Up Your Bus Number

A Reproducible Data Science Workflow

Kjell & Amy Wooding

INSTITUT TUTTE INSTITUTE

https://bit.ly/bus_number

Source: Speed (1994). Special Edition.
https://www.imdb.com/title/tt0111257/

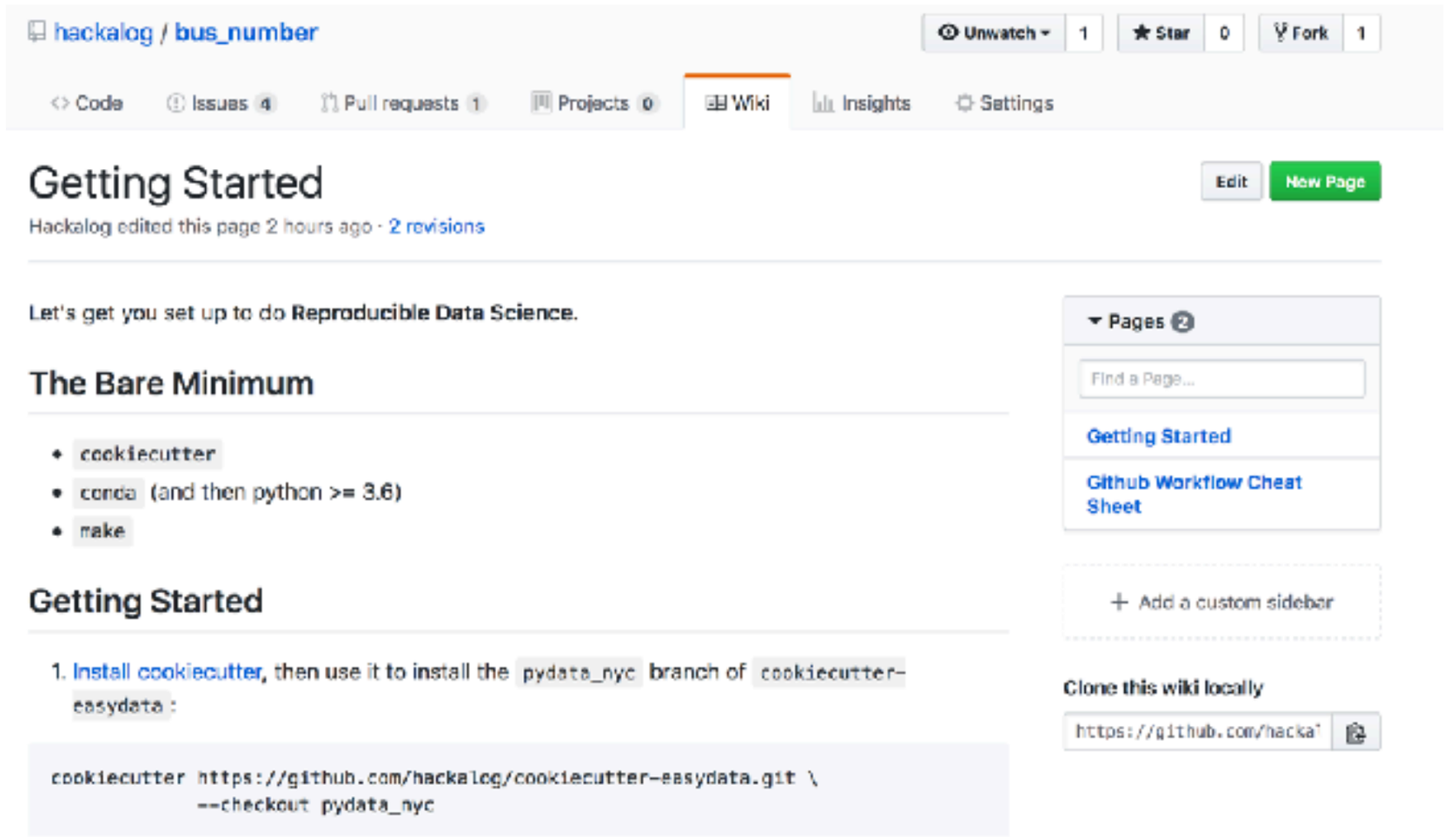
Bus Number:

The number of people who need to get hit by a bus before your data science project becomes *irreproducible*.

This might be **zero**.

While you wait...

https://bit.ly/bus_number_help



https://bit.ly/bus_number

Who Needs Reproducible Data Science?

https://bit.ly/bus_number

Meet Bjørn

- Bjørn is a dot-com millionaire. Currently he heads the Ikea R&D kitchen in Sweden.

- Bjørn employs a large number of Finnish line cooks. He can't understand a word they say.

- Bjørn needs a **trained model** to do real-time translation from Finnish to Swedish.

- Even though test kitchens have notoriously high turnover, reproducibility means Bjørn can ask his sous-chef to **deploy updated models** whenever needed.

