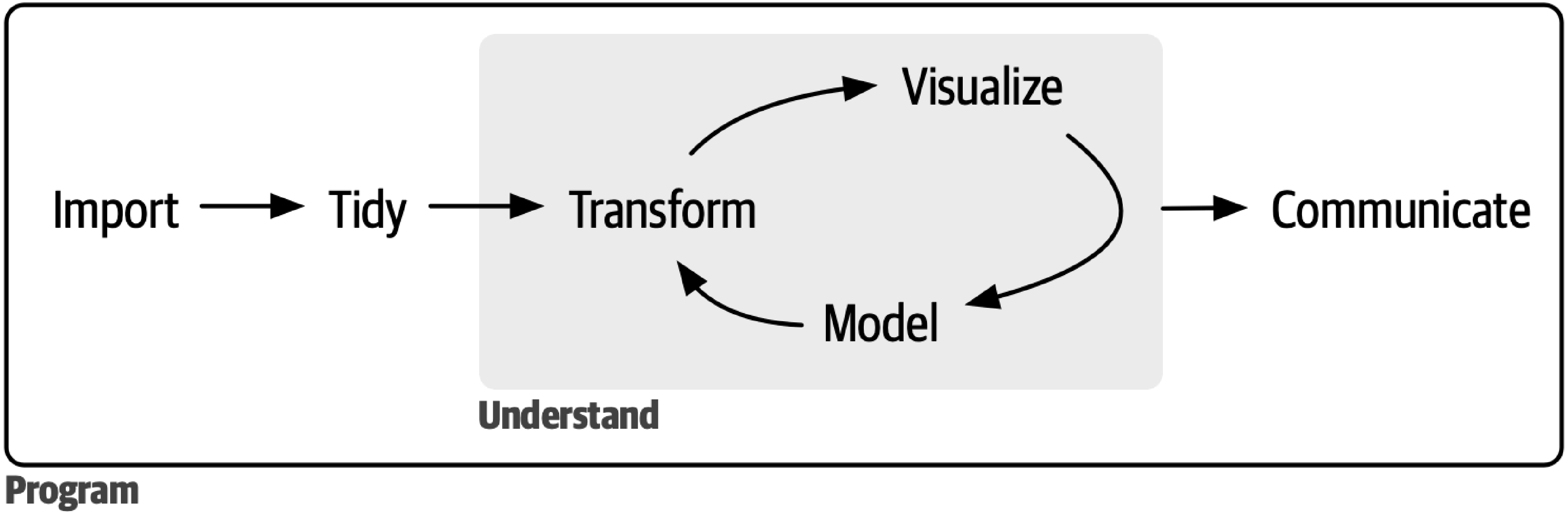
9 Data science workflow

(AST230) R for Data Science

Md Rasel Biswas

Data science tools



A typical data science project

Data science tools

Import

- Reading data from different sources, e.g., SAS, SPSS, Stata, Excel, SQL, etc.

Tidy

- When your data is tidy, each column is a variable and each row is an observation

Transform

- Transformation includes
  - narrowing in on observations of interest (like all people in one city or all data from the last year),
  - creating new variables that are functions of existing variables (like computing speed from distance and time), and
  - calculating a set of summary statistics (like counts or means).

Data science tools

Visualize

- Visualization is a fundamentally human activity.
- A good visualization will show you things you did not expect or raise new questions about the data.
- A good visualization might also hint that you're asking the wrong question or that you need to collect different data.

Model

- summarizing data
- Models are complementary tools to visualization.
- Once you have made your questions sufficiently precise, you can use a model to answer them.

