
Data Science
@thomaskoentges
Image: https://unsplash.com/@wocintechchat

Tools Needed

Data Science

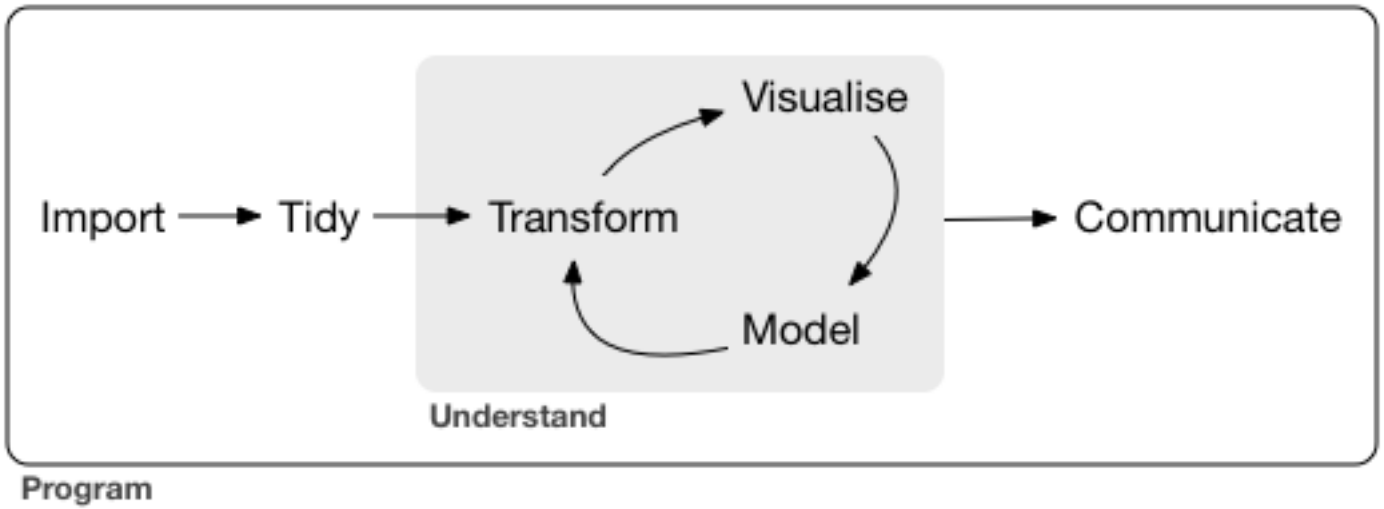


Image taken from https://r4ds.had.co.nz/introduction.html

The Different Steps

Data Science

1. Import
2. Tidy (make it rectangular)
3. Transform (Tidying + Transforming = Wrangling / Cleaning)
4. Visualise(fundamentally human)
5. Model (fundamentally computational)
6. Communicate

The First Steps

Data Science

1. Import
2. Tidy (make it rectangular)
3. Transform (Tidying + Transforming = Wrangling / Cleaning)

80% routine and boring, 20% weird and frustrating!
Unfortunately often preponderance of the work (data is messy)!

Data Analysis

Data Science

Hypothesis Generation (Exploratory Analysis)
Uses data and domain knowledge
Generate hypotheses that explain the data
Evaluated informally using skepticism

Hypothesis Confirmation (Confirmatory Analysis)
Needs mathematical model to generate falsifiable predictions
“Preregister” analysis plan (use every observation only once)