Source: http://muppet.wikia.com/wiki/File:Swedish-chef.png

https://bit.ly/bus_number

Meet Mark

- Mark used to be an astronaut. After a rough year, he decided to change careers, and now works for a fashion magazine in NYC.

- Mark has to keep up with thousands of new fashions every week. More if it's Fashion Week.

- Mark wants to **publish a paper and software library** for accelerating fashion review using science.

- Reproducibility means his **peer reviewers** can reproduce his work, his paper will get accepted, his library will get pulled into scikit-learn, and he will transform the fashion industry forever.

Source: Youtube / 20th Century Fox

https://bit.ly/bus_number

Meet Annie

- Annie writes high frequency trading software. She moved to NYC after getting tired of the fast pace of life on the West Coast.

- Annie is constantly doing **EDA** on financial data in order to get that edge she can exploit for her firm.

- With thousands and thousands of analyses to sort through, there's no way Annie can remember everything she has done to her data.

- Reproducible data science gives Annie a way to **keep track** of the data she has worked with, the models she has built, and the experiments she has run, so that successful experiments can make it to the trading floor.

Source: https://www.imdb.com/title/tt0111257

https://bit.ly/bus_number

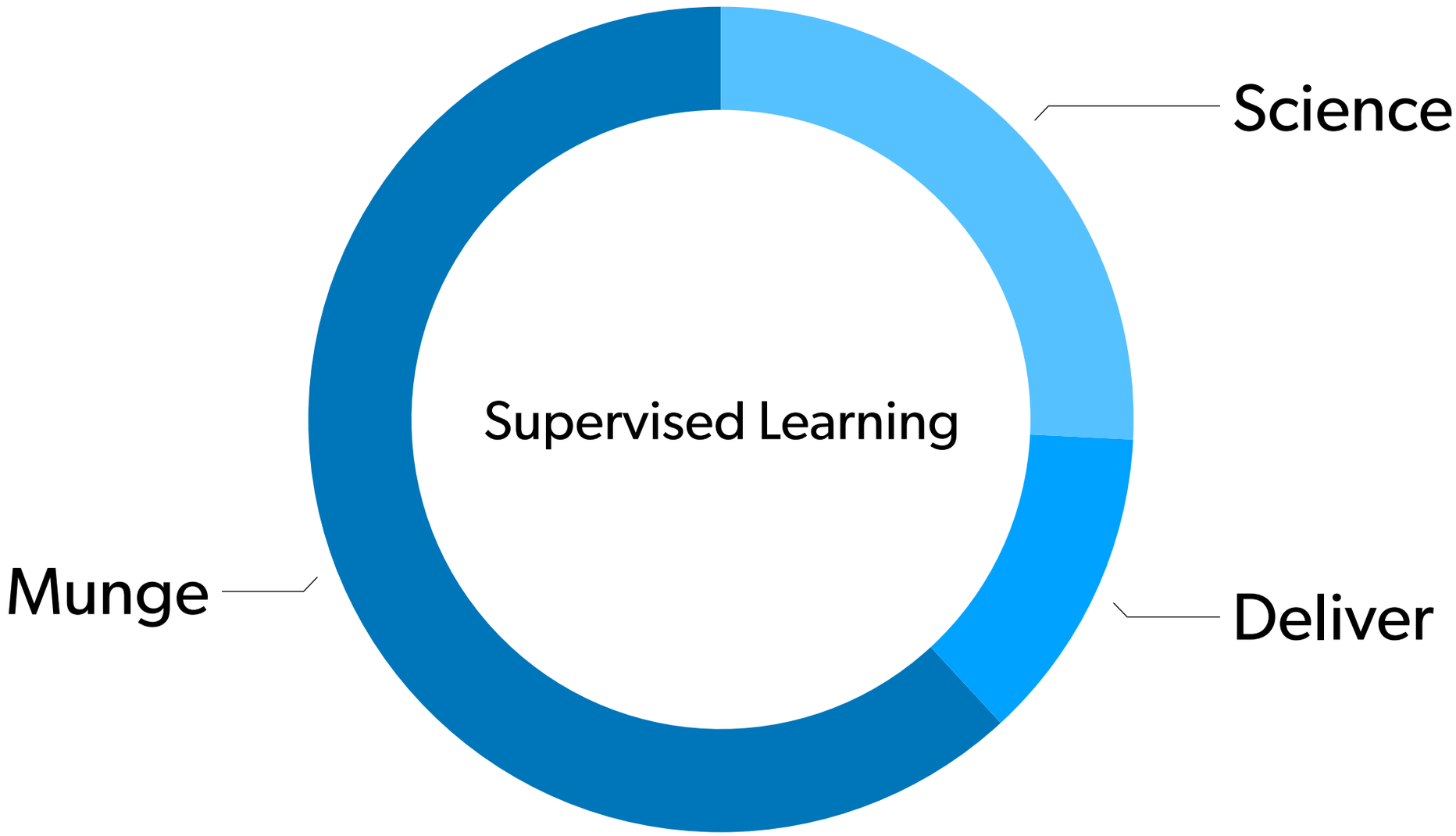
Goal	Customer	Deliverables	Why Reproducible?
Exploratory Data Analysis	You	Intuition, knowledge, interactive widgets, reusable code	To help future you
Build, Deploy Models	Your Organization	Trained models, reliable metrics	Easy to update, easy to hand off
Publish or Share	Other (Data) Scientists	Paper, slides, code	Science without reproducibility is magic