Communicate

`tidyverse`

- `tidyverse` is a collection of R packages
  - `ggplot2`, `tibble`, `tidyr`, `readr`, `purrr`, `dplyr`, and many more
  - `tidyverse_packages()` → shows the complete list of `tidyverse` packages
- `tidyverse` packages share a common philosophy of data and R programming and are designed to work together naturally
  - Coding philosophy is different for functions of `tidyverse` packages compared to the base R packages

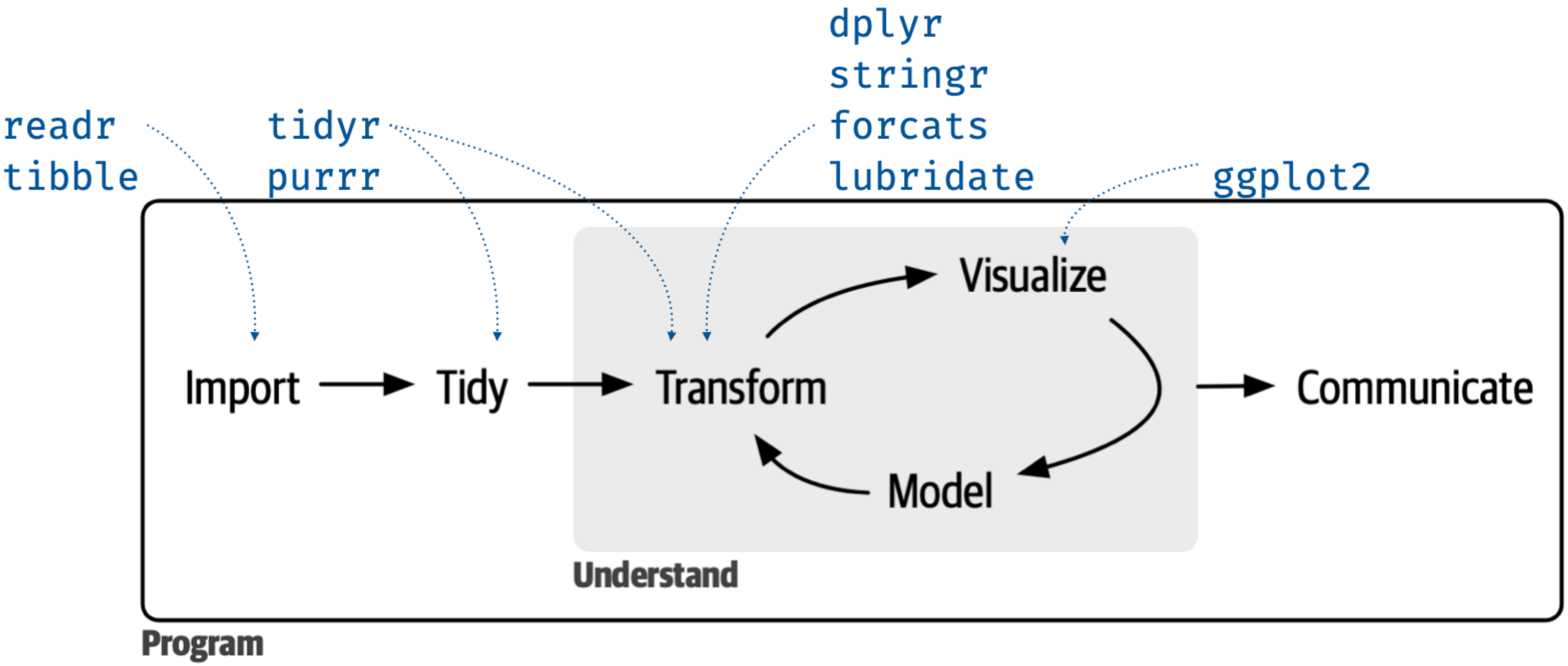
tidyverse

- **Hadley Wickham** and his colleagues have been working on `tidyverse` packages at *RStudio Inc.*
  - Wickham H and Grolemund G (2017). *R for data science* O'Reilly.
- To load all packages of `tidyverse`

```
library(tidyverse)
```

- No need to load `ggplot2` package if you load `tidyverse` family of packages

tidyverse


dplyr
stringr
forcats
lubridate
readr
tibble
tidyr
purrr
ggplot2
Visualize
Import
Tidy
Transform
Communicate
Model
Understand
Program