Data

@thomaskoentges

Image: https://unsplash.com/@ev

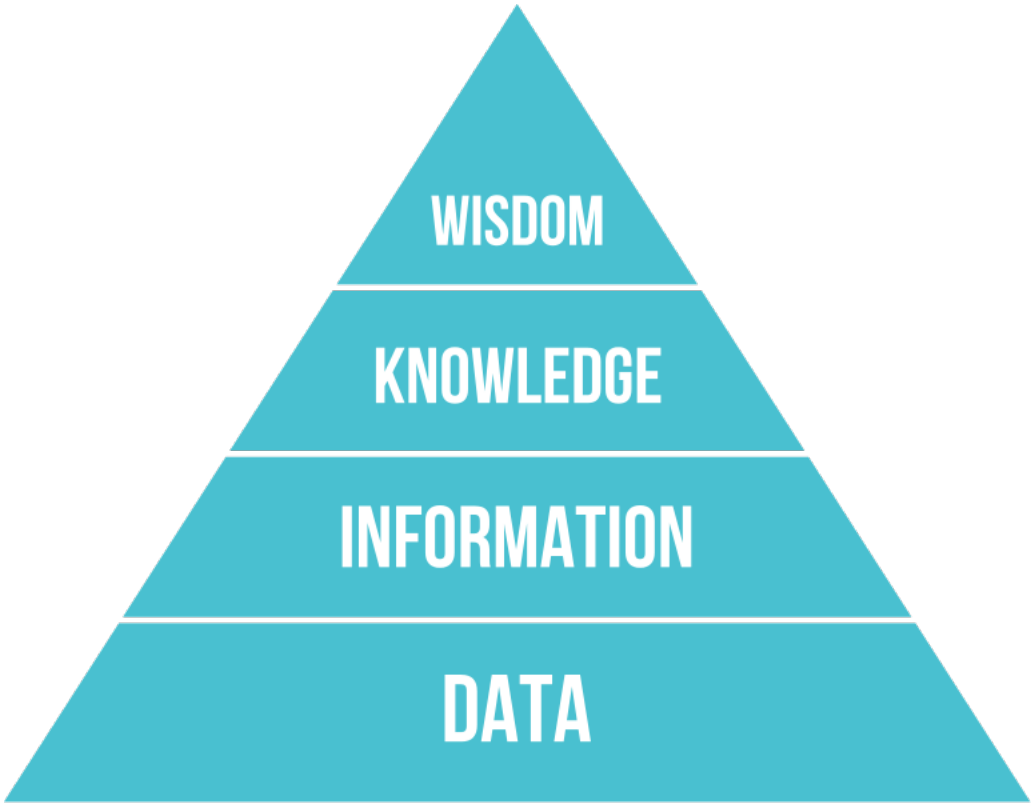
What is Data?

Data

- Recorded facts
- Based on collection
- Information stored and processed by computers
- Bits of information / fragments of information

Data + Context → Information

DIKW Pyramid



https://commons.wikimedia.org/wiki/File:DIKW_Pyramid.svg

Structured Data

Data

- Data where a property-value mechanism was implemented

Property: Name ← Value: Thomas
Property: Job ← Value: Lecturer
Property: Favourite board games ← Value: Settlers, Chess, Go

Linked Data

Data

- Describes a method to **connect structured data** using the web
- Often uses **common vocabularies** and **formats**
- Can be expressed in **triples**

RDF Triple

Data

Subject: Thomas

Predicate: Teaches

Object: Introduction to Digital Humanities

RDF Triple

Data

Subject: https://orcid.org/0000-0002-9425-5850

Predicate: http://linkedscience.org/teach/ns#teacherOf

Object: (http://studium.fmi.uni-leipzig.de/studium/wahlbereich-informatik/#10-207-0001)


Data Formats
@thomaskoentges
Image: https://unsplash.com/@sigmund

Different Formats

Data Formats

- MD: MarkDown
- TXT: Text file
- CSV: Comma Separated Value
- TSV: Tab Separated Value
- XML: Extensible Markup Language
- JSON: JavaScript Object Notation

XML

Data Formats

```
<search>
 <result-count type="integer">407</result-count>
 <results type="array">
  <result>
   <id type="integer">23171387</id>
   <title>Rugby World Cup. 2 January 2011</title>
   <collection type="array">
    <collection>TAPUHI</collection>
    <collection>New Zealand Cartoon Archive</collection>
    <collection>Hodgson, Trace, 1958- :Digital cartoons</collection>
   </collection>
  </result>
     </results>
</search>
```