https://bit.ly/bus_number

jupyter notebook

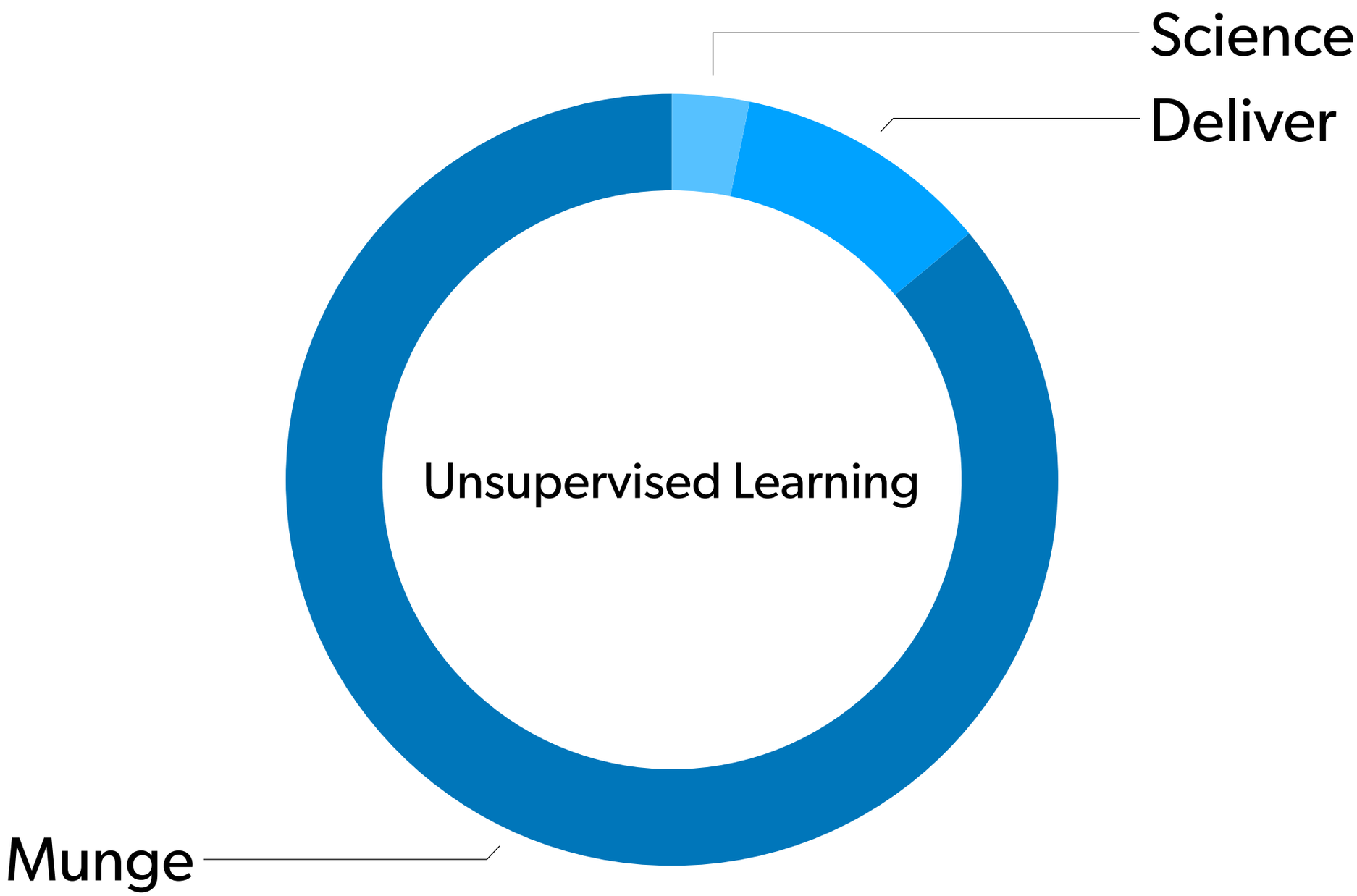
https://bit.ly/bus_number

How do you spend your “Data Science” time?



https://bit.ly/bus_number

How do you spend your “Data Science” time?



https://bit.ly/bus_number

Source: Youtube: Cårven Der Pümpkîn: https://www.youtube.com/watch?v=2Qj8PhxSnhg

The Right Tools for the Job

https://bit.ly/bus_number

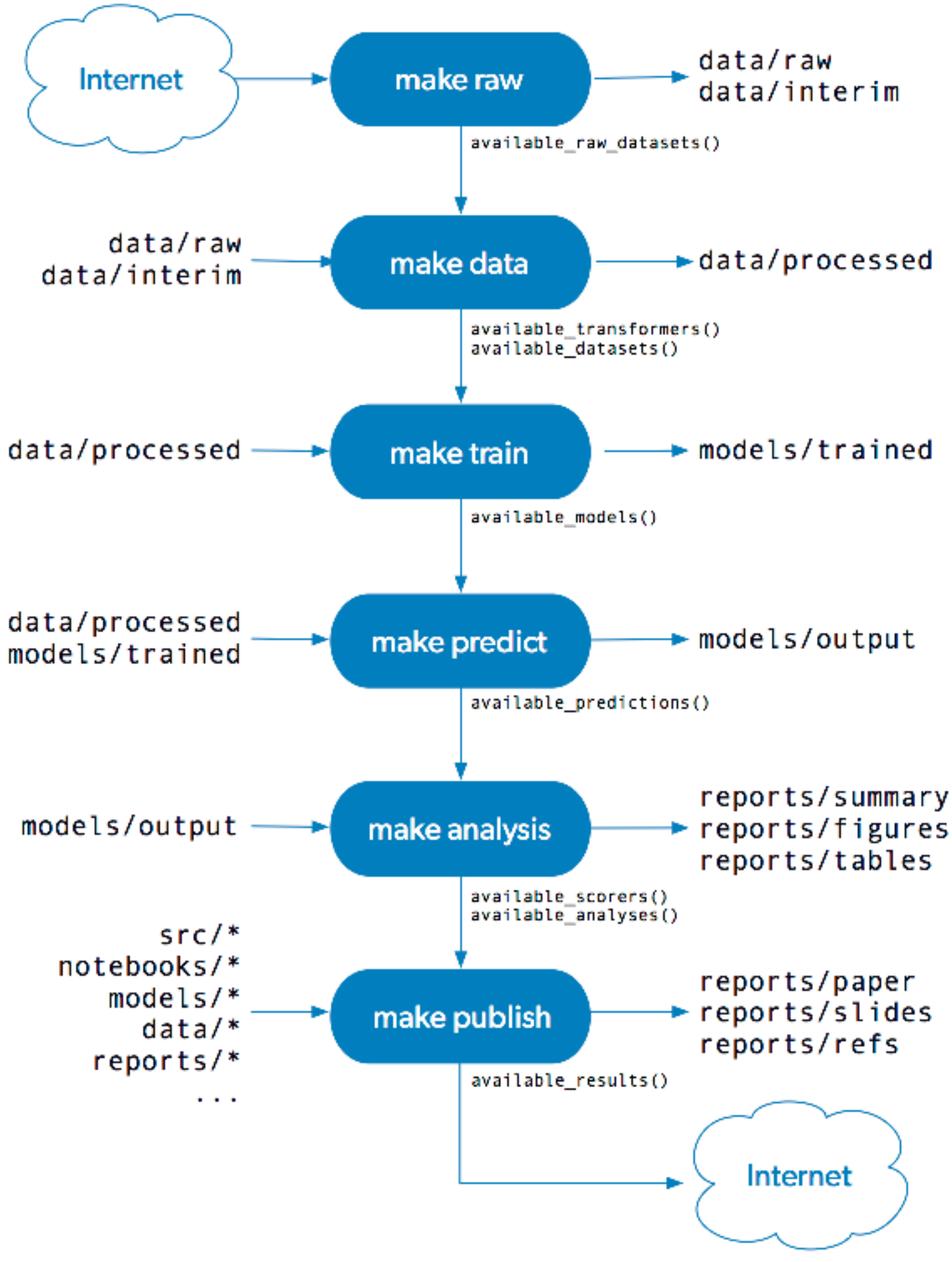
The Right Tools for the Job

- Revision Control: **git** and (github/gitlab/bitbucket)
- Language: **python 3.6+**
- Package Manager and Virtual Environments: **conda**
- Frameworks: **scikit-learn, joblib**
- IDE: **jupyter notebook**
- Scripting: **make**
- Templates: **cookiecutter**