JSON

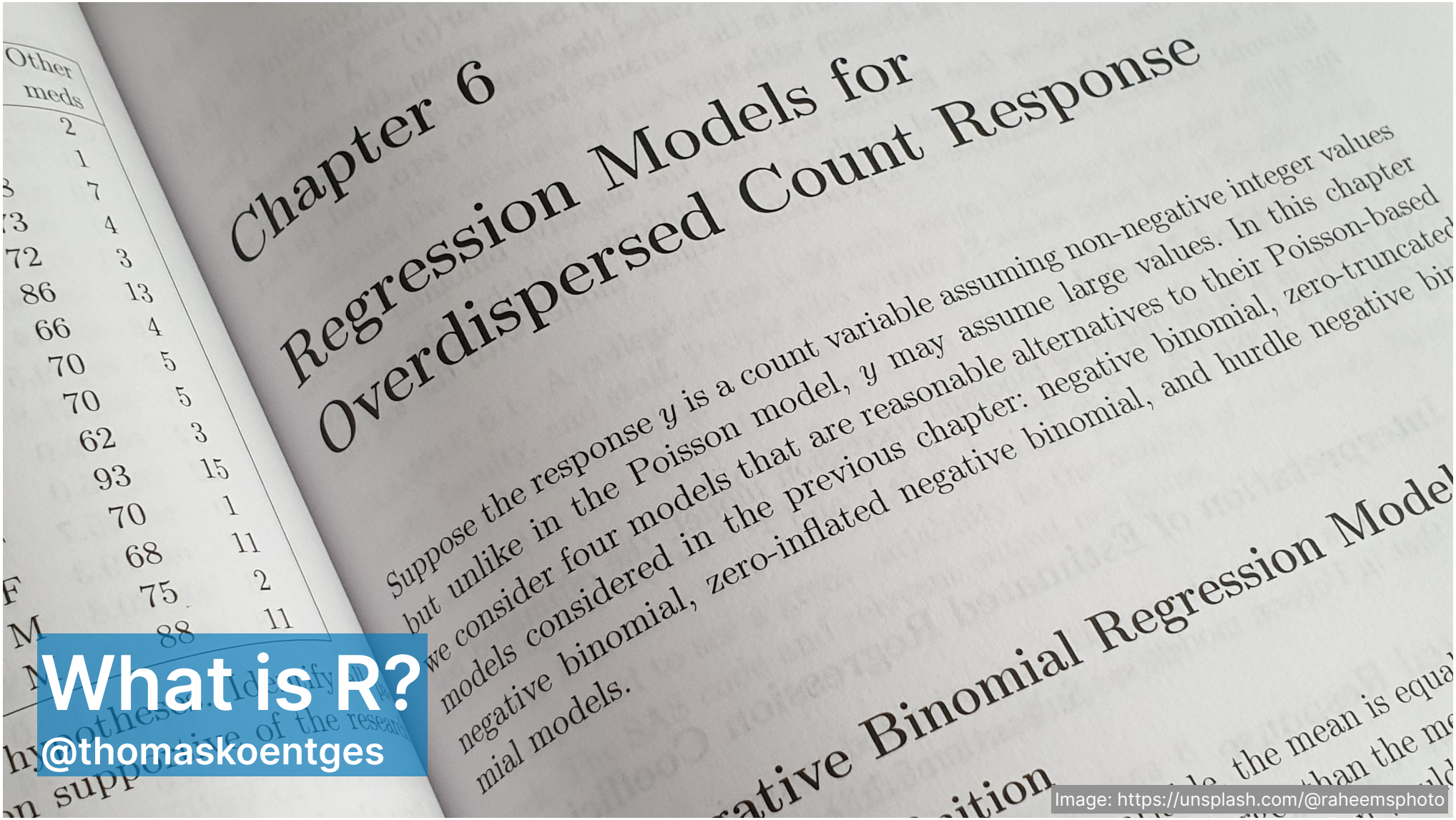
Data Formats


```
{"search":
{"result_count":407,"results":[
{"id":23171387,"title":"Rugby World Cup. 2 January 2011","collection":
["TAPUHI","New Zealand Cartoon Archive","Hodgson, Trace, 1958- :Digital
cartoons"]},
{"id":22795854,"title":"Rugby Southland. 2 January 2011","collection":
["TAPUHI","New Zealand Cartoon Archive","Winter, Mark, 1958- :[Digital cartoons
published in the Southland Times and other papers]"]
]}
}
```