https://bit.ly/bus_number

Data Science is a DAG

https://bit.ly/bus_number

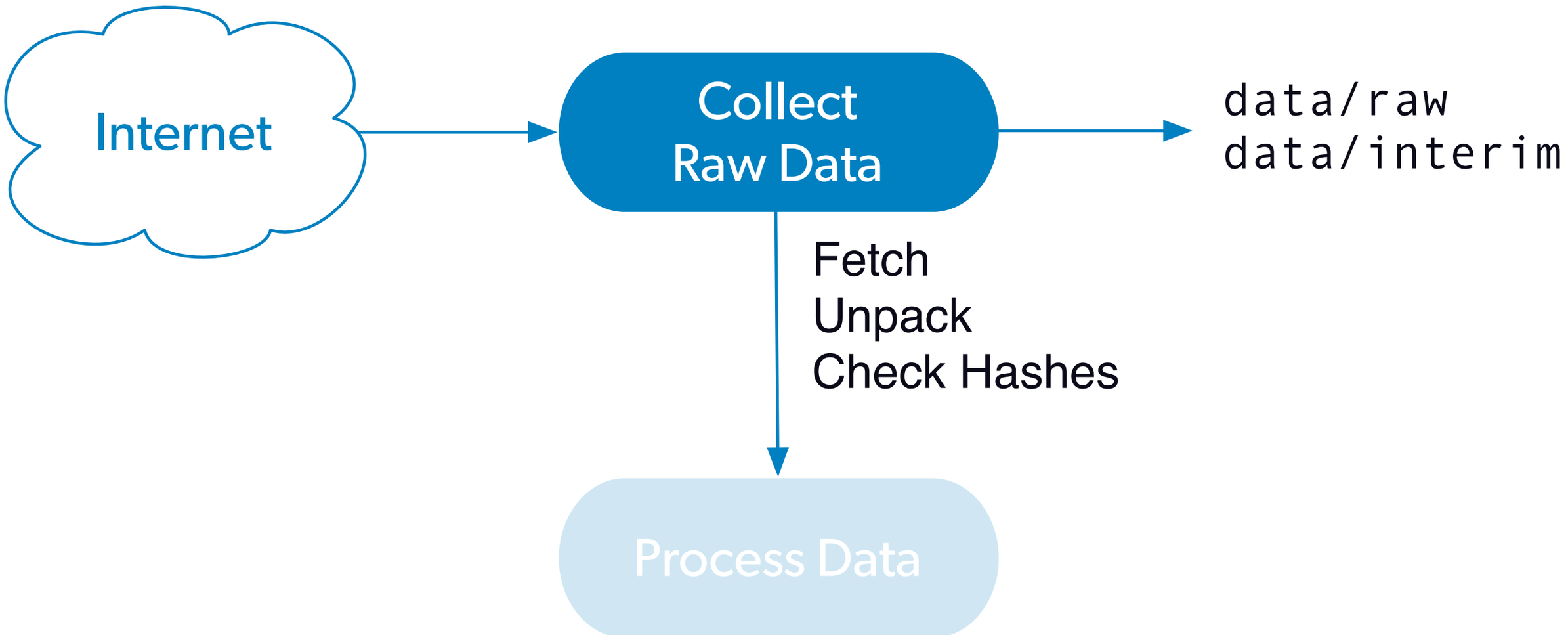


https://bit.ly/bus_n

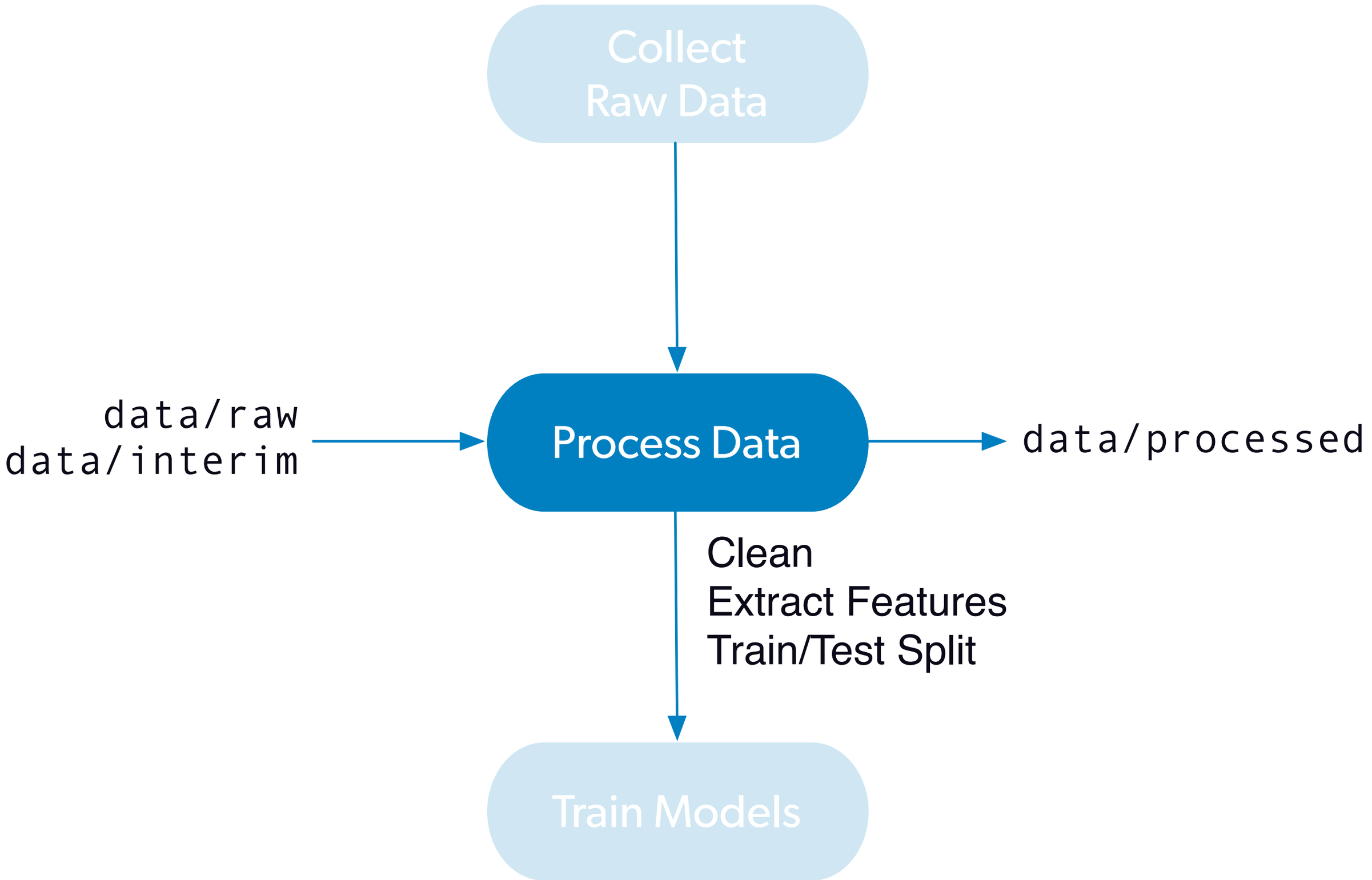
“Raw Data is Read-Only”

Sing it with me

https://bit.ly/bus_number



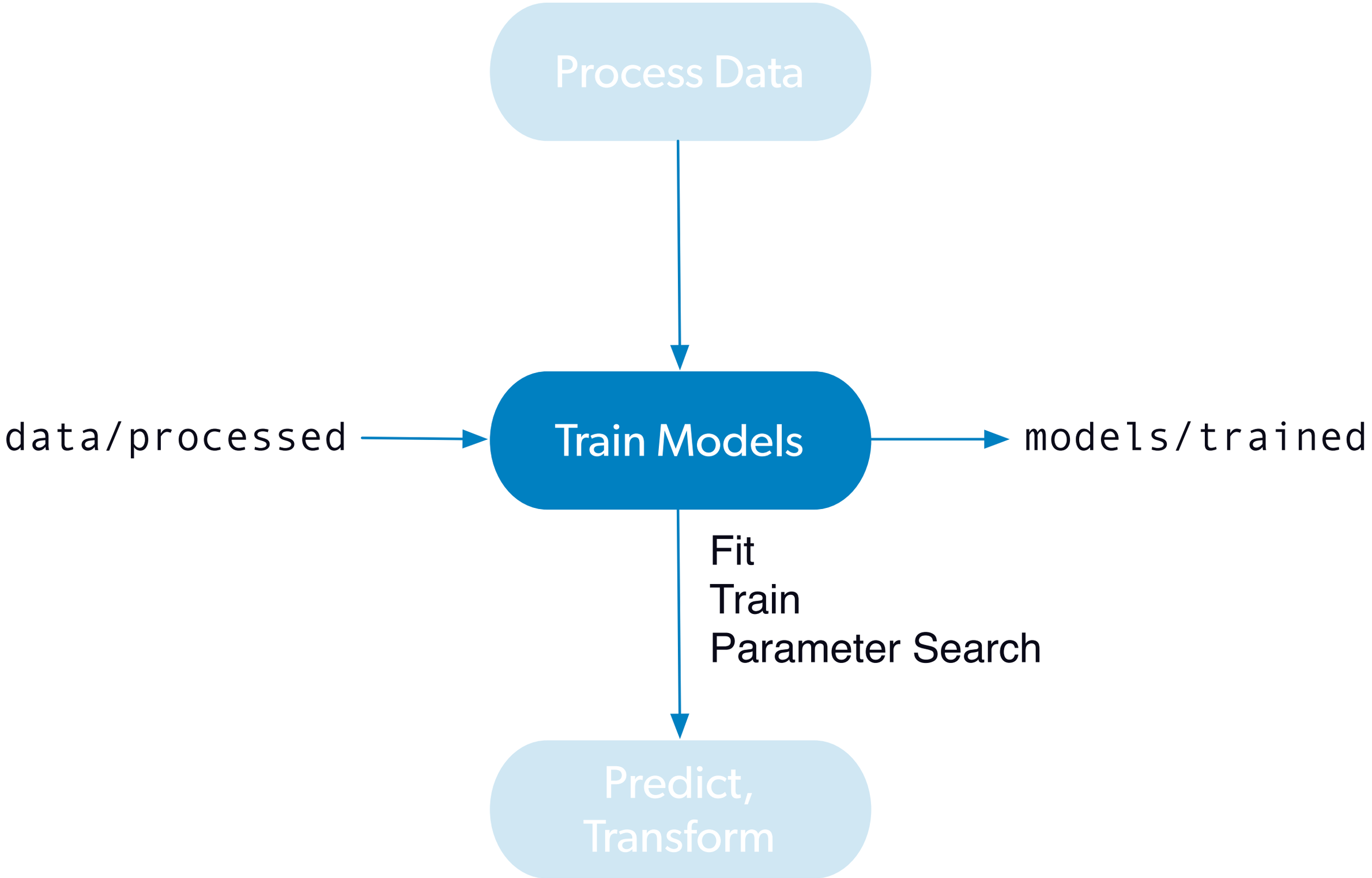
https://bit.ly/bus_number



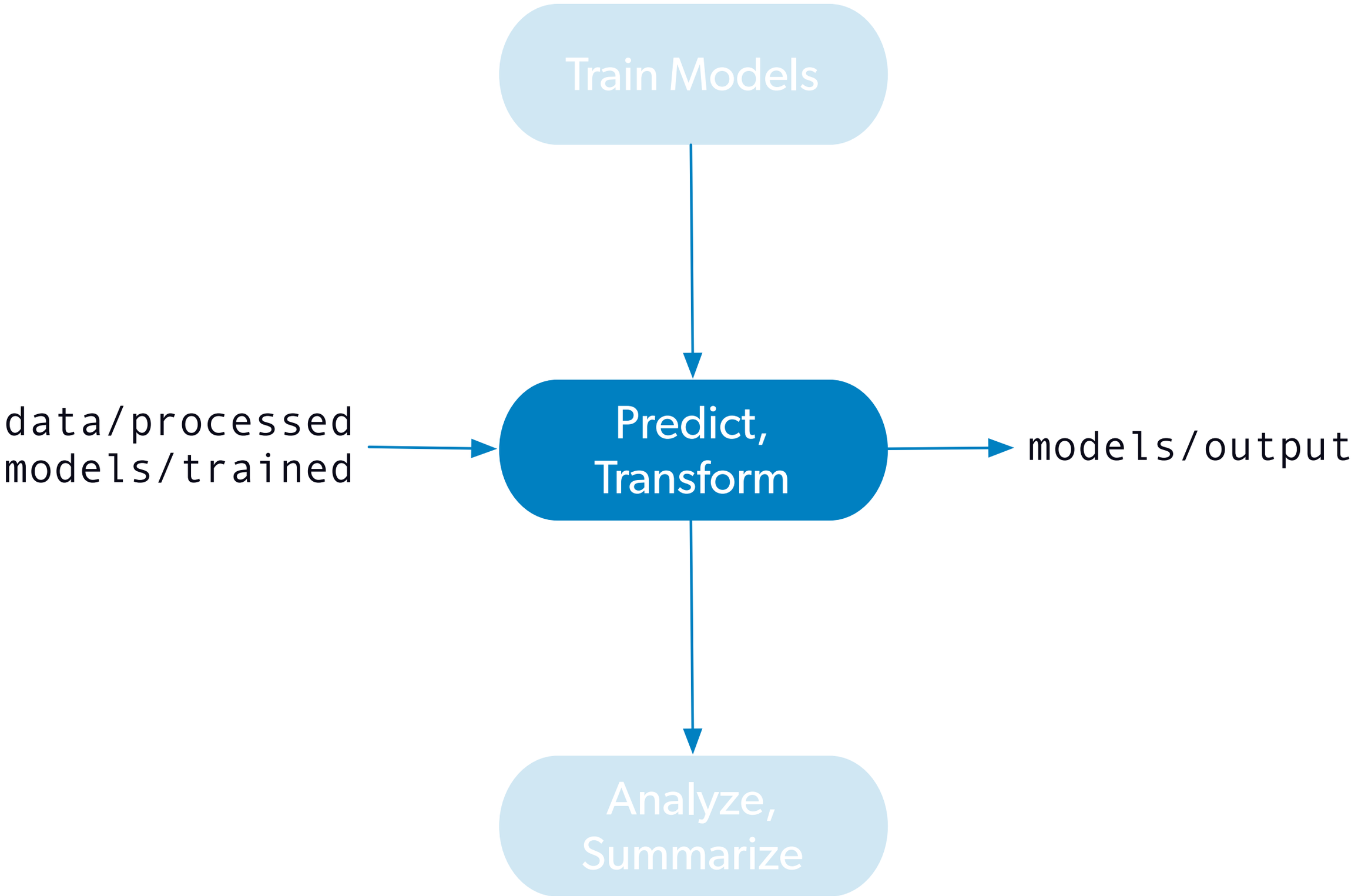
https://bit.ly/bus_number

We’re gonna (data) science the *@#! out of this