Other Important Formats

Data Formats

- PDF
- TIFF
- PNG (50 4E 47)
- XSLX (zipped XML files)
- SPSS System Data File Format Family (.sav)


Other meds
2
1
7
4
3
13
4
5
5
3
15
1
11
2
11
Chapter 6
Regression Models for Overdispersed Count Response
Suppose the response y is a count variable assuming non-negative integer values but unlike in the Poisson model, y may assume large values. In this chapter we consider four models that are reasonable alternatives to their Poisson-based models considered in the previous chapter: negative binomial, zero-truncated negative binomial, zero-inflated negative binomial, and hurdle negative bi
ative Binomial Regression Model
What is R?
@thomaskoentges
Image: https://unsplash.com/@raheemsphoto

What is R?

```
Error in match.fun(FUN) : argument
"FUN" is missing, with no default
```

Features

What is R?

- Dynamically typed language (mainly for data analysis)
- R Foundation for Statistical Computing
- Big community of statisticians
- Bad reputation: slow, data import struggle (some people suggest using python to import and clean data, then use R for analysis)

R: “Hello World!”

What is R?

```
print("Hello World!”)
```

R Shiny: “Hello World!”

What is R?

```
library(shiny)
server <- function(input, output) {
          output$helloWorld <- renderText({
           "hello world”
           })}
ui <- fluidPage(
        sidebarLayout(
          sidebarPanel(),
          mainPanel(textOutput(“helloWorld"))
          ))
shinyApp(ui = ui, server = server)
```

R vs. SPSS

What is R?

- Free
- Object-oriented and interpreted programming language
- Can be used as open-source analytics package that can replace or complement commercial analytic tools
- “no GUI”
- Processes all kinds of data
- Extensible
- No helpline, but very large community and stackoverflow

- Expensive
- Proprietary data analytics tool
- GUI
- Processes rectangular data (which is pressed into tables)
- Not as easily extensible
- Helpline

Closing Remarks

@thomaskoentges

Image: https://unsplash.com/@jasmund

5-Star Linked Open Data

Closing Remarks

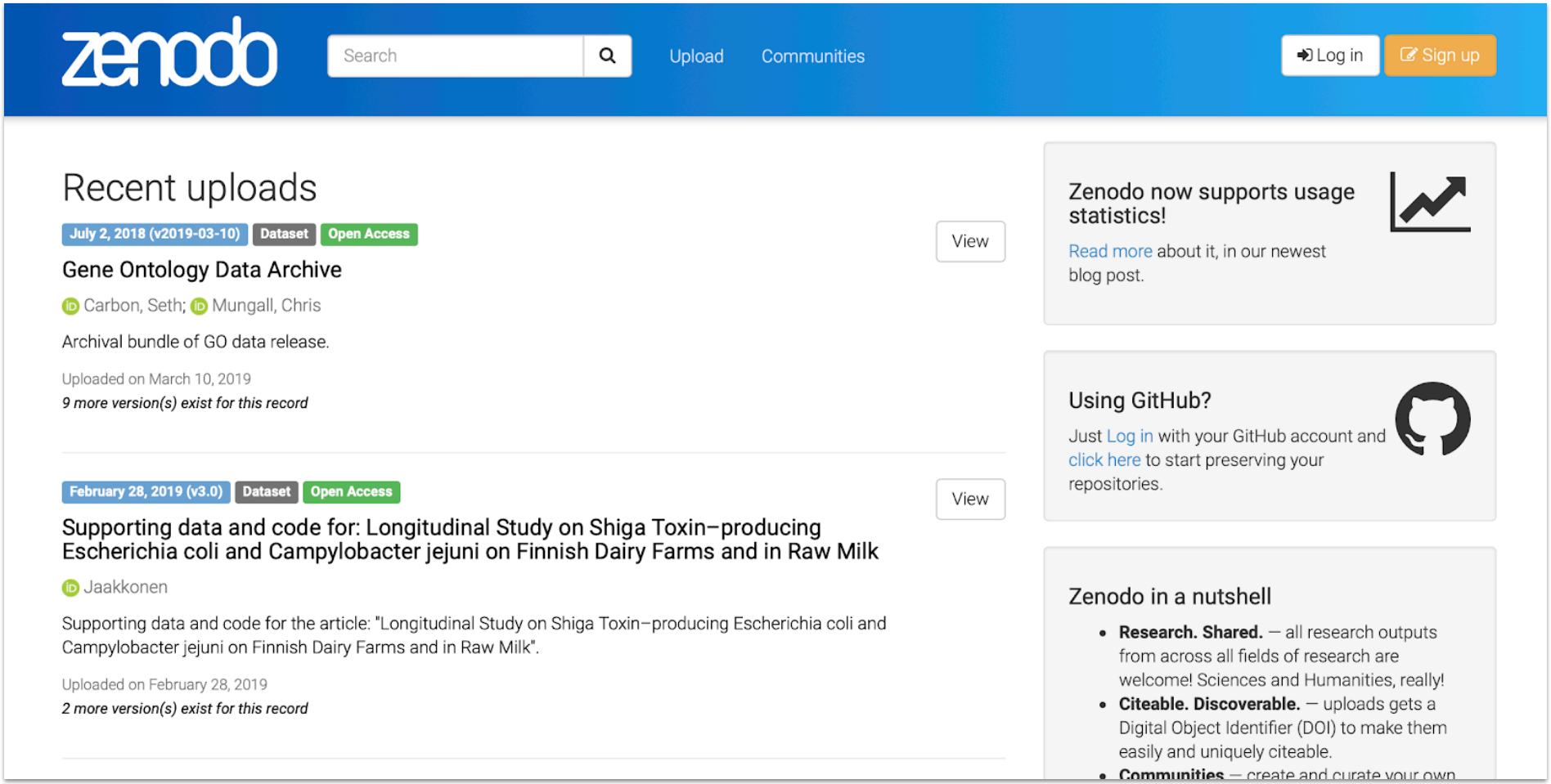
★Make your data available on the Web under an open license

★Make it available as structured data

★Make it available in a non-proprietary open format

★Use URIs

★Link your data to other data to provide context


zenodo
Search
Upload
Communities
Log in
Sign up
Recent uploads
July 2, 2018 (v2019-03-10)
Dataset
Open Access
View
Gene Ontology Data Archive
Carbon, Seth; Mungall, Chris
Archival bundle of GO data release.
Uploaded on March 10, 2019
9 more version(s) exist for this record
February 28, 2019 (v3.0)
Dataset
Open Access
View
Supporting data and code for: Longitudinal Study on Shiga Toxin–producing Escherichia coli and Campylobacter jejuni on Finnish Dairy Farms and in Raw Milk
Jaakkonen
Supporting data and code for the article: "Longitudinal Study on Shiga Toxin–producing Escherichia coli and Campylobacter jejuni on Finnish Dairy Farms and in Raw Milk".
Uploaded on February 28, 2019
2 more version(s) exist for this record
Zenodo now supports usage statistics!
Read more about it, in our newest blog post.
Using GitHub?
Just Log in with your GitHub account and click here to start preserving your repositories.
Zenodo in a nutshell
Research. Shared. — all research outputs from across all fields of research are welcome! Sciences and Humanities, really!
Citeable. Discoverable. — uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
Communities — create and curate your own

Open Source & Open Data

The Rolling Swarm