Source: https://www.imdb.com/title/tt3659388



https://bit.ly/bus_number



https://bit.ly/bus_number

I'm only here for the Pretty Pictures

Source: https://www.imdb.com/title/tt3659388

Predict,
Transform

models/output

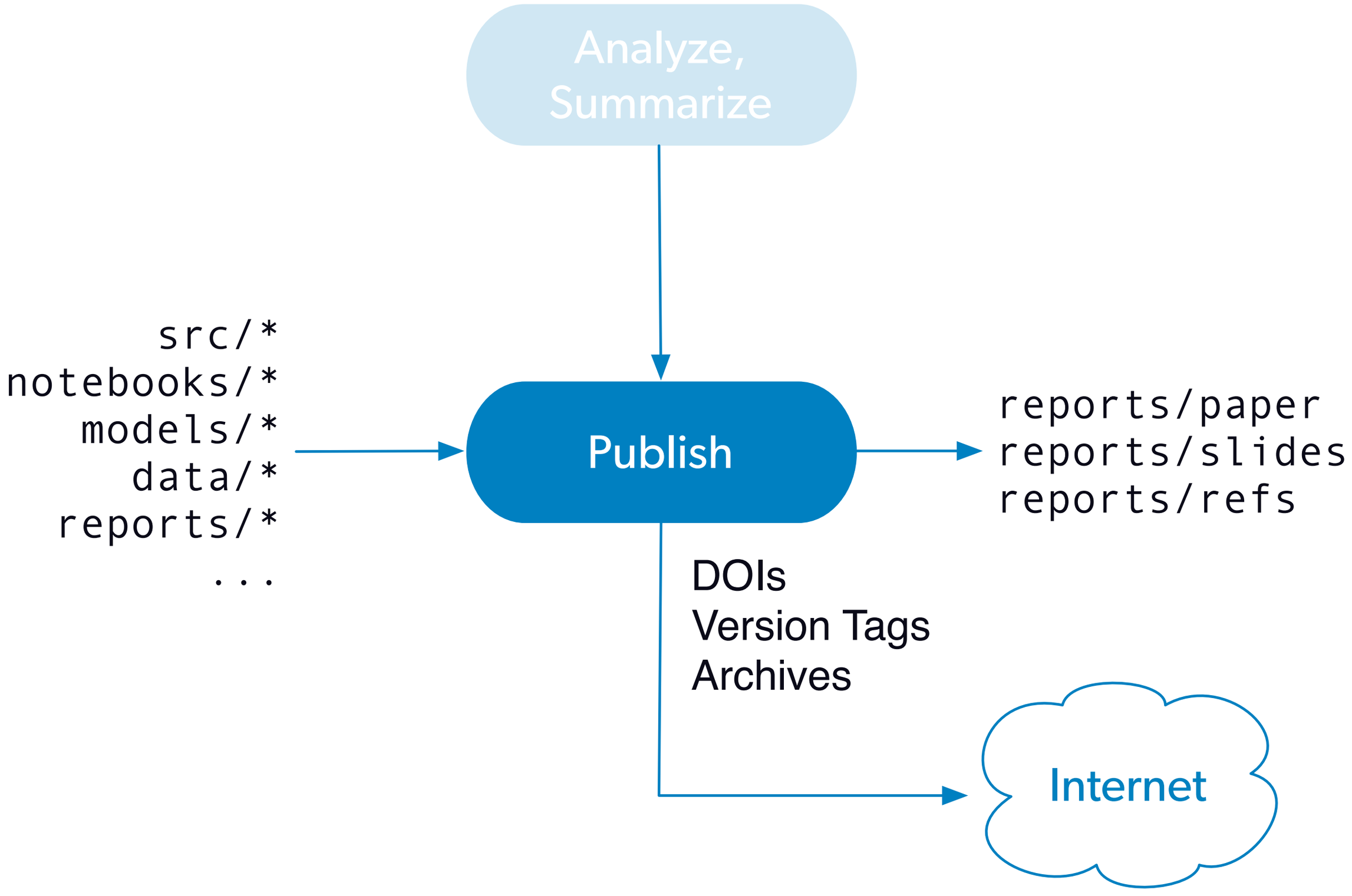
Analyze,
Summarize

reports/summary
reports/figures
reports/tables

Summary Statistics
Data Tables
Figures

Publish

https://bit.ly/bus_number



https://bit.ly/bus_number

What's your Bus Number?

Source: https://www.imdb.com/title/tt0111257

https://bit.ly/bus